UNIVERSITAT
POLITÈCNICA
DE VALÈNCIA

CARMA 2018

2nd International Conference on Advanced Research Methods and Analytics

July 12-13, 2018 · Valencia, Spain

Congress UPV

2nd International Conference on Advanced Research Methods and Analytics (CARMA 2018)

The contents of this publication have been evaluated by the Scientific Committee according to the procedure described in the preface. More information at http://www.carmaconf.org/

Scientific Editors

Josep Domenech
María Rosalía Vicente
Desamparados Blazquez

Publisher

2018, Editorial Universitat Politècnica de València
www.lalibreria.upv.es / Ref.: 6447_01_01_01

ISBN: 978-84-9048-689-4 (print version)
Print on-demand

DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8742

Preface

Domenech, Josep [a]; Vicente, María Rosalía [b]; Blazquez, Desamparados [a]
[a] Dept. Economics and Social Sciences, Universitat Politècnica de València, Spain. [b] Dept. Applied Economics, Universidad de Oviedo, Spain

Abstract

Research methods in economics and social sciences are evolving with the increasing availability of Internet and Big Data sources of information. As these sources, methods, and applications become more interdisciplinary, the 2nd International Conference on Advanced Research Methods and Analytics (CARMA) is an excellent forum for researchers and practitioners to exchange ideas and advances on how emerging research methods and sources are applied to different fields of social sciences as well as to discuss current and future challenges.

Keywords: *Big Data sources, Web scraping Social media mining, Official Statistics, Internet Econometrics, Digital transformation, global society.*

1. Preface to CARMA2018

This volume contains the selected papers of the Second International Conference on Advanced Research Methods and Analytics (CARMA 2018) hosted by the Universitat Politècnica de València, Spain during 12 and 13 July 2018. This second edition consolidated CARMA as a unique forum where Economics and Social Sciences research meets Internet and Big Data. CARMA provided researchers and practitioners with an ideal environment to exchange ideas and advances on how Internet and Big Data sources and methods contribute to overcome challenges in Economics and Social Sciences, as well as on the changes in the society because of the digital transformation.

The selection of the scientific program was directed by Maria Rosalia Vicente, who led an international team of 33 scientific committee members representing 28 institutions. Following the call for papers, the conference received 73 paper submissions from all around the globe. All submissions were reviewed by the scientific committee members under a double blind review process. Finally, 40 papers were accepted for oral presentation during the conference. This represents an overall paper acceptance rate of 54%, ensuring a high quality scientific program. It covers a wide range of research topics in Internet and Big Data, including nowcasting people mobility and economic indicators, applications of Big Data methods in retail and finance, using search and social media data, among others.

CARMA 2018 also featured two special sessions on “Big Data for Central Banks” and “Using Big Data in Official Statistics,” chaired by Juri Marcucci and Gian Luigi Mazzi, respectively. Both sessions gave a complementary institutional perspective on how to use Internet and Big Data sources and methods for public policy and official statistics. The perspective from the private sector was contributed by Norbert Wirth, who talked about “Data Science development at scale” in his keynote speech.

The conference organizing committee would like to thank all who made this second edition of CARMA a great success. Specifically, thanks are indebted to the authors, scientific committee members, reviewers, invited speakers, session chairs, presenters, sponsors, supporters and all the attendees. Our final words of gratitude must go to the Faculty of Business Administration and Management of the Universitat Politècnica de València for supporting CARMA 2018.

2. Organizing Committee

General chair

Josep Domènech, Universitat Politècnica de València

Scientific committee chair

María Rosalía Vicente, Universidad de Oviedo

Local arrangements chair

Desamparados Blazquez, Universitat Politècnica de València

3. Sponsors

BigML
DevStat

4. Supporters

Universitat Politècnica de València
Facultad de Administración y Dirección de Empresas
Departamento de Economía y Ciencias Sociales

4. Scientific committee

Concha Artola, Banco de España
Nikolaos Askitas, IZA – Institute of Labor Economics
Jose A. Azar, IESE Business School
Silvia Biffignandi, University of Bergamo
Petter Bae Brandtzaeg, SINTEF
Jonathan Bright, Oxford Internet Studies
José Luis Cervera, DevStat
Piet Daas, Statistics Netherlands
Pablo de Pedraza, Universidad de Salamanca / University of Amsterdam
Giuditta de Prato, European Commission – JRC Directorate B
Rameshwar Dubey, Montpellier Business School
Enrico Fabrizi, DISES – Università Cattolica del S. Cuore
Juan Fernández de Guevara, IVIE and University of Valencia
Jose A. Gil, Universitat Politècnica de València

Felix Krupar, Max-Planck-Institute for Innovation and Competition
Caterina Liberati, University of Milano-Bicocca
Juri Marcucci, Bank of Italy
Rocio Martinez Torres, Universidad de Sevilla
Esteban Moro, Universidad Autónoma de Madrid / Universidad Carlos III
Michela Nardo, European Commission – Joint Research Centre
Enrique Orduña, Universitat Politècnica de València
Bulent Ozel, University of Zurich / Universitat Jaume I
Andrea Pagano, European Commission – Joint Research Centre
Ana Pont, Universitat Politècnica de València
Ravichandra Rao, Indian Statistical Institute
Pilar Rey del Castillo, Instituto de Estudios Fiscales
Anna Rosso, DEMM University of Milan
Vincenzo Spiezia, OECD
Pål Sundsøy, NBIM/Norway
Sergio L. Toral Marin, Universidad de Sevilla
Antonino Virgillito, Italian Revenue Agency
Sang Eun Woo, Purdue University
Zheng Xiang, Virginia Tech

Index

Full papers

Abstracts

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8292

Blockchain-backed analytics: Adding blockchain-based quality gates to data science projects

Herrmann, Markus; Petzold, Jörg and Bombatkar, Vivek
Technology & Data, GfK SE, Germany

Abstract

A typical analytical lifecycle in data science projects starts with the process of data generation and collection, continues with data preparation and pre-processing and heads towards project specific analytics, visualizations and presentations. In order to ensure high quality trusted analytics, every relevant step of the data-model-result linkage needs to meet certain quality standards that furthermore should be certified by trusted quality gate mechanisms.

We propose "blockchain-backed analytics", a scalable and easy-to-use generic approach to introduce quality gates to data science projects, backed by the immutable records of a blockchain. For that reason, data, models and results are stored as cryptographically hashed fingerprints with mutually linked transactions in a public blockchain database.

This approach enables stakeholders of data science projects to track and trace the linkage of data, applied models and modeling results without the need of trust validation of escrow systems or any other third party.

Keywords: *Blockchain; Data Science; Data Management; Trusted Data; Trusted Analytics.*

1. Trusted analytics

A typical analytical lifecycle in data science projects starts with the process of data creation and collection, continues with data preparation and pre-processing ("data wrangling") and heads towards project specific analytics, visualization and presentation, i.e. the results. To enforce trusted analytics, every step of the data lifecycle and applied analytics of an analytics project needs to meet certain quality standards. While these standards may vary broadly among academia and industries, there is one common challenge for every field of trusted analytics: How to publicly document trusted data and analytics in an immutable way? And if possible, without the involvement of any third party ensuring the trust.

Why is this considered a challenge? In academia, trusted analytics is achieved by peer-reviewed processes and bibliographical documentation. In data-driven industry sectors, on the other hand, the massive amount of decentralized data being generated daily and the huge number of data science and analytics projects worldwide cannot be evaluated and documented by any manual or human review system in a reasonable amount of time.

With the recently matured possibilities of machine learning – and in general the field of artificial intelligence - the documentation of the data-model-result relationship will become more and more relevant and consequently requires a scalable and immutable data and information documentation solution as a quality gate.

A decentralized storage system based on blockchain technology is able to introduce such quality gates to data science projects. For this reason we propose "blockchain-backed analytics"; whereby the data, applied methods and relevant results are stored as cryptographically hashed fingerprints in an immutable blockchain database.

Delivering this scalable and easy-to-use generic approach means being able to track and trace the linkage of data, models and modeling results, without the need of involving escrow systems or any other third party.

Our work builds on existing research in the fields of blockchain-based data protection and identity management, where blockchain technology is being applied to secure the management of digital identities and protect data ownership (Zyskind et. al, 2015). Accordingly blockchain-backed analytics is an extension of blockchain-based identity management techniques to data science projects and is therefore particularly relevant for data-driven academic research and industry projects.

2. Blockchain technology

To date the application of the blockchain technology is predominantly influenced by Satoshi Nakamoto's design of the cryptocurrency Bitcoin, which is based on the consensus in a distributed system, achieved with a Proof-of-Work algorithm (Nakamoto, 2008).

In this regard, a blockchain is a distributed database that is continuously keeping records of transactions in a logical order and in sync across participants, i.e. instances. Multiple transactions are bundled and stored in a block, whereas new blocks are sequentially appended to the previous block(s), with each block containing a list of cryptographically signed transactions with timestamps. In order to ensure integrity of the blockchain, each new block contains a pointer to a distinct hash value of the previous block and – in most cases – the root hash of the Merkle tree of all transactions of the previous block as well. This can be considered as a hash chain of all transactions of the block (ibid.).

The sequence of inter-linked blocks then forms a blockchain with the inherit feature that every block can be traced back to the initial first block of the chain. This also implies that any later modification or deletion of single transactions or entire blocks would result in a hash mismatch in hash pointers and Merkle trees and therefore break the chain.

A blockchain network can be private, where access and read/write permissions can be restricted, or public with unrestricted access and read/write permissions. Although the most popular applications of public blockchains to date are cryptocurrencies, the technology is by far not limited to this use case (Davidson, 2016).

The proposed public blockchain database approach seems to be most suitable for blockchain-backed analytics due to one of its core characteristics: the immutability of its records.

2.1. Immutability of the blockchain

The data stored in a blockchain database is immutable in the sense that once a record has been written, it cannot be modified or deleted afterwards. This can be put down to the process of validating transactions and adding them to a new block, which is commonly referred to as "mining".

Among others, there are two popular categories of mining algorithms in public blockchains: Proof-of-work (PoW), as applied with Bitcoin mining (Nakamoto, 2008) and Proof-of-stake (PoS), as proposed for the cryptocurrency Ethereum (Buterin, 2017).

These categories of mining algorithms both provide consensus among the distributed parties of the blockchain about the validity of the transactions and therefore, the final commit to the database. Whereas each set of algorithms contains advocates and attack vectors[1], both sets also share characteristics which makes it almost impossible to determine the consensus process for the purpose of fraud or self-interest and therefrom derived ensure the immutability of existing blockchain records.

Considering that a broad distribution of mining instances is crucial to reduce the risk of manipulation, it is recommended to use a large (i.e. widespread distributed) public blockchain for blockchain-backed analytics. Alternatively a private or permissioned blockchain can be applied; in particular for big consortiums that aim to retain control over configuration parameters of the blockchain, e.g., to reduce transaction costs (Davidson, 2016).

2.2. Blockchain database capabilities

Despite its database structure, a distributed blockchain database is not primarily intended to be used as traditional database storage, mostly due to matters of the distributed technical design and mining process. Notably with the traditional Bitcoin blockchain, there are a number of known scalability limitations, such as the limited number of transactions per block, the limited throughput of transactions and the high latency until a transaction is confirmed (Croman et al., 2016). In addition, classical blockchains are usually not capable of any traditional querying capabilities (as opposed to RDBMS or NoSQL data stores) and in most cases only allow the lookup of existing – and thus valid – transactions.

For this reason, public blockchain databases mostly serve as distributed ledgers (especially for cryptocurrencies) with the property of providing synchronized, auditable and verifiable transactional data across multiple users and distributed networks (without the need of the involvement of third parties to validate transactions). They are not designed as data storage.

3. Blockchain-backed analytics

The idea of blockchain-backed analytics consists of creating an immutable linkage between the three core components of an analytics project: data, model and result.

The data component can be any kind of data that has either been used to train (i.e. to build) a model, or to apply a model. The model component can be any kind of data science model

[1] In order to manipulate the PoW consensus, the computing power (i.e. the hash rate) of a fraudulent participant needs to exceed 51% of the hash rate of the overall network. To fix a PoS consensus, a complex randomized process has to be determined. Hence, the probability of exactly hitting one of these attack vectors is negatively correlated with the size of the network and will tend to theoretically zero in large blockchain networks (Buterin, 2017).

represented as a function, script, library, binary executable, containerized application or even as virtual machine image; whereas the format of the result is determined by the model.

3.1 Blockchain signatures of components

Each component will be registered as a secure cryptographically hashed fingerprint as a transaction property to a public blockchain database, together with a pointer to the transaction identifier containing the component it continues from. The registration process consists of two steps:

1. Creation of the fingerprint (i.e. a secure cryptographical hash) of the component
2. Signature of the transaction to a public blockchain, consisting of:
   a. The hash of the component ($\mathbf{h_x}$)
   b. The transaction identifier ($\mathbf{t_x}$) of the linked component (optional for the data component)

The fingerprint should be created with a hash function in compliance with the Advanced Encryption Standard (AES)[2] and have a key length of no less than 128-Bit. The transaction properties can be submitted as a hexadecimal string, or ideally in JavaScript Object Notation (JSON) format (Fig.1.).

Figure 1: Transaction properties

```
{
 "properties":
  { "data": [{"name": "data ", "hash": "hx(data)"}],
    "model": [{"name": "model", "hash": " hx(model)", "data": "tx(data)"}],
    "result": [{"name": "result", "hash": " hx(result)", "model": "tx(model)"}]
  }
}
```

In short, this approach stores the linked chain of analytical components with an immutable public blockchain transaction, whereas each component can be always identified by its unique hash and transaction identifier. Records of the relationship of projects, hashes and transactions have to be kept separately.

3.2 Component linkage verification

In order to track a component, the blockchain can be queried by a given transaction or wallet identifier for a specific transaction that includes either data, model or result

[2] Advanced Encryption Standard (AES), Retrieved May 12th, 2018, from: https://nvlpubs.nist.gov/nistpubs/FIPS/NIST.FIPS.197.pdf; (doi:10.6028/NIST.FIPS.197. 197).

information as transaction data. The data-model-result relationship can then be traced by the linkage of components that is being reflected in the result's transaction properties.

Furthermore, such a verification procedure could also simply be used to ensure the source integrity of data, models and results on an individual basis; by retrieving the component's signature and verifying the fingerprints against the fingerprints of the original or linked component. Notably, filtering queries (e.g. finding all datasets a specific model has been applied with) are in general not possible without parsing the entire blockchain.

3.3 Blockchain ecosystem

With the increasing popularity of cryptocurrencies, a vast set of blockchain implementations have emerged, with Bitcoin being the first in 2009. The blockchain technology best suited for a specific analytics project depends on individual requirements such as payload size, block time, transaction fees and the public availability of the blockchain. As a ledger for information verification, almost any blockchain technology that allows querying transactions and including transaction properties is principally applicable for blockchain-backed analytics.

Overall we can recommend the Ethereum blockchain as an ecosystem for blockchain-backed analytics. Component hashes can be either stored as raw transactional data (i.e. transaction property) in hexadecimal format, or alternatively integrated into a "smart contract", a programmatic feature of the Ethereum blockchain. Furthermore, as it is one of the largest public blockchain transaction networks worldwide, a widespread distribution of Ethereum nodes is guaranteed.

3.4 Costs analysis

Using a public blockchain network always involves costs to process a transaction, i.e. a fee must be paid before a transaction can be processed and validated.

With respect to the Ethereum ecosystem, the total fee for a single transaction adds up the base transaction price (currently 21000 "gas") and the costs for additional payload (currently 4 gas for a zero byte, 68 gas for a non-zero byte).[3] Considering that an AES 256-bit hash (equal to 32 bytes) can be expressed as a hexadecimal string with 64 characters, a complete data-model-result linkage documentation requires approximately 1 kilobyte of additional payload in hexadecimal format. In sum, the payload of all three components as additional hex-encoded raw transaction data of three transactions on the Ethereum

[3] Ethereum Homestead Documentation: Estimating transactions costs on the Ethereum blockchain, Retrieved May 12th, 2018, from: http://ethdocs.org/en/latest/contracts-and-transactions/account-types-gas-and-transactions.html.

blockchain resulted in about $0.20 total fees with a fast confirmation time (less than 30 seconds) in May 2018.[4]

In addition to transaction costs, blockchain-backed analytics involves computational costs for hashing the components. Since parallelizing the execution of computing a single hash is not possible, the computational costs of hashing a component only vary with the individual CPU performance; not with the number of physical CPUs or cores.

Our own performance tests with two popular secure cryptographic hashing algorithms (BLAKE2 & SHA-256) using commodity hardware have shown that large components with even a one terabyte file size can be hashed under 30 minutes and smaller components with sizes of up to one gigabyte within just a few seconds (Table 1.).

Table 1: Computational costs (time) for hashing different file sizes

CPU: Intel(R) Xeon(R) CPU E5-2698 v4 @ 2.20GHz

Algorithm	**1 MB**	**10 MB**	**100 MB**	**1 GB**	**10GB**	**100 GB**	**1 TB**
BLAKE 2	0.004s	0.027s	0.181s	1.481s	15.584s	2m 25s	~25m
SHA-256	0.015s	0.110s	0.622s	6.204s	63.167s	10m 7s	~90m

Source: Own performance tests (2018).

4. Discussion

Whilst we believe that blockchain-backed analytics is a scalable and easy-to-use approach to ensure trusted analytics, there are several considerations which should be made.

For example, additional transaction costs for registering components in a public blockchain are not insignificant, although they are very low for single transactions. But it should be noted that – especially in the field of artificial intelligence – self-learning and self-evolving machine learning or deep learning models need to be tracked at every step of the model evolution process. However choosing the right blockchain technology (e.g. with optimal block size and transaction costs) for the specific project requirements and consolidating multiple components into a single transaction can help to optimize the costs of a blockchain-backed analytics project.

[4] Own registration of a result-component on the Ethereum blockchain on May 12th, 2018: https://etherscan.io/tx/0xd2749d1bcd7983769ba4801265c65fce8e92df7476f57df01bffcb148e5f0b32.

In theory, our approach is scalable to any kind of data size in terms of scalability and usability for big data applications, but in practice it is limited to the costs of hashing the components.

As described, the computational costs for hashing a component is not considered to be a significant overhead for component sizes up to a few gigabyte, but they have to be taken into account for big data. However in many big data environments data is mostly stored in distributed file systems, such as the Hadoop Distributed File System (HDFS), where the identification of distributed chunks of data is being achieved by inherit file checksum mechanisms that are already been applied during the data ingestion process.[5]

When applying blockchain-backed analytics with data or results stored in HDFS, the component does not need to be hashed again, because the available block checksums could be re-used as distinct block hashes in order to create a Merkle tree of all relevant blocks.

A similar approach also applies for containerized applications with inherit hashing mechanism, such as Docker images, where a fingerprint of the image is automatically created during the image build process.[6] Consequently, it is possible to easily integrate parts or entire analytical ecosystems in the format of a Docker image digest as distinct model component into blockchain-backed analytics.

Following our approach, where the data itself is not stored on the blockchain, an additional overhead process of maintaining a documentation of the relationship of projects, hashes and transactions has to be taken into account. However recent database solution developments with blockchain capabilities (e.g. decentralization and immutability) on top of traditional database capabilities (e.g. querying, indexing, search), could ease the adoption of blockchain-backed analytics, due to the omission of additional hashing procedures and documentation in off-chain references.[7] A similar ease of use could also apply for current developments of distributed (file) system solutions that are directly attached to a blockchain.[8]

[5] Hadoop Checksum, Retrieved May 12th, 2018, from: https://hadoop.apache.org/docs/current/hadoop-project-dist/hadoop-common/FileSystemShell.html#checksum.

[6] Docker Engine Reference, Docker images digests, Retrieved May 12th, 2018, from: https://docs.docker.com/engine/reference/commandline/images/#list-image-digests.

[7] e.g. solution “BigChainDB”, Retrieved May 12th, 2018, from: https://www.bigchaindb.com.

[8] e.g. solution “The Interplanetary File System”, Retrieved May 12th, 2018, from: https://ipfs.io.

With respect to continuous improvements in the development and integration of blockchain-based technologies, we are confident that our generic proposal of tracing the data-model-result linkage of analytical projects can be easily extended to broader ecosystems, such as continuous integration systems as part of the application lifecycle management.

References

Buterin, V. (2017). Proof of Stake FAQ., *Ethereum Wiki*., Retrieved May 12th, 2018, from: *https://github.com/ethereum/wiki/wiki/Proof-of-Stake-FAQ*.

Croman K. et al. (2016). On Scaling Decentralized Blockchains., *In: Clark J., Meiklejohn S., Ryan P., Wallach D., Brenner M., Rohloff K. (eds) Financial Cryptography and Data Security. FC 2016. Lecture Notes in Computer Science, vol 9604.* Springer, Berlin, Heidelberg.

Davidson, S., et al (2016). Economics of Blockchain., Retrieved May 12th, 2018 from: *https://dx.doi.org/10.2139/ssrn.2744751*.

Nakamoto, S. (2008). Bitcoin: A Peer-to-Peer Electronic Cash System., Retrieved May 12th, 2018, from: *http://bitcoin.org/bitcoin.pdf*.

Zyskind, G., et al (2015). Decentralizing Privacy: Using Blockchain to Protect Personal Data., *2015 IEEE Security and Privacy Workshops*, San Jose, CA, pp. 180-184.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8295

Algorithmic Trading Systems Based on Google Trends

Gómez-Martínez, Raúl; Prado-Román, Camilo; De la Orden de la Cruz, María del Carmen
Departamento de Economía de la Empresa, Universidad Rey Juan Carlos de Madrid

Abstract

In this paper we analyze five big data algorithmic trading systems based on artificial intelligence models that uses as predictors stats from Google Trends of dozens of financial terms. The systems were trained using monthly data from 2004 to 2017 and have been tested in a prospective way from January 2017 to February 2018. The performance of this systems shows that Google Trends is a good metric for global Investors' Mood. Systems for Ibex and Eurostoxx are not profitable but Dow Jones, S&P 500 and Nasdaq systems has been profitable using long and short positions during the period studied. This evidence opens a new field for the investigation of trading systems based on big data instead of Chartism.

Keywords: *Big data, behavioral finance, investors' mood, artificial intelligence, Bayesian network, Google Trends.*

1. Introduction

Algorithmic trading systems invest in financial markets in an unattended and constant way, sending buy and sell orders to the market for a financial instrument, according to a complex mathematical algorithm. Most of the trading systems that are operating nowadays follows Chartism rules, but the irruption of big data in asset management has opened a new approach for algorithmic trading.

There are numerous studies that demonstrate that investor mood is affected by multiple factors, changes over time and may be conditioned by experience or training (Cohen and Kudryavtsev, 2012). These changes in mood provide evidence of anomalies in the behavior of stock markets (Nofsinguer, 2005). Corredor, Ferrer and Santamaría (2013) claim that investor mood has a significant effect on stock performance.

We find that weather affect to the stock market returns (Hirshleifer and Shumway, 2003, Jacobsen and Marquering, 2008) as sunny climates are associated with an optimistic mood and then positive returns. Seasonal patterns like vacations that implies the effect of "sell in May and go away" or the "Halloween" effect (Bouman and Jacobsen, 2002; Marshall 2010) means that securities market yield should be greater from November to April than from May to October. Even the Moon (Yuan, Zheng and Zhu, 2006) implies different returns according to the different phases of the moon observing differences from 3% to 5% in yield from one phase to another.

The sports results are another item that modifies investors mood. Edmans, García and Norli (2007) studied the results of football, cricket, rugby and basketball and others have focused on the NFL (Chang, Chen, Chou and Lin, 2012), football (Berument, Ceylan and Gozpinar, 2006; Kaplanski and Levy, 2010) and on cricket (Mishra and Smyth, 2010). Gómez and Prado (2014) performed a statistical analysis of the following stock markets session return after national team football matches. The results obtained show that after a defeat of the national team, we should expect negative and lower than average prices on the country's stock market, the opposite occurring in the case of a victory.

At this stage, if investor mood varies and affects financial markets and their liquidity (Liu, 2015), the challenge that arises is how to measure mood to predict market trend (Hilton, 2001) which leads us to consider a Big Data approach:

Wu et al. (2013) use big data to predict market volatility, Moat et al. (2013) use the frequency of use of Wikipedia to determine investor feelings, whereas Gómez (2013) elaborated a "Risk Aversion Index" based on the stats of Google Trends for certain economic and financial terms that relate to market growth. Through an econometric model, he shows that Google Trends provide relevant information on the growth of financial markets and may generate investment signs that can be used to predict the growth of major

European stock markets. According to this approach, we could create an algorithmic trading system that issues buy and sell orders by measuring the level of aversion to risk, if an increase in tolerance towards risk implies a bull market and an increase in aversion to risk a bear market.

In this paper we will describe Big Data trading algorithmic systems that, instead of Chartism rules, use Artificial Intelligence (AI) models based on Google Trends to predict de evolution of main world stock index.

2. Methodology and Hypothesis

The following statistics are mainly used to measure the perform of an algorithmic trading system (Leshik and Crall, 2011):

- Profit/Loss: the total amount generated by the system from its transactions over a certain period.
- Success rate: Percentage of successful transactions out of the total transactions, if if the percentage is above 50%, the system is profitable and the higher the percentage, the better the system.
- Profit Factor: this rate shows the relationship between earnings and losses, by dividing total earnings by total losses. A rate higher than 1 implies positive returns and the higher the rate, the better.
- Sharpe Ratio: relates profitability to volatility, the higher the ratio, the better the performance of the system (Sharpe, 1994).

InvestMood[1] developed in January 2017 trading algorithmic systems for the following index: Ibex 35, Eurostoxx 50, Dow Jones, S&P 500 and Nasdaq.

According to Gomez (2013), the volume of searches registered in Google on financial terms has explanatory capacity and predictive on the evolution of the markets. Since 2004 in which Google Trends began to publish these statistics, it is observed that bearish markets imply high level of searches of terms such as crash, recession or short selling, while bull markets imply low levels of this searches. Bearing this in mind, InvestMood have created big data algorithmic trading systems that open long or short positions following an artificial intelligence model in which the predictors are Google Trends stats while the target variable is the next evolution of those index (up / down).

The process of the algorithm is the following one: Every first day of the month these systems trains the artificial intelligence models, using a monthly sample of Google Trends

[1] For mor information visit: http://www.investmood.com/

of dozens of economic-financial terms, and issue a prediction for next month's trend. Google Tends consults have been limited to financial matters and don not have any restriction by localization. The system maintains a long or short open position until there is a new prediction in the opposite direction.

From this point, the hypothesis to study is the following one:

H1: A big data algorithmic trading system based on artificial intelligence models over investors' mood can generate positive returns.

We will validate this hypothesis if we reach three evidences:

1. Profit/Loss amount is positive including license costs and trading commissions.
2. Success rate is higher than 50%
3. Profit factor is higher than 1

3. Data

Goggle Trends[2] has historic data available from 2004 in a monthly base. As the first models was trained on 2017, January 1st the first models were trained using 156 observations.

The prospective analysis of this paper starts in January 2017 and ends in February 2018, so we have 14 months for the study and therefor 14 different models, each one for one month.

All the quotes and stats used in this study has been provided by Trading Motion. Trading Motion[3] is a Fintech who allows users of 23 brokers all over the world to operate in an unattended way using trading algorithmic systems developed by 74 professional developers. All these developers follow Chartism rules except InvestMood, that using the Rey Juan Carlos University has developed its systems using IA on Investors' Mood.

Trading systems studied in this paper are running on Trading Motion form January 2017. After three months testing the systems they were available for the clients for April 2017.

4. Results

The URL available for the stats of the systems concerning this study are:

Ibex 35: https://www.tradingmotion.com/explore/System/PerformanceSheet?Id=17652

Esx 50: https://www.tradingmotion.com/explore/System/PerformanceSheet?Id=17705

[2] Visit: https://trends.google.com/trends/

[3] For moro information visit: https://www.tradingmotion.com/

DJ: https://www.tradingmotion.com/explore/System/PerformanceSheet?Id=17651

S&P 500: https://www.tradingmotion.com/explore/System/PerformanceSheet?Id=17654

Nasdaq: https://www.tradingmotion.com/explore/System/PerformanceSheet?Id=17653

The models created for the trading systems has been trained using the algorithms of dVelox, a data mining tool developed by IT firm Apara[4]. These algorithms build a Bayesian Network (Bayes, 1763) like the following one, used for the Nasdaq model trained in 2018 February 1st:

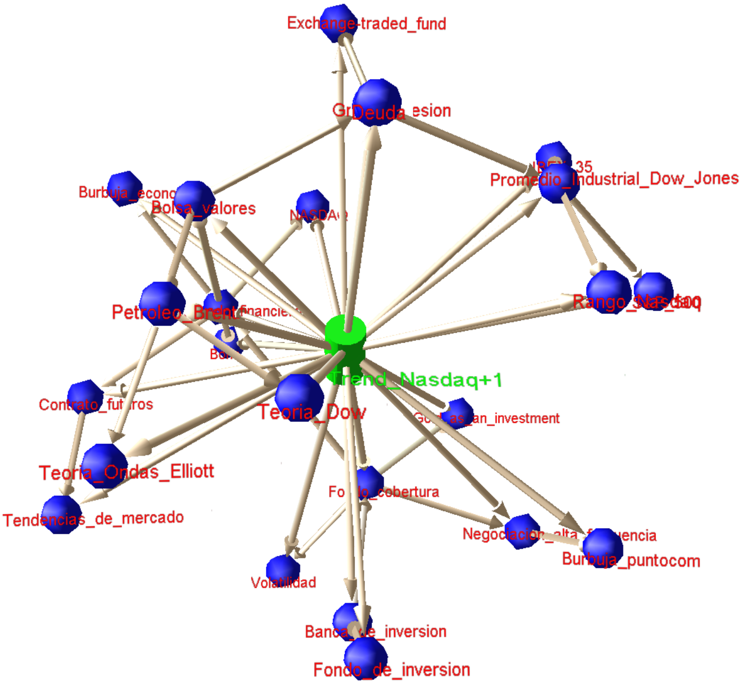


Figure 1. Bayesian Network for Nasdaq trading system. Source: dVelox (2018).

Table 1 sums up the performance of each one of the five trading systems that have been running form January 2017 to February 2018. In this table we observe that systems for Ibex and Eurostoxx are not profitable, so we cannot validate de H1 hypothesis for these indexes. Notwithstanding, the systems created for Dow Jones, S&P 500, and Nasdaq has been profitable, they have a success rate higher than 50% and a profit factor higher than 1, so we can validate H1 for this American indexes.

[4] For more information visit: http://www.apara.es/es/

Table 1. Performance of Big Data trading algorithmic systems on Investors' Mood

Index	Profit/Loss	Success rate	Profit Factor	Sharpe Ratio
Ibex 35	-824,00 €	47,80%	0,99	-0,64
Eursotoxx 50	-560,00 €	50,50%	1,00	-0,12
Dow Jones	16.919,00 €	58,20%	1,36	1,52
S&P 500	20.803,00 €	60,70%	1,44	1,86
Nasdaq	34.628,00 €	62.40%	1,54	2,43

Source: Trading Motion (2018)

5. Conclusions

In this study, we used an innovative approach to check the capability of the behavioral finance and the Investors' Mood to predict the evolution of the financial markets. The study is based on big data and uses artificial intelligence to predict the evolution of Ibex 35, Eurostoxx 50, Dow Jones, S&P 500 and Nasdaq indexes. We can check that these "pure investors' sentiment" systems can be profitable for the American indexes while the result is poor for European ones.

First conclusion is that Google Trends is a good investors' sentiment metric for the American indexes studied, closer than global sentiment if Google Trends has been no limited by location. The poor results for Ibex or Eursotoxx suggests a limitation in Google Trends stats for this models in further investigation.

Second conclusion from this study is that trading systems can be developed using an alternative approach to common systems based on technical analysis. This study has shown how the trading system for Dow Jones, S&P 500 and Nasdaq, based on the predictions of an artificial intelligence model that uses investors' mood from Google Trends to predict is capable to generate positive returns in a long/short strategy.

All this opens an interesting field of research in the development of algorithmic trading.

References

Bayes, T. (1763). An Essay towards solving a Problem in the Doctrine of Chances. Philosophical Transactions of the Royal Society of London 53: 370-418. doi:10.1098/rstl.1763.0053.

Berument, H., Ceylan, N. B., y Gozpinar, E. (2006). Performance of soccer on the stock market: Evidence from turkey. The Social Science Journal, 43(4), 695-699. doi:10.1016/j.soscij.2006.08.021

Bouman, S., y Jacobsen, B. (2002). The Halloween indicator, "sell in may y go away": Another puzzle. The American Economic Review, 92(5), 1618-1635.

Chang, S., Chen, S., Chou, R. K., y Lin, Y. (2012). Local sports sentiment y returns of locally headquartered stocks: A firm-level analysis. Journal of Empirical Finance, 19(3), 309-318. doi:10.1016/j.jempfin.2011.12.005

Cohen, G. y Kudryavtsev, A., (2012). Investor Rationality y Financial Decisions. Journal of Behavioral Finance, 13(1), 11-16.

Corredor, P., Ferrer, E. y Santamaría, R. (2013): El sentimiento del inversor y las rentabilidades de las acciones. El caso español. Spanish Journal of Finance y Accounting, 42 (158), 211-237

Edmans, A., García, D., y Norli, Ø. (2007). Sports sentiment y stock returns. The Journal of Finance, 62(4), 1967-1998.

Gómez Martínez, R. (2013). Señales de inversión basadas en un índice de aversión al riesgo. Investigaciones Europeas De Dirección y Economía De La Empresa, 19(3), 147-157. doi:http://dx.doi.org/10.1016/j.iedee.2012.12.001

Gómez Martínez, R., y Prado Román, C. (2014). Sentimiento del inversor, selecciones nacionales de fútbol y su influencia sobre sus índices nacionales. Revista Europea De Dirección y Economía De La Empresa, 23(3), 99-114. doi:http://dx.doi.org/10.1016/j.redee.2014.02.001

Hilton, D.J., (2001). The Psychology of Financial Decision-Making: Applications to Trading, Dealing, y Investment Analysis. Journal of Psychology y Financial Markets, 2(1), 37-53.

Hirshleifer, D., y Shumway, T. (2003). Good day sunshine: Stock returns y the weather. The Journal of Finance, 58(3), 1009-1032. Retrieved from http://www.jstor.org/stable/3094570

Jacobsen, B., y Marquering, W. (2008). Is it the weather? Journal of Banking y Finance, 32(4), 526-540. doi:http://dx.doi.org/10.1016/j.jbankfin.2007.08.004

Kaplanski, G., y Levy, H. (2010). Exploitable predictable irrationality: The FIFA world cup effect on the U.S. stock market. The Journal of Financial y Quantitative Analysis, 45(2), 535-553. Retrieved from http://www.jstor.org/stable/27801494

Leshik, E. y Crall, J., (2011). An Introduction to Algorithmic Trading: Basic to Advanced Strategies. Wiley.

Liu, S., (2015). Investor Sentiment y Stock Market Liquidity. Journal of Behavioral Finance, 16(1), pp. 51-67.

Marshall, P. S. (2010). In Kaynak E. H.,TD (Ed.), Sell in may y go away? probably still good investment advice!. HUMMELSTOWN; PO BOX 216, HUMMELSTOWN, PA 17036 USA: INT MANAGEMENT DEVELOPMENT ASSOCIATION-IMDA.

Mishra, V., y Smyth, R. (2010). An examination of the impact of India's performance in one-day cricket internationals on the Indian stock market. Pacific-Basin Finance Journal, 18(3), 319-334. doi:10.1016/j.pacfin.2010.02.005

Moat, H., Cure, C., Abakan, A., Kennett, D. Y., Stanley, H. E., y Pries, T. (2013). Quantifying Wikipedia usage patterns before stock market moves. Scientific Reports, Retrieved fromhttp://www.nature.com/srep/2013/130508/srep01801/pdf/srep01801.pdf

Narayan, S. y Narayan, P.K., (2017). Are Oil Price News Headlines Statistically y Economically Significant for Investors? Journal of Behavioral Finance, 18(3), pp. 258-270.

Nofsinguer, J.R., (2005). Social Mood y Financial Economics. Journal of Behavioral Finance, 6(3), pp. 144-160.

Sharpe, W., F., (1994) The Sharpe ratio properly used it can improve investment management, J Portf Manag, 21, pp. 49–58

Yuan, K., Zheng, L., y Zhu, Q. (2006). Are investors moonstruck? lunar phases y stock returns. Journal of Empirical Finance, 13(1), 1-23. doi:http://dx.doi.org/10.1016/j.jempfin.2005.06.001

Wu, Kesheng and Bethel, Wes and Gu, Ming and Leinweber, David and Ruebel, Oliver, A Big Data Approach to Analyzing Market Volatility (June 5, 2013). Algorithmic Finance (2013), 2:3-4, 241-267. Available at SSRN: https://ssrn.com/abstract=2274991 or http://dx.doi.org/10.2139/ssrn.2274991

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8301

How to sort out uncategorisable documents for interpretive social science? On limits of currently employed text mining techniques

Philipps, Axel
Institute of Sociology & Leibniz Center for Science and Society (LCSS), Leibniz University Hanover, Germany.

Abstract

Current text mining applications statistically work on the basis of linguistic models and theories and certain parameter settings. This enables researchers to classify, group and rank a large textual corpus – a useful feature for scholars who study all forms of written text. However, these underlying conditions differ in respect to the way how interpretively-oriented social scientists approach textual data. They aim to understand the meaning of text by heuristically using known categorisations, concepts and other formal methods. More importantly, they are primarily interested in documents that are incomprehensible with our current knowledge because these documents offer a chance to formulate new empirically-grounded typifications, hypotheses, and theories. In this paper, therefore, I propose for a text mining technique with different aims and procedures. It includes a shift away from methods of grouping and clustering the whole text corpus to a process that sorts out uncategorisable documents. Such an approach will be demonstrated using a simple example. While more elaborate text mining techniques might become tools for more complex tasks, the given example just presents the essence of a possible working principle. As such, it supports social inquiries that search for and examine unfamiliar patterns and regularities.

Keywords: *text mining; interpretive social science; qualitative research; standardised and non-standardised methods; social science.*

1. Introduction

Before starting to answer the title of the paper, the exact nature of text mining needs to be identified. Text mining is a combination of statistical and linguistic approaches of text analysis that has lately gained attention in the field of digital humanities. An important forerunner was the Italian literary scholar Franco Moretti (2007) with his concept of "distant reading". He proposed that scholars who are used to employing in-depth interpretations (close reading) are unable to read and study the ever-increasing amount of data that is produced worldwide. Because of this, he recommends a different approach. In contrast to printed books, Moretti accesses digitally-accessible texts and identifies patterns in large corpora. This kind of distant reading includes a growing number of visualisations such as maps, graphs, and trees (Jänicke et al., 2015). Such visualisations usually show relations between such things as actors, names and places; text mining tools, in contrast, concentrate on linguistically small units: words and phrases. Text mining can be defined as a set of "computer-based methods for a semantic analysis of text that help to automatically, or semi-automatically, structure text, particular very large amounts of text" (Heyer, 2009: 2). So, such applications practically count, relate, rank, cluster, and classify single and groups of words in large text corpora and present the outcomes in frequency graphs, word clusters, and networks.

In recent years there has also been a growing interest in text mining for social science research. Various works (i.e. DiMaggio, Nag & Blei, 2013; Marres 2017; Philipps, Zerr & Herder, 2017) present mostly exploratory studies using algorithmic information extraction approaches to demonstrate the power of such tools for text analysis in the social sciences. Proponents of these computer-based methods primarily address qualitatively-oriented social scientists for two reasons (i.e. Evans & Aceves, 2016; Wiedemann, 2013). Firstly, such tools help researchers, who mainly work with textual data, to deal with the increasing number of digitally-accessible texts. Secondly, it is argued that, in a similar way to the grounded theory approach (Glaser & Strauss, 1967), text mining is employed to identify patterns. However, these propositions are slightly misleading. This is a rather unbalanced representation of qualitative and interpretive social research and might explain, to some extent, why (semi)-automatic analysis of textual data has, up to now, been widely ignored in interpretive social sciences (for more details see Philipps, 2018).

This paper therefore primarily takes a closer look at how text mining analyses textual data and in what respect that analysis differs from methods commonly employed by interpretively-oriented social scientists. In this respect, I suggest a different aim and operating procedure for text mining which is more appropriate for interpretive social science. It includes a shift from standardised procedures of classification and clustering of large text corpora to detecting documents that do not fit to applied constructed concepts. To demonstrate this approach, I am presenting an exemplary working principle of low complexity. Later, adapted text mining techniques might become tools for more complex tasks. These seek to support interpretive social science that examine unfamiliar patterns and the regularities of socially-produced meanings.

2. Analysing textual data with text mining and in interpretive social science

Text mining techniques comprise of a wide range of methods from frequency and co-occurrence analysis to sentiment analysis and then to more complex approaches such as topic models and machine learning (Marres 2017; Wiedemann, 2016, 2013). While frequency and co-occurrence counts and identifies the use of words and the relationship between groups of words in large text corpora topic models, machine learning transforms words into numbers and computes statistical interferences in textual data. By no means can these methods be successfully employed to detect thematic shifts or networks of knowledge structures on a trans-textual level in social research studies (i.e. Adam & Roscigno, 2005; Blei & Lafferty, 2006). However, applying text mining requires the setting of some parameters before research is started. For frequency and co-occurrence analyses, for example, researchers need to determine relevant words or groups of words in advance. For a sentiment analysis they have to define classes, ranging from extremely negative to extremely positive. In addition, most machine-learning algorithms demand supervised training (intermediate results are controlled and evaluated by analysts during processing) and even for unsupervised topic models (without interference of external data or human control) researchers have to determine the exact number of clusters to be computed. Hence, current text mining methods have certain characteristics in common; before analysis, researchers define, even to the smallest degree, what is relevant and can potentially be found in textual data. Based on these (standardised) parameter settings, whole text corpora are classified, ranked, or grouped.

However, standardised approaches are, for a great deal of interpretively-oriented social scientists, the opposite to how they were trained. For the most part, they learned and share the basic premise of interpretive social science working with non-standardised methods. This means that a researcher should approach their object of investigation with an open mind and be prepared for surprises. Hence, these researchers seek to situationally understand meanings produced in interactional settings – being ready to overcome previous classifications and schemes. They aim to generate assumptions based on identified content-related, functional and formal aspects of the examined empirical material (for more details see Soeffner, 1999). Nonetheless, while these interpretively-oriented social researchers avoid standardised settings, they employ heuristic models to interpret textual data. They work with commonly-known (scientific) classifications and typifications in order to see how useful this knowledge is for understanding the meaning of given textual data and, at the same time, they search for unfamiliar regularities and patterns. Thus, these researchers translate and describe the world of the observed “into one that we find comprehensible” (Abbott, 2004: 31) and only if they discover so far incomprehensible phenomena do they seek to grasp the underlying working principle and meaning in the form of new but empirically-grounded typifications, hypotheses, and theories.

Against this background, I presume that currently-operating text mining applications for classification and information extraction are often insufficient to be “complementing techniques” (Wiedemann, 2013: no page) for most social scientists with special training in interpretive methods. Under certain circumstances text mining might enable qualitatively-oriented researchers to learn about the variety and development of relevant categories. It is also reasonable to assume that machine learning algorithms which demonstrate knowledge about statistical characteristics of language and text-external knowledge manually coded by analysts (e.g. categories or example sets) will help to retrieve or annotate information in unknown material. However, in all these cases text mining is used to classify and group the entire textual data based on determined parameter settings. We therefore need to think of additional text mining strategies more adjusted to interpretative social science and its basic premise.

3. Adjusting text mining for interpretive social science

Text mining applications might become more relevant for interpretive social science, I suppose, if they enable researchers to divide a large corpus of documents into those with and without comprehensible patterns and components. Such information will stimulate the power of interpretive social inquiry, interpretively explore hidden patterns and unveil unfamiliar meaning. The working principle of such a search strategy might be best described with Max Weber’s (1949) limiting concept of ideal types: “It is a conceptual construct (*Gedankenbild*) which is neither historical reality nor even the ‘true’ reality. It is even less fitted to serve as a schema under which a real situation or action is to be subsumed as one *instance*. It has the significance of a purely ideal *limiting* concept with which the real situation or action is *compared* and surveyed for the explication of certain of its significant components” (Weber, 1949: 93, italics in the original work). Thus, ideal types are not the final outcome of empirical investigations but are used as an heuristic limiting concept to identify the significant aspect of real situations or actions. Practically, if an ideal type has not fully-grasped all aspects of the social phenomena, the researcher will pay full attention to this and mark it for further interpretation. In Weber’s book *Economy and Society* (2013) he, for example, applied ideal types in a “procedure of the ‘imaginary experiment’” (10) comparing a purely rational constructed course of actions with the concrete course of events: “By comparison with this it is possible to understand the ways in which actual action is influenced by irrational factors of all sorts, such as affects and errors, in that they account for the deviation from the line of conduct which would be expected on the hypothesis that the action were purely rational" (Weber 2013: 6). Thus, he intellectually constructs an ideal type of pure rationality to grasp favouring or hindering circumstances which are devoid of subjective meaning "if they cannot be related to action in the role of means or ends" (Weber 2013: 7). Generally speaking, with ideal types as limiting concepts he describes a common strategy among interpretive social scientists to approach their object of investigation in that one employs conceptual constructs to understand social phenomena and by paying attention to unfamiliar regularities and patterns (in Weber's terms: deviations). The latter phenomena are of special interest because their interpretation offers a chance to broaden or even to rewrite established scientific knowledge. However, one has to note that Weber was interested in understanding and explaining social action motivationally. The construction of ideal types thus is not restricted to a rational course of actions.

Applying this search strategy to text mining, a modified variant might become central for interpetative social research working with large digitally-accessible text corpora. In contrast to currently operating mining techniques which classify and group an entire text corpus, an adaptation would use constructed concepts to identify documents which show characteristics assumed in the formulated concept and those that do not fit. Therefore, in contrast to present computer-based applications working with linguistic models and theories, an adjusted text mining technique would operate with preliminary ideas and assumptions, formulated by interpretively-oriented researchers. In particular, for a large corpus of documents the latter will come up with a constructed concept after analysing some selected documents and heuristically employ this to sort out documents that display conceptually anticipated features and relations. In the next step, researchers examine and interpret the specificity of the remaining documents. In this process they might adjust existing concepts or formulate others.

In addition, from the perspective of the humanities one could also say such a modified text mining technique mimics the hermeneutic circle (see Gadamer, 2004). Suggestions formulated in a first round of interpreting textual data are used to identify what is comprehensible and what is not. Incomprehensible textual data will be analysed in further interpretive rounds producing altered or additional suggestions which become the basis for more interpretive sequences. The process will come to an end with working interpretations (constructed concepts) to understand the textual data of interest. Nonetheless, like the hermeneutic circle the process will be impossible to finish as other researchers might find more appropriate readings for understanding certain textual data in the future.

4. An example for sorting out uncategorisable documents

Often interpretively-oriented social scientists work with and interpret a small number of documents. However, sometimes they are confronted with a large corpus of textual data such as an archive of interview transcripts, protocols, letters and other forms of written documents. There are various ways of dealing with such conditions. With Merkens (2004), one might select some documents according to specific characteristics (i.e. relevant for the research goal) and concentrate on these cases or apply the theoretical sampling strategy starting with a few documents and selecting further documents for interpretation based on minimal and maximal contrasts. Theoretical sampling comes to an end if additional analyses of documents reveal no further information. However, there always remain documents that are not interpreted and may contain unexplored patterns and meanings. Under such circumstances an adapted version of text mining technique would offer an opportunity to search these documents for deeper analyses.

In the following paragraphs, I present an instance of low complexity to give an idea of how such an variant of text mining can support interpretively-oriented research projects. It does not involve a reprogrammed text mining application but rather it demonstrates a possible working principle. The case in point is an investigation of applied approaches to promote unconventional ideas in 93 grant proposals sent to a major research-funding organisation in Germany in 2013 (for more detail on method and findings see Philipps, forthcoming). The study started by skimming through the textual data and selecting proposals for deeper analysis. Without any predefined assumptions about specific approaches to unconventional ideas, I began to read a number of grant proposals to get an idea of these. Based on a preliminary impression of the material, I then employed closer readings in a contrastive manner. Using maximal and minimal contrast cases, I searched for specific structural and rhetorical patterns in the rationales of the grant proposals. My interpretation of research proposals continued until typical approaches to unconventional ideas could be identified and separated. After scrutinising 20 proposals and skimming through further applications I came up with a typology of distinct approaches. In an additional and laborious step the typology of identified argumentative patterns was separated into segments and described in a codebook. After a group of interpreters applied segment descriptions to a randomly selected sample of proposals and discussed disagreements and questions, amended codes were used by the author to annotate all 93 research grant proposals. Finally, the manual coding process enabled us to categorise all documents and search for cases with different argumentative patterns or other aspects.

Especially for studies with a greater corpus, automatic text mining would be another option searching for empirically-identified patterns before establishing a codebook and manually annotating the remaining documents. However, such a search strategy requires a limiting concept to sort out documents that show conceptually-suggested patterns and those that do not. In my study, such a concept might, for example, be typical wordings that appear with the identified approaches. Applicants who promoted ideas of solving practical problems typically discussed “drawbacks” or “disadvantages” of earlier solutions and what “benefit” or “advantage” their solution offers in contrast. Concentrating on these wordings, of course, is one-sided and does not fully capture all possible variants and other typical aspects. However, by producing two groups of documents (with and without these certain wordings), one can reduce the number of proposals demanding deeper analysis. In the case of this research project, a simple retrieval of these terms shows that 48 grant proposals used at least one of the terms if not all of them. Combining this result with the already examined proposals (n=20) 33 uncategoriseable documents remain. Hence, this procedure already condenses the number of non-examined documents from 73 down to 33. Apart from applying additional limiting concepts to further reduce the amount of these documents it should be clear that such a search strategy assists interpretively-oriented social scientists to single out documents for further examination.

5. Conclusion

In this paper, I discussed how standardising procedures of current text mining techniques differs in respect of methodological premises commonly employed by interpretively-oriented social scientists. Without question, text mining features such as ranking, grouping or classifying textual data are useful for many research questions in social sciences. However, I presume an adjusted mining techique will greatly support interpretive social science if it shifts from standardised procedures of classification and the clustering of large text corpora to detect documents that do not fit into applied constructed concepts. It is also important to note that such a mining technique would not be based on linguistic theories and information management concepts but on suggestions offered by interpretively-oriented social scientists. As demonstrated at a low level, such an approach can help interpretive social inquiries to single out documents and examine them for unfamiliar patterns and regularities of socially-produced meanings. Nonetheless, as the complex topic of this paper shows it is still a long way from translating the methodological premises of interpretive social sciences into working additional text mining techniques.

References

Abbott, A. (2004). *Methods for Discovery. Heuristics for the Social Sciences*. New York: W. W. Norton & Company, Inc.

Adams, J., & Roscigno, V. J. (2005). White Supremacists, Oppositional Culture and the World Wide Web. *Social Forces,* 84(2), 759–778.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic Topic Models. *Proceedings of the 23rd International Conference on Machine Learning, ACM*, 113–120.

DiMaggio, P., Nag, M., & Blei, D. (2013). Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Art Fundings. *Poetics*, 41(6), 570–606.

Evans, J. A., & Aceves, P. (2016). Machine Translation: Mining Text for Social Theory. *Annual Review of Sociology*, 42, 21–50.

Gadamer, H.-G. (2004). *Truth and Method*. 2nd rev. ed. Trans. J. Weinsheimer & D. G. Marshall. New York: Crossroad.

Glaser, B. G., & Strauss, A. L. (1967). *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Brunswick & London: AldineTransaction.

Heyer, G. (2009). Introduction to TMS 2009. In G. Heyer (Ed.), *Text Mining Services. Building and Applying Text Mining Based Service Infrastructures in Research and Industry. Proceedings of the Conference on Text Mining Services 2009 at Leipzig University* (pp.1–14). Leipzig: LIV.

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R. Borgo, F. Ganovelli & I. Viola (Eds.), *Eurographics Conference on Visualization (EuroVis)-STARs.* The Eurographics Association.

Marres, N. (2017). *Digital Sociology: The Reinvention of Social Research*. Hoboken: John Wiley & Sons.

Merkens, H. (2004). Selection Procedures, Sampling, Case Construction. In U. Flick, E. von Kardoff, & I. Steinke (Eds.), *A Companion to Qualitative Research* (pp. 165–171). London: Sage.

Moretti, F. (2007). *Graphs, Maps, Trees. Abstract Models for Literary History*. London: Verso.

Philipps, A. (forthcoming). Wissenschaftliche Orientierungen. Empirische Rekonstruktionen an einer Ressortforschungseinrichtung. München & Weinheim: Juventa.

Philipps, A. (2018). Text Mining-Verfahren als Herausforderung für die rekonstruktive Sozialforschung. *Sozialer Sinn. Zeitschrift für hermeneutische Forschung*, 19(1): 191–210.

Philipps, A., Zerr, S., & Herder, E. (2017). The Representation of Street Art on Flickr. Studying Reception with Visual Content Analysis. *Visual Studies*, 32(4), 382–393.

Soeffner, H.-G. (1999). Verstehende Soziologie und sozialwissenschaftliche Hermeneutik. In R. Hitzler, J. Reichertz, & N. Schröer (Eds.), *Hermeneutische Wissenssoziologie* (pp. 39-49). Konstanz: UVK.

Weber, M. (2013). *Economy and Society: An Outline of Interpretive Sociology*. (Translation of *Wirtschaft und Gesellschaft*, 4th ed., 1956). Berkeley, Los Angeles, & London: University of California Press.

Weber, M. (1949). Objectivity in Social Science and Social Policy. (Translation of *Die 'Objektivität' sozialwissenschaftlicher Erkenntnis*, 1904) In E. Shils, & H. Finch (Eds.), *The Methodology of the Social Sciences* (pp. 49–112). Glencoe: The Free Press.

Wiedemann, G. (2016). *Text Mining for Qualitative Data Analysis in the Social Sciences: A Study on Democratic Discourse in Germany*. Wiesbaden: Springer.

Wiedemann, G. (2013). Opening up to Big Data: Computer-Assisted Analysis of Textual Data in Social Sciences [54 paragraphs], in: *Forum Qualitative Sozialforschung/Forum Qualitative Research*, 14, Art. 13, http://nbn-resolving.de/urn:nbn:de:0114-fqs1302231.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8302

A proposal to deal with sampling bias in social network big data

Iacus, Stefano Maria [a]; Porro, Giuseppe [b]; Salini, Silvia [a] and Siletti, Elena [a]
[a] Department of Economics, Management and Quantitative Methods, Università degli Studi di Milano, Italy, [b] Department of Law, Economics and Culture, Università degli Studi dell'Insubria, Italy

Abstract

Selection bias is the bias introduced by the non random selection of data, it leads to question whether the sample obtained is representative of the target population. Generally there are different types of selection bias, but when one manages web-surveys or data from social network as Twitter or Facebook, one mostly need to focus with sampling and self-selection bias.

In this work we propose to use official statistics to anchor and remove the sampling bias and unreliability of the estimations, due to the use of social network big data, following a weighting method combined with a small area estimations (SAE) approach.

Keywords: *Big data; Well-being; Social indicators; Sentiment analysis; Self-selection bias; Small area estimation.*

1. Introduction

Despite, social media users could be thought of as the world's largest focus group, and be analysed as such (Hofacker et al., 2016), when one deals with such a data it seems that one cannot to take into account the selection bias. Indeed dealing with Twitter data (as for Iacus et al. (2017)), it is obvious that the sample is done by people that have Internet access, that have decided to open an account on Twitter and that are active users.

In statistical literature, studies largely addresses this bias using the propensity score (PS) approach (Rosenbaum & Rubin, 1983) or the Heckman approach (Heckman, 1979). Both methods attempt to match the self-selected intervention group with a control group that has the same propensity to select the intervention, but they both relay on information that dealing with data such as Twitter data are not disposable. In web-survey context, these issues have been addressed by some strategies based on weighting procedures and model-based approach (Bethlehem & Biffignandi, 2012), nevertheless, also these proposals relies on the availability of unit level information from big data sources, that nowadays are still a mirage, and that, dealing with aggregated big data, are always impossible to achieve.

2. Our proposal

We propose to manage sampling bias, due to the use of aggregated data from social networks, combining a weighting method with a small area estimation model. Our proposal start from this consideration: SAE models have been traditionally used to check and remove unreliability from direct estimations, because if we use direct estimations from Twitter data, those can suffer of selection bias as introduced above, first of all we use a weighting method and then we check and remove their unreliability using SAE models.

In big data context SAE models have been recently used, employing this new kind of data as a covariates when official statistics are missing or they are poor. Porter et al. (2014) use Google trends searches as covariates in a spatial FH model, while in Falorsi et al. (2017) the time series query share extracted always from Google Trends is used as covariate to improve the SAE model estimates for Italian regional young unemployed. Marchetti et al. (2015) use big data on mobility as covariates in a FH model to estimate poverty indicators; where accounting for the presence of measurement error they follow the Ybarra & Lohr (2008) approach. Moreover, themselves have proposed the use official data to verify and remove the self-selection bias due to the use of big data, but in addition to the suggestion, no concrete proposals has been made. Finally, Marchetti et al (2016) use data coming from Twitter as covariate to estimate Italian households' share of food consumption expenditure.

In order to proceed in our direction we have to take into account some topics. When we deal with big data, we often have not a really unit level data to use for direct estimations.

To overcome this problem we can consider different hierarchical levels of aggregations. As an example, we can think of Italian provinces as a unit level for regions. In this way, should be clear that the use of small sample techniques are suitable. Going back to the example: also if we manage million of tweet, if we consider provinces as statistics unit, this number will always be very small. A good and desirable property of big data is the high time frequency, however this feature is often disregarded for the official statistics. In this work, we consider data with the same frequency, but the opportunity to use data with a different time frequencies could be an interesting methodological challenge for the future. Lastly, dealing with timely and spatial information, we should take into account both time and space correlations too. Following these addresses, we now present our method step by step, and we propose an application as toy example.

2.1. The method

About SAE model, we consider area level models, because we assume to have area level covariates. Furthermore, because these data are available for several periods of time *T* and for *D* domains, to consider also eventually time and space correlations, we have chosen a spatio-temporal Fay-Herriot (STFH) model, proposed by Marhuenda et al. (2013). Thus, for domain *d* and *t* time periods, let μ_{dt} be the target parameter, the STFH model, as all the FH models, has two stages, where in first stage, the "sampling model" is defined as follow:

$$\hat{\bar{y}}_{dt}^{DIR} = \mu_{dt} + e_{dt} , \qquad e_{dt} \sim N(0, V(\hat{\bar{y}}_{dt}^{DIR})) , d = 1, \dots, D, t = 1, \dots, T \tag{1}$$

where e_{dt} are the sampling errors that are assumed to be independent and normally distributed, and $V(\hat{\bar{y}}_{dt}^{DIR})$ is the sampling variance of the direct estimator. Especially, we consider as direct estimator the regional sampling mean, weighted by some characteristics to overcome the non-sampling structure of our data

$$\hat{\bar{y}}_{dt}^{DIR} = \frac{1}{\sum_{i=1}^{n_{dt}} w_{idt}} \sum_{i=1}^{n_{dt}} y_{idt} w_{idt} \tag{2}$$

where n_{dt} is the number of provinces in region *d* at time *t*, and w_{idt} are the weights used. For the sampling variance we use the same weights.

While in second stage, the "linking" model is

$$\mu_{dt} = \boldsymbol{x}'_{dt}\beta + u_d + v_{dt} , \qquad u_d \sim N(0, \sigma_1^2) , v_{dt} \sim N(0, \sigma_2^2) \tag{3}$$

it relates all areas through the regression coefficients, $\boldsymbol{x}_{dt}$ is a column vector containing the aggregated values of *k* covariates for the *d*-th area in *t*-th period, and β is the vector of coefficients. u_d are the area effects, that follow a first order spatial autocorrelation, SAR(1), process with variance σ_1^2, spatial autocorrelation parameter ρ_1 and proximity matrix **W** of dimension $d \times d$. Especially, **W** is a row-standardized matrix obtained from an initial proximity matrix $\mathbf{W}^{\mathrm{I}}$, whose diagonal elements are equal to zero and the residual entries equal to one, when the two domains are neighbours, and zero otherwise. Normality for u_d is required for the mean squared error, but not for point estimations. Furthermore v_{dt} represents the area-time random effects that are i.i.d. for each area d, following the first order autoregressive, AR(1), process with autocorrelation parameter ρ_2 and variance parameter equal to σ_2^2.

The final model is defined as $$\hat{\bar{y}}_{dt}^{DIR} = \boldsymbol{x}'_{dt}\beta + u_d + v_{dt} + e_{dt} \qquad (4)$$

Then, $\boldsymbol{\theta} = (\rho_1, \sigma_1^2, \rho_2, \sigma_2^2)$ is the vector of unknown parameters involved in the STFH model. Marhuenda et al (2013) give the empirical best linear unbiased estimator (EBLUE) of β, and the empirical best linear unbiased predictors (EBLUPs) of u_d and v_{dt}. Both are obtained by replacing a consistent $\widehat{\boldsymbol{\theta}}$ in the respectively BLUE and BLUPs introduced by Henderson (1975). Also due to Marhuenda et al (2013) is the parametric bootstrap procedure for the estimation of the mean squared error (MSE) of the EBLUPs, that for B bootstrap replies has the following form $$MSE(\hat{\mu}_{dt}) = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{\mu}_{dt}^{b} - \mu_{dt}^{b}\right)^2 \qquad (5)$$

In this way the point estimates $\hat{\mu}_{dt}$ of μ_{dt} can be supplemented with (5) as measure of uncertainty.

3. Application

The assessment of well-being mostly at local level is an important task for policy makers, because they increasingly need to target their policies and actions not to the nation, but at local domains. Unfortunately very often there is a lack of this level data, even more if the interest is for high frequency data too. To fill this gap, the use of data from social networks can be considered a good option to improve well-being knowledge. In this section considering a well-being index from Twitter data and some official statistics, we implement the proposed approach to check and, if necessary, to remove unreliability of estimations.

3.1. A Subjective Well-being Index with the Twitter Data

Since 2012 Iacus et al. (2015) propose to apply iSA (integrated Sentiment Analysis, Ceron et al. (2016)) method, to derive a composite index of subjective well-being that capture different aspects and dimensions of individual and collective life. This index named Social Well Being Index (SWBI) monitor the subjective well-being expressed by the society through the social networks, epsetially, in Iacus et al. (forthcoming) the SWBI index is provided for the Italian provinces from 2012 to 2016 and combined with the “Il Sole 24 Ore Quality of Life index”. SWBI is not the result of some aggregation of individual well-being measurements, but it directly measures the aggregate composition of the sentiment throughout the society at province or regional level. For this reason, about the weights, we can't consider users characteristics as traditionally, but aggregated to area ones. SWBI has been inspired by the definitions introduced by the think-tank NEF (New Economic Foundation), for its Happy Planet Index (New Economics Foundation, 2012), and it is defined as a manifold, dynamic combination of different features, with indicators which look beyond the single item questions and capture more than simply life satisfaction.

The eight SWBI dimensions concern three different well-being areas: personal well-being, social well-being and well-being at work. Data source are tweets written in Italian language and from Italy and data are accessed through Twitter's public API. A small part of these data (around 1to 5% each day) contain geo-reference information which allows to build the SWBI indicator at province level in Italy.

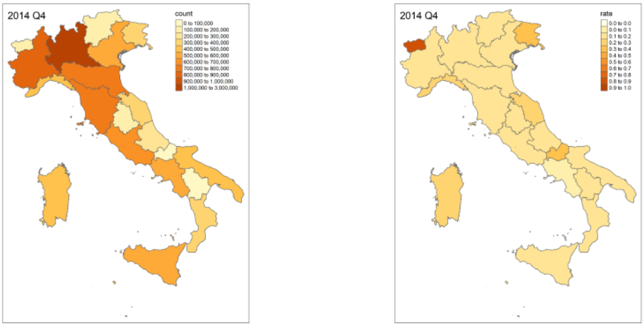


Figure 1. Twitter counts, on the left, and Twitter rates, on the right, in 2014-Q4.

In the application presented here, we consider SWBI quarterly data, with an area aggregation at the Italian provincial and regional levels. Now we shortly describe the dimension of these data. To compute the the SWBI index we consider more than two hundred million tweets (201,496,621) in 22 quarters from 2012 to June 2017. Despite this huge total number, we have to reveal a decrease in the number of tweets for all the four quarters of 2015 and the first quarter of 2016, anyway we stress that also for these quarters the counts were still in the thousands (minimum count equal to 1727 tweets in 2016-Q1 for Basilicata). To have a more realistic view of the situation, we consider a Twitter rate: the ratio between the number of analysed tweets and the number of inhabitants in the region in the same period. Considering the simple counts, we would have seen that most of the SWBI info comes from Lombardia, Piemonte, Emilia and Toscana, while considering also the resident population, we note that information is also substantial in particular or small regions such as Friuli, Sardegna, Valle d'Aosta and Molise, for the last two we remark their large variability during the period too. While the dispersion for big regions like Lazio and Lombardia seems to be smaller. However, we can conclude that during our observational period the average Twitter rate is equal to about 20% of the population, with a mean value always greater than 9% (minimum for Campania), for all regions. Looking at Figure 1, as example, it is clear that, considering the Twitter rate (on the right), all the Italian regions are homogeneously observed with the exception of the Valle d'Aosta which have a higher but almost anomalous rate.

3.2. The implemented model

To implement the proposed model, as toy example, we consider only the `wor` dimension of the SWBI index, at quarterly time level, from 2012-Q1 to 2017-Q2. We compute regional values following (2) and using as weights: the broadband coverage and the Twitter rate. The broadband coverage is provided by “Il Sole 24 Ore” and Infratel Italia, this coverage can be considered as the opportunity to access internet in the different provinces. While the Twitter rate, computed in each period and province level, can be a proxy of the use of Twitter.

The weighted quality of job has remained stable, with very little variability between the regions, the distributions are very compressed, until the second half of 2015. From 2015-Q3 weighted `wor` grows, and especially from the second half of 2016, this dimension attained values greater than 80. Even the differences between the regions are more evident: the distributions are less crushed. We remark that the shapes of the quality of work weighed or not, computed as simple means, are quite similar. It seems that difference for the weighted `wor` index among regions are small, instead are greater in time. Considering the different rankings obtained by the two indices, in the 29% of the cases there are no differences and only for the 14% of the cases, there is a difference (Δ) in the rankings greater than 5 positions. Regions with the greater Δ are Trentino, Campania, Marche, and Sardegna, for the first two we remark the greatest position improvement, while for the last two we remark the greatest worsening of position. Focusing on time, we recorded the major Δ in 2014-Q1, 2015-Q1 and both the considered 2017 quarters. The mean of the ranking Δ is equal to 2.01(SD = 2.41).

Referring to the model (4) we use as direct estimator of regional quality of job the weighted `wor` and its sampling variance. Because in one Italian region, Valle d'Aosta, there is only one province, we decided to drop this region from our data. In the recked STFH model data are available for T = 22 time instants, from the first quarter of 2012 to the second quarter of 2017, and the dominions are the D = 19 considered Italian regions. The considered area level auxiliary variables were, before any process of selection, in the job context: the unemployment and the inactivity rate, computed both in relation to the labour force, as traditionally, and to the resident population; while in the socio demographic context: the birth, the mortality and the natural rate. All the covariates come from official statistics distributed by ISTAT (http://istat.it/), as representatives for all the Italian regions at quarters time level. The row-standardized proximity matrix $\mathbf{W}^{c}$ of dimension 19×19 , has been obtained from an initial proximity matrix $\mathbf{W}^{Ic}$, whose diagonal elements are equal to zero and the residual entries equal to one, when the two regions had some common borders, and zero otherwise. Since in Italy, there are two regions corresponding to two islands, for them we take as neighbours the regions with direct naval connections.

3.3. Results and discussion

After fitting the model the selected covariates were the unemployment rate, calculated traditionally dividing by the labour force and the mortality rate. The coefficients were both negative: regions with larger unemployment and mortality rate have smaller quality of job. The estimated spatial autocorrelation ρ_1 is significant enough with a small and negative value of about -0.02. While the temporal autocorrelation parameter ρ_2 is still significant and greater with a positive value equal about to 0.86. The value equal to zero for $\hat{\sigma}_1^2$ is coherent with analysis of distribution discussed above. Quality of job change in time but less or not at all between regions.

Comparing the resulting EBLUPs obtained by fitting the STFH model and the direct estimates, weighted or not, we can conclude that the direct weighted estimates are approximately design unbiased. Looking at the rankings, what change if we use EBLUPs estimates instead of direct, weighted or not, estimates ? Comparing the rankings obtained with the simple means `wor` and those with EBLUPs estimates, we find that in the 31% of the cases the position is the same and in the 14% of the cases the position Δ is greater than 4. Regions and time situation is the same as when we compared above simple means with weighted means. The mean of the ranking Δ is equal to 1.97 (SD = 2.3). While comparing the rankings obtained with the weighted means `wor` and those with EBLUPs estimates, the situation is very different: in the 88% of the cases the position is the same and in less than 1% of the cases the position Δ is greater than 4. Only in one case we have a great ranking Δ (Marche in 2015-Q3, Δ = 7). The average of the differnces in this case is equal to 0.16 (SD = 0.54). This means that moving to weighted estimates to EBLUPs estimates the ranking are quite the same.

In SAE literature, traditionally use coefficients of variations (CV) to analyse the gain of efficiency of the EBLUPs estimates. National statistical institutes are committed to publish statistics with a minimum level of reliability (<20%). The CVs of our three indices, where those computed for the STFH model have been obtained using boostrap and (5), except few exceptions, are always lower 20%. For 14 regions the CVs are still less than 10%, while the highest CVs values have been obtained only in few quarters for 5 regions: Calabria, Friuli, Lazio, Marche, and Trentino. What is clear is that whenever we observe a peak of CVs, the EBLUPs estimates improve reliability, but also considering only weighted indices this happen. CVs obtained for EBLUPs estimates are quite always lower than both others, but in some cases they are larger, for very small values, than those obtained for weighted estimations, examples for Trentino in 2014-Q4 and 2015-Q1. Thus, EBLUPs based on STFH model are always more reliable than direct simple mean.

More results and discussion will be introduced during the conference presentation.

References

Bethlehem, J. & Biffignandi, S. (2012) *Handbook of Web Surveys*. John Wiley and Sons, Inc., New York, DOI 10.1002/9781118121757

Ceron, A., Curini, L. & Iacus, S.M. (2016) iSA: A fast, scalable and accurate algorithm for sentiment analysis of social media content. *Information Sciences*, 367-368, 105-124.

Falorsi, S., Fasulo, A., Naccarato, A. & Pratesi, M. (2017) Small area model for Italian regional monthly estimates of young unemployed using Google trends data. ISI2017-Marrakech.

Heckman, J.J. (1979) Sample selection bias as a specification error. Econometrica 47(1), 153-161.

Henderson CR (1975) Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2), 423-447.

Hofacker. C.F., Malthouse. E.C. & Sultan, F. (2016). Big data and consumer behavior: imminent opportunities. *Journal of Consumer Marketing*, 33(2), 89-97.

Iacus, S.M., Porro, G., Salini, S. & Siletti, E. (2015) Social networks, happiness and health: from sentiment analysis to a multidimensional indicator of subjective well-being. ArXiv e-prints 1512.01569

Iacus, S.M., Porro, G., Salini, S. & Siletti, E. (2017) How to exploit big data from social networks: a subjective well-being indicator via twitter. In: Petrucci, A. & Verde, R. (eds) *SIS 2017. Statistics and data science: new challenges, new generations*. Proceedings of the Conference of the Italian Statistical Society, Firenze University Press, Firenze, 537-542.

Iacus, S.M., Porro, G., Salini, S. & Siletti, E. (forthcoming) Social networks data and subjective well-being: an innovative measurement for italian provinces. Italian Journal of Regional Studies.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rinzivillo, S., Pappalardo, L. & Gabrielli, L. (2015) Small area model-based estimators using big data sources. Journal of Official Statistics 31(2), 263-281.

Marchetti, S., Giusti, C. & Pratesi, M. (2016) The use of twitter data to improve small area estimates of households' share of food consumption expenditure in italy. AStA Wirtschafts und Sozialstatistisches Archiv 10(2), 79-93.

Marhuenda, Y., Molina, I. & Morales, D. (2013) Small area estimation with spatio-temporal Fay Herriot models. *Computational Statistics & Data Analysis*, 58, 308-325.

Molina, I. & Marhuenda, Y. (2015) sae: An R Package for Small Area Estimation. *The R Journal*, 7(1), 81-98, https://journal.r-project.org/archive/2015/RJ-2015-007/index.html.

New Economics Foundation (2012) *The happy planet index: 2012 report. a global index of sustainable well-being*. Tech. rep.

Porter, A.T., Holan, S.H., Wikle, C.K. & Cressie, N. (2014) Spatial Fay Herriot models for small area estimation with functional covariates. *Spatial Statistics* 10, 27-42.

Rosenbaum, P.R, Rubin, D.B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41-55.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8303

Big data analytics in returns management – Are complex techniques necessary to forecast consumer returns properly?

Asdecker, Björn and Karl, David
Chair of Operations Management and Logistics, University of Bamberg, Germany

Abstract

The more people shop online, the more consumer returns e-tailers face. In order to plan the returns management process capacity adequately, it is necessary to forecast the expected amount of returned parcels. Big data analytics provides a vast number of methods to perform such tasks. However, it should be noted that particularly small- and medium-sized e-tailers lack the capabilities and resources to employ such complex techniques. Against this background, this paper analyses the performance of several data analysis methods that differ in application complexitiy using real data from an apparel e-tailer. On the one hand, we find that –as expected– complex methods outperform simple ones. On the other hand, and from a practitioner's perspective probably even more interesting, we also conclude that a binary logistical regression as the simplest analyzed method may already provide satisfactory results. The findings indicate that the use of big data analytics is of great value to effectively and efficiently manage consumer returns – even if not the most sophisticated state-of-the-art method is used.

Keywords: *returns management; product returns; e-commerce; forecast models.*

1. Introduction and motivation

In today’s retailing world, more people shop online more frequently. Consequently, e-commerce revenues have been skyrocketing in the last two decades and an end to this development is not foreseeable. In 2015, for instance, US consumers spent more than $340.41 billion online; up from $4.98 billion in 1998 (United States Census Bureau, 2018). While e-tailing clearly offers numerous advantages over traditional brick-and-mortar retailing, there is one major disadvantage: supply and demand are geographically separated. Therefore, consumers are unable to see, feel, and test the products before purchase. In other words, they can not try before they buy, which almost inevitably leads to consumer returns.

From a business perspective consumer returns are a major cost driver and pose a serious threat on an e-tailer’s profitability. According to Stock et al. (2006), expenditures can be as high as $30–35 per return, while return rates may well exceed 50 % for fashion items (Asdecker et al., 2017). These already staggering numbers still reflect only part of the problem. In addition to the direct costs there are indirect effects that influence customer value.

Existing research shows that the returns process is part of the post purchase-experience and herein influences customer satisfaction and retention (Petersen & Kumar, 2009). Laseter and Rabinovich (2012) argue that the improvement of the product return experience is based on the following three principles: (1) lower customer efforts to return the product, (2) offer customized solutions that fit the customers’ needs, and (3) exceed customer expectations when processing returns. The third principle specifically refers to the desired outcome (e.g., compensation of made payments), ease of contact, and recovery responsiveness (Mollenkopf et al., 2007). The latter can be operationalized with the time necessary to process a return. To speed returns up, operations require the most accurate capacity planning, which, in turn, is based on forecasts. The importance of this task cannot be underestimated: The better the forecasts are, the more effective and efficient will cosumer returns be processed.

Literature provides various econometric, statistical and/or data mining methods that can be employed to predict returns. Some are higly complex while others are more straightforward and easier to apply. In a world where many decision makers strive for the one optimal alternative, complex state-of-the-art methods seem to be the best choice. However, they also demand sophisticated skills, additional capabilities and financial resources, which are confined, particularly in small- and medium-sized e-tailing companies (Coleman et al., 2016). In a challenging paper, Banks (1993, p. 360) concludes: “I would guess that intelligent use of simple tools will achieve 95 % of the knowledge that could be obtained through more sophisticated techniques, at much smaller cost. Also the simple tools can be applied more quickly to all problems, whereas the complex tools are unlikely to be ubiquitously used." While the paper was written in a different era the general message remains. Against this background, we address the following research-leading question:

- **How are simple data analysis methods performing compared to complex, more sophisticated ones when predicting consumer returns?**

Unlike other publications that try to identify factors that influence the likelihood of consumer returns (e.g., Toktay et al., 2004, Asdecker et al., 2017, Srmiti, 2018), this paper compares the performance of different forecasting methods. To the authors' best knowledge, the publication that is the closest to the one at hand originates from Urbanke et al. (2015). They developed a decision support system that identifies transactions with a high likelihood to return before the actual sale takes place and demonstrated the approach's applicability using a large dataset from a German fashion e-tailer. Within their study they compare seven forecasting techniques, namely the principal component analysis, linear discriminant analysis, randomized truncated singular value decomposition, feature selection based on univariate chi-squared statistic, random projection, non-negative matrix factorization, and a specific feature extraction technique that ignores nominal indicators. While they search for the technique with maximum precision, they do not compare simple with complex approaches.

The remaining article will provide an overview of different forecasting approaches, followed by a report on the performance of the previously introduced techniques. Finally, we conclude with a summary and an outlook on future research.

2. Theoretical background and return forecasting techniques

The data science and statistics literature provides a variety of different methods or techniques that can be used to to predict consumer returns. Since consumers decide to either return or keep the delivered items the dependent variable is binary in nature. This article considers five approaches that are briefly described as follows.

2.1. Binary logistic regression

The binary logistic regression is the simplest method to be taken into account. It is an extension of the linear regression, where the dependent variable is binary (1=return, 0=keep). The independent variables can be either continuous (interval/ratio) or categorial (ordinal/nominal) in nature. For each observation, the binary logistic determines the probability that the dependent variable takes value "1" (Hastie et al., 2009).

2.2. Linear discriminant function analysis

The linear discriminant function analysis, which was first performed by Fisher (1936), shares great similarities with the logistic regression. The analysis requires at least two a priori known groups to which observations shall be assigned. The basic idea is to create a linear combination of independent variables, which best classifies the available data. Thereby, it determines a score for each observation which is then compared to a critical discriminant score in order to carry out the classification (return vs. keep).

2.3. Artifical neuronal network: multiylayer perceptron

Artificial neuronal networks are based on a set of connected nodes, the so called artificial neurons, which are organized in layers. Connected artificial neurons can exchange signals with each other. The receiving artificial neuron will then process it and, in turn, signal artificial neurons connected to it. The ultimate goal is to find a function that best assigns input data to the correct output. To achieve this goal in the returns management context, this study uses multilayer perceptrons, a class of feedforward neural networks. Herein, the information flows exclusively from the input layer through hidden layers with a certain amount of units to the output layer with no feedback flow. Training is done through backpropagation, which is a supervised learning technique that compares the outputs of the network with the known actual values (Hastie et al., 2009).

2.4. Decision tree learning: C5.0 algorithm

Decision trees are hierarchical structures of branches, representing conjunctions of certain characterisics, and leafs, representing class labels. The goal of this technique is to create a decision tree that best classifies the available observation. For this purpose many decision tree algorithms have been presented. This analysis refers to the C5.0 algorithm. C5.0 is the faster, more efficient successor of the widely-employed C4.5 algorithm (Pandya, 2015).

2.5. Ensemble learning technique

The ensemble learning method uses several algorithms to improve predictive performance. It determines the result of every single algorithm and interprets it as a hypothesis for the final verdict (Polikar, 2006). In this study, we used the training data to determine the three most accurate techniques from the following selection: artificial neuronal network, different decision trees (C5.0, QUEST, CART, CHAID), binary logistic regression, linear discriminant analysis, Bayesian network, and nearest neighbor. This resulted in three hypotheses that were given a vote proportional to the confidence, which equals the probability that the postulated hypothesis of a specific algorithm is accurate.

With regard to the required expertise and software, the binary logistic regression and the two-group discriminant analysis are quite straightforward and implemented in common statistical programs with simple and intuitive user interfaces. The remaining three are more complex and therefore require more data mining knowledge as well as less intuitive software (e.g., R, Python, MATLAB, or specific data mining tools such as the IBM SPSS Modeler).

3. Comparison of the different forecasting techniques

The introduced prediction techniques will be tested using real data from a medium-sized German fashion e-tailer that requested confidentiality concerning its name. We will first describe the characteristics of the provided data and then determine the predictive performance of the previously introducted techniques.

3.1. The dataset

The dataset contains shipment and returns information from April 2012 to April 2013. More up-to-date data would be desirable but is unrealistic because many merchants consider returns data as proprietary and are unwilling to share this kind of information. During the analyzed period, the e-tailer sent 220,474 shipments and received 131,907 returns, which equals an α-returns rate of 59.8 % (Asdecker, 2015). While the returns rate might change over time due to different customer behavior or successful avoidance strategies, it should be noted that data age is irrelevant to this type of comparative investigation. The e-tailer shared the following information:

- Package ID: unique ID of the shipment
- Price: Total value of shipped goods in Euro
- Number of articles: total number of articles in the shipment
- Delivery time: time between order placement and delivery in days
- Customer type: categorial variable that indicated the customer type (1=female, 2=male, 3=family, 4=company, 5=unknown)
- Customer age: customer age at the time of order in years
- Account age: age of the customer account at the time of order in days
- Return: binary variable to indicate whether the enclosed return label had been used (1=return, 0=no return)

3.2. Predictive performance

We divided the shared dataset in two parts. The first twelve months, that is, the data from April 2012 to March 2013, were used to derive the prediction model. The last month, April 2013, is the actual test data. Within that period, the e-tailer sent 22,069 shipments and received 13,166 returns. Consequently the α-returns rate is 59.7 %. The criterion used to assess the quality of the forecast is the total absolute error (TAE), which is the summed absolute difference between the predicted and the actual value. This equals the total number of cases that were mispredicted, which means they were either classified as a return even though nothing was returned or vice versa. All five predicting methods were implemented in IBM SPSS Modeler 18, leading to the following results:

- Binary logistic regression, TAE=7,329 (33.21 %)
- Linear discriminant function analysis, TAE=7,372 (33.40 %)
- Artifical neuronal network I (MLP, one hidden layer, six units/layer), TAE=7,364 (33.37 %)
- Artifical neuronal network II (MLP, two hidden layers, six units/layer), TAE=7,146 (32.38 %)
- Decision tree learning: C5.0 algorithm, TAE=6,973 (31.60 %)
- Ensemble learning technique, TAE=6,963 (31.55 %)

Accordingly, the ensemble learning technique provides the best results and predicts 68.45 % of the analyzed cases correctly. The three techniques derived from the training data and used in the ensemble were C5.0, CHAID, and QUEST, which are all decision trees. This finding was expected because the utilization of multiple algorithms within an ensemble usually provides better predictive results than any of the algorithms separately (Polikar, 2006). More surprising is the good performance of the simpler, less sophisticated techniques. Binary logistic regression and linear discriminant function analysis correctly predict 66.79 % and 66.60 % of cases, respectively. This is equivalent to the artificial neuronal network with one hidden layer (66.63 %) and only slightly worse than the neuronal network with two hidden layers (67.62 %).

On the one hand, this generally shows that data mining can be helpful when it comes to plan the return processes and determining the necessary capacity in the returns department. On the other hand, this study also highlights that it does not always have to be the most sophisticated method to generate an acceptable consumer return forecast. In fact, a cost-benefit analysis might favor simple methods, since more sophisticated and complex methods require more resources and data mining knowledge. This holds particulary true for the binary logistic regression, which is not only easy to conduct but also allows for a detailed analysis regarding the factors that affect consumer return behavior. Table 1 summarizes the SPSS report for the statistically significant binary logistic regression ($p=0.000$, Nagelkerke’s $R^2=0.212$) derived from the twelve months training data.

In the model, the probability of a return increases with the total value of the shipped goods, the number of items in a shipment and the account age, whereas it decreases with the delivery time. Packages that are delivered to women have the highest return probability, followed by families, companies, men and unknown recipients. The standardized coefficients b and the Wald statistics show the factor’s relative impact: the biggest effect has the price, followed by the number of items and the customer type.

Table 1. Results of the binary logistic regression model

Variable	b	SE	Exp(b)	Wald	Odds Ratio
Constant*	0.655	0.006	1.926	12,514.985	
Price*	1.267	0.010	3.550	15,552.654	3.550
Nr. of articles*	0.325	0.006	1.384	2,532.507	1.384
Delivery time*	-0.026	0.005	0.975	23.871	0.975
Sent to: Mr.[a]*	-0.437	0.026	0.646	272.223	0.646
Sent to: Family [a]	-0.139	0.086	0.870	2.597	0.870
Sent to: Company[a]	-0.303	0.178	0.739	2.904	0.739
Sent to: Unkown [a]*	-0.554	0.190	0.574	8.525	0.574
Customer age*	-0.065	0.005	0.937	172.983	0.937
Account age*	0.036	0.005	1.036	48.129	1.036

Legend: [a] = reference category: Mrs; * = significant on .05-level;
Nagelkerke’s R^2 = 0.212; all continuous variables standardized

4. Summary and outlook

Overall, the good classification performance of the parametric methods (binary logistic regression and linear discriminant analysis) surprises. In fact, their performance is only 1.66/1.85 percentage points worse than the ensemble technique as the best nonparametric method. However, their results are easy to interpret and understand. Therefore, simple models such as the binary logistic regression might be the better choice in business practice, especially for small and medium-sized e-tailers that face limited data mining capabilities and financial resources. This holds particularly true because they can also be used for the initiation of preventive returns management measures.

This study is based on real data provided by a German e-tailer. Nevertheless, the scope of the analyzed data was very limited. It would be desirable if future studies had access to additional information, e.g., a customer’s order and return history, shopping basket composition, to substantiate the presented results. With a larger amount of data, it is very likely that the investigated complex techniques can better exploit their advantages. Moreover, this analysis focused on the prediction of return shipments which is important to plan the reverse logistics process. In further research, it may be of interest to take a closer look inside the shipments to extend the analysis to single articles/items.

References

Asdecker, B. (2015). Returning mail-order goods: analyzing the relationship between the rate of returns and the associated costs. *Logistics Research* 8(3), 1–12.

Asdecker, B., Karl, D., & Sucky, E. (2017). Examining Drivers of Consumer Returns in E-Tailing with Real Shop Data. *Proceedings of the 50th Annual Hawaii International Conference on System Sciences (HICSS)*, 4192–4201.

Banks, D. (1993). Is Industrial Statistics Out of Control? *Statistical Science* 8(4), 356–377.

Coleman, S., Göb, R., Manco, G., Pievatolo, A., Tort-Martorelle, X., & Reis, M. S. (2016). How Can SMEs Benefit from Big Data? Challenges and a Path Forward. *Quality and Reliability Engineering International* 32(6), 2151–2164.

Fisher, R. A. (1936). The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics* 7(2), 179–188.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning. Data Mining, Inference, and Prediction* (2nd ed.). New York: Springer.

Laseter, T. M., & Rabinovich, E. (2012). *Internet Retail Operations – Integrating Theory and Practice for Managers.* Boca Raton: CRC Press, Taylor & Francis Group.

Mollenkopf, D. A., Rabinovich, E., Laseter, T. M., & Boyer, K. K. (2007). Managing Internet Product Returns: A Focus on Effective Service Operations. *Decision Sciences* 38(2), 215–250.

Petersen, J. A., & Kumar, V. (2009). Are Product Returns a Necessary Evil? Antecedents and Consequences. *Jounal of Marketing* 73(3), 35–51.

Pandya, R. (2015). C5.0 Algorithm to Improved Decision Tree with Feature Selection and Reduced Error Pruning. *International Journal of Computer Applications* 117(16), 18–21.

Polikar, R. (2006). Ensemble based systems in decision making. *IEEE Circuits and Systems Magazine* 6(3), 21–45.

Srmiti, K. (2018). Predicting Online Returns. In Kumar, A. & S. Saurav (Eds.), *Supply Chain Management Strategies and Risk Assessment in Retail Environments* (pp. 181–194). Hershey: IGI Global.

Stock, J., Speh, T., & Shear, H. (2006). Managing product returns for competitive advantage. *MIT Sloan Management Review* 48(1), 57–62.

Toktay, L. B., van der Laan, E. A., & de Brito, M. P. (2004). Managing Product Returns: The Role of Forecasting. In Dekker, R., Fleischmann, M., Inderfurth, K. & L. N. Van Wassenhove (Eds.), *Reverse Logistics* (pp. 45–64). Berlin: Springer.

United States Census Bureau (2018). U.S. Retail Trade Sales - Total and E-commerce (1998-2015). Retrieved February 22, 2018, from http://www2.census.gov/retail/releases/current/arts/ecommerce.xls.

Urbanke, P., Kranz, J., & Kolbe, L. (2015). Predicting Product Returns in E-Commerce: The Contribution of Mahalanobis Feature Extraction. *Proceedings of the 36th International Conference on Information Systems (ICIS)*, 1–12.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8305

Facebook, digital campaign and Italian general election 2018. A focus on the disintermediation process activated by the web

Calò, Ernesto Dario; Faggiano, Maria Paola; Gallo, Raffaella; Mongiardo, Melissa
Department of Communication and Social Research (CoRiS), Sapienza University, Rome, Italy

Abstract

The digital media convergence has innovated all the information and communication channels. ICTs have adapted to the reticular structure, which can be interpreted as a paradigm of today's Network Society, invoking a reorganization of man-machine relationship, interpersonal interactions and the ways of collecting, processing and storing data. Political communication and its processes of participation are not exempt from this evident change. The aim of this paper is to investigate the transformation dynamics that have taken place through the Internet and social networks, as effectively more democratic tools that are able to stimulate a bottom-up and grassroots participation, in some respects a disintermediate participation compared to the typical unidirectionality of analogue media. In identifying the specific characteristics of the interaction spaces offered by the web, we have examined the digital campaign strategy devised by the political parties during the Italian general election (4th March 2018). The results have returned an unprecedented scenario, in which the social media strategy plays a leading role in the different dimensions of political marketing. The web is proposed to become the ideal social arena for the meeting between political offer and demand.

Keywords: *Internet Data; Networked Politics; Digital Campaign; Political Marketing; Social Media Strategy; Italian General Election.*

1. Introduction

Last 4th March 2018 general election was held in Italy. Political candidates have planned complex political marketing strategies and voters have been exposed to a communication river, resulting from an intense campaign. The Internet has hosted most of the communication flows: the structure made of "nodes" and "bridges" of the *Network Society* (Castells, 1996) has poured online, fueling a universe of exceptional relationships. The possibility of Internet users to share interaction spaces, beyond physical proximity, is a peculiar characteristic and the main advantage of this digital *medium*. The web, during its evolutionary path, has abandoned the rigidity of a *top-down* communication to enable new *bottom-up* communication flows of *peer-to-peer* interaction, also in the political sphere and in civic engagement.

The objective of this paper is to study the peculiarities of this electoral campaign, observing the role of the Internet and, more specifically, of Facebook, the social network that most of all has been the preferred platform for the dissemination of targeted messages (top-down), for their circulation and for the active user response (bottom-up), as "consumers of the political offer". Moreover, in considering the overall media scenario, we asked ourselves whether there has been a permanent disintermediation between political offer and demand, where political actors have proposed to bypass traditional mediation to enable a direct channel with the individual citizen through the web. Lastly, it is appropriate to investigate the political marketing and web marketing activities, with specific reference to social media marketing, carried out during the election campaign to reach a discouraged "voter-consumer", tired of the promises not kept[1] and eager to participate actively in the dialogue with the "seller", the same necessary dialogue that occurs today between firms and consumers in the field of products and services businesses.

2. Electoral campaign, Internet and the networked politics

2.1. Towards a new structural change

The crisis of political participation in the most advanced democracies (also visible from the phenomenon of abstentionism) has been observed by a wide range of theoretical and empirical research (Dalton & Wattenberg, 2000; Kriesi et al., 2013; Dalton, 2014). The results of these studies have highlighted the problems related to the mediatization of

[1] According to Eurobarometer, in fact, Italian parties collect the trust of just 9% of citizens. The government and parliament achieve a 15% confidence rate on the national population (Standard Eurobarometer 86, November 2016).

politics. Press and television have helped to create distances[2], increasing the level of distrust of the political class. The *media-system* has not fulfilled the functions of *watchdog*, instead it contributed to a phenomenon known as the spectacularization of politics, in which the media are co-authors of the definition of the *agenda-setting* (McCombs & Shaw, 1972). In an attempt to recover at least part of the lost ground, it was necessary to restore those relational spaces typical of direct contact. With the advent of the Internet, political actors have stimulated new forms of communication, based on the principles of political marketing, relational marketing and social media marketing; while voters have experienced a new process of post-media individualization, with partial replacement of the old and obsolete mediation mechanism (Chaffee & Metzger, 2001). The gradual evolutionary process of digital modernization has invested at the same time the political organizations, the information system and the electorate (Norris, 2000).

At the origins of this innovative model is the absolute success of «Obama for America», the first and biggest digital campaign organization ever (Johnson, 2009). It has proved its effectiveness through the wider mobilization network in history: on the election day, on 4th November 2008, about 25% of Obama's voters were connected to him through his online networks (Masket, 2009; Plouffe, 2010). In the wake of this success, Trump and Clinton moved towards it in 2016, while, in Europe, we witnessed the accomplishments of the Spanish movements *Ciudadanos* and *Podemos*, in 2015 and 2016, of the French movement *En Marche!*, for Macron 2017, and the English one, for Corbyn 2017. In Italy, in 2013, the *Movimento 5 Stelle* (M5S), started by Beppe Grillo, was the forerunner of this new way of meaning the political organization. An inclusive political organization, that nurtures itself from below, promoting unprecedented forms of disintermediation and participation. But in 2013, only the M5S had a complete and effective digital strategy. This strategy was, in some respects, necessary, because, unlike the other parties, they had very limited visibility within the mainstream analogue media contexts (TV and press). Their success in 2013, thanks to the web, has shown, even in Italy, that it was necessary to reevaluate the old mechanisms of *broadcasting* communication, in favor of a participatory model guided by digital technologies. The infrastructure and organizational architecture of politics should have moved online, in the networked politics, where *individualized*, post-bureaucratic and participatory activism (Cepernich, 2017) resides. Moreover, the peculiarity of the Italian general election 2018 is identifiable in the legislative intervention that has abolished public funding for parties. This has encouraged the transition to an online digital campaign, taking advantage of the (relative) gratuitousness of the web spaces and of the new engagement and

[2] According to an Eurobarometer survey on trust in the media-system, 52% of Italians have confidence in media information. This is in line with the European average (53%) but well below the Scandinavian countries (Finland 88%, Denmark 77% , Sweden 77%), the Netherlands and Portugal (73%) or Austria (72%). (Special Eurobarometer 452, Media pluralism and democracy, November 2016).

fundraising initiatives. In this way the primacy of public opinion has been established, especially that which is formed online, where the *hubs* (super-nodes) of the network, in the role of *influencers* (Weimann, 1991), move the consensus as new *opinion leaders* in a two-step flow of communication (Lazarsfeld et al., 1944) on the web. Political marketing, therefore, has equipped itself with data scientists, demoscopic surveys, data-driven strategies, news management, etc. The Internet increasingly becomes the favorite place for political communication, a place inhabited by a voter who is no longer a passive spectator, but an active-citizen and a *digital-militant* (Novelli, 2018).

2.2. Applied social tools

The Facebook profile is an extremely widespread digital tool, which refers to a strong interpersonal and disintermediate component. The existence of *avatars* within a peer user network is the most obvious expression of a process of personalization of the media space. The "logged in" voter enters the world of *mass self-communication* (Castells, 2009), where the contents are increasingly seen in a subjective and self-selective key. This is what determines the *filter bubbles* of the self-referential information (Pariser, 2011), as the effects of a semi-automatic selection of contents, individually or collectively, as a consequence of the hyper-selective exposure of subjects and of the functioning of the algorithmic logic in the search results. The individual becomes at the same time a producer, a distributor and a selective consumer of contents, in his personal space that he takes care of at will. From a marketing point of view, social networks are a precious reservoir of *big data*, from which to draw to set up the digital campaign. Profiling tools and techniques prove extremely useful not only in commercial campaigns but also in electoral campaigns. The feedback provided by Facebook, as well as by other social networks, allows to quantify a significant part of the response of the communication’s receiver, returning some measures on the effectiveness of the strategies. Moreover, in confirming the importance of online communication, we refer to the Audiweb data on internet diffusion in Italy: from the June 2017 survey it emerged that Italians who claim to have at least one device with the possibility to access the Internet are 43 million, about 90% of the population between 11 and 74 years.

3. Research methodology

The effectiveness of digital communication tools is also found in the field of empirical social research. The huge amount of aggregated data offered by the web reduces the problems inherent to the detection and sampling techniques, as well as favoring dialogue and compatibility between different softwares, by virtue of the common use of computer language. In observing the new forms of online political interaction and communication, we gathered all the textual, graphic and audiovisual material (it was almost always a mixed material) produced by the official Facebook pages of the main parties that make up the current tripolar political scenario: *Movimento 5 Stelle*, *PD* (Democratic Party) and *Lega*. The time frames in which the data have been collected concern the first and fourth (last) week of the electoral campaign (exactly from 5th to 11th February and from 26th February to 3rd March). A total of 1,397 Facebook posts were studied. Lastly, we have constructed and analyzed through a matrix a large series of variables, which have been useful to return important information for the observation of the phenomenon, after considering the actual electoral results and what we set out to investigate.

4. Findings

In light of the results of the general election, we focused on the three main political parties: indeed, the M5S obtained 32.9% of the votes, followed by the PD (18.8%) and the Lega (17.4%)[3]. Data about the subscriptions to the Facebook spaces (both parties and candidates profiles) and the traffic of messages directed through these channels (table 1.), have confirmed the importance of a campaign strategy specifically devised for the world wide web. The parties that have concentrated massive resources on the Internet have increased their visibility, beyond a mere "showcase website". The participation modalities of Facebook users differ in intensity. In fact, among these differences it creeps the phenomenon named by Morozov (2011) "*couch activism*", which is a level of "gaunt" involvement, that does not necessarily translate into concrete activism.

Table 1. Number of Facebook subscribers/number of posts produced by the parties (weeks 1-4)

Facebook pages/profiles	Subscribers	Total posts
Lega – Salvini Premier	395,541	1,022
Partito Democratico	259,069	98
Movimento 5 Stelle	1,284,671	277
Matteo Salvini[a]	2,162,960	
Matteo Renzi[a]	1,125,786	
Luigi Di Maio[a]	1,560,451	

a: Facebook personal profiles. Source: Facebook.com (5th March 2018), our elaboration.

[3] It is worth pointing out, however, that in considering the agreements done by the center-left coalitions (PD, *Più Europa*, *Civica Popolare*, *Italia Europa Insieme*) and those done by the center-right (Lega, *Forza Italia*, *Fratelli d'Italia*, *Noi con l'Italia - UDC*), the scenario changes considerably: the center-left has reached a total of 23%, while the center-right has obtained 37%.

Regarding the process of political *leaderization*, the processed data (table 2.) show that in the Lega the leader's communication prevails over that one of the party. The messages of the Lega are oriented to the construction and narration of the public image of the leader, Salvini. They help reduce the distances between voters and candidate. The PD has used a balanced strategy between the leader and the party's image. The M5S, instead, has privileged the importance of the party in relation to the candidate, Di Maio, confirming to be a deliberately inclusive movement, with a process of disintermediation that, before the Facebook channel, passes from *Il Blog delle stelle* (ex Beppe Grillo's Blog) and from the *Rousseau* platform, as two virtual spaces of direct interaction..

Table 2. ***Leaderization*** **strategies per party**

Party/leader	*Continuum* party-leader (0-100)[a]
Lega/Matteo Salvini	86
Partito Democratico/Matteo Renzi	48
Movimento 5 Stelle/Luigi Di Maio	37

a: the constructed index summarizes 10 items regarding the communication of the leader and the party. The value 0 represents the complete centrality of the party, while the value 100 represents the maximum process of *leaderization*. Our results are in line with those presented by *Ipsos* and *TWIG* about Facebook and Twitter messages. Source: our elaboration plus *Ipsos* http://www.ipsos.com/it-it/elezioni-politiche-2018-analisi-del-voto.

In observing the data on the topics faced by the parties through Facebook, it is clear that the Lega (Figure 1) has characterized its communication on the security issue, with specific reference to the government of immigration. The main communication strategy of the Salvini's party has relied on *fear arousing appeals* (Fabris, 1997), namely those messages that aim to impress the public through the exposure to negative contents, in an implicit attempt to spread worry and subsequently propose a solution to the problem.

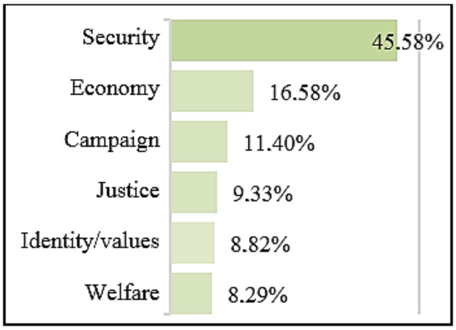


Figure 1. Main topics in the speeches of the Lega.

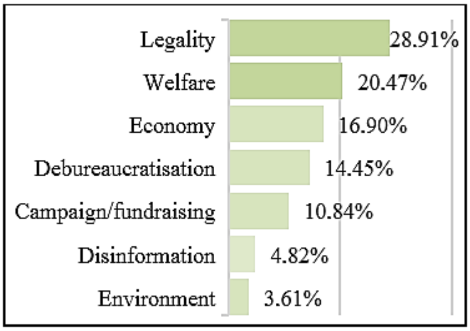


Figure 2. Main topics in the speeches of the M5S.

The M5S (Figure 2), instead, proposes itself as a new reforming force, capable of combating corruption, lack of transparency and the high costs of politics. Their target constituency is largely made up of the weakest sections of the population, for whom massive welfare measures have been proposed (the "*citizenship income*" is the most relevant and well-known example among their proposals). Lastly, the PD (Figure 3) has tried to leverage the typical values of the Progressive Left, underlining past political achievements and the need to ensure continuity in the Country's path.

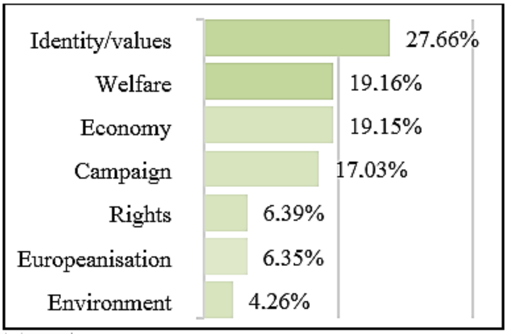


Figure 3. Main topics in the speeches of the PD.

5. Conclusions

The online digital campaign has proven to be of great strategic importance for the parties. The exponential spread of web messages has amplified the echo of communications. As visible from the data, M5S and Lega have used the internet to create an image coherent with the electorates' expectations, setting a digital campaign (thanks to the use of profiling techniques and algorithms) based on the urgencies perceived by voters. In some cases the annulment of the distances between the candidate and the voter has turned in an almost *one-to-one communication*. The PD, instead, despite having proclaimed the values of a "progressive" Left, has demonstrated to be maliciously bound with the "old" media-system, missing the appointment with a fruitful dialogue with his targets. Their conspicuous loss of consensus, as ruling party before elections, passes from this inadequate online digital strategy. In conclusion, the Internet has certainly renewed the ways through which political offer meets political demand, providing an interactive and partially disintermediate exposure space available to candidates and voters, in order to return useful two-way feedback for both actors involved. In fact, as Mazzoleni argued (1998), political marketing must "favor the adaptation of a candidate to his/her potential electorate, make him/her known to the largest number of voters, create differences compared to the opponents, optimize the number of votes that can be earned during the campaign".

References

Bongrand, M. (1993). Le marketing politique. Paris: PUF.

Castells, M. (1996). The Rise of a Network Society. Oxford (UK): Blackwell.

Castells, M. (2009). Communication Power. New York: Oxford University Press.

Chaffee, S.H. & Metzger, M.J. (2001). The End of Mass Communication?. *Mass Communication & Society*, 4, 365-379.

Cepernich, C. (2017). Le campagne elttorali al tempo della networked politics. Bari-Roma: Laterza Editore.

Dalton, R.J. (2014). Citizen Politics. Public Opinion and Political Parties in Advanced Industrial Democracies. Thousand Oaks (CA): CQ Press-Sage.

Dalton, R.J. & Wattenberg, M.P. (2000). Parties without Partisans. Political Change in Advanced Industrial Democracies. New York: Oxford University Press.

Fabris, G. (1997). La pubblicità. Teoria e prassi. Milano: FrancoAngeli.

Johnson, D.W. (2009). Campaigning for President 2008. Hoboken: Taylor & Francis.

Kriesi, H., Lavenex, S., Esser, F., Matthes, J., Bühlmann, M., Bochsler, D. (2013). Democracy in the Age of Globalization and Mediatisation. Basingstoke: P. Macmillan.

Lazarsfeld, P.F., Berelson, B., Gaudet, H. (1944). The People’s Choice. How the Voter Makes Up His Mind in a Presidential Campaign. New York: Columbia University Press.

Masket, S. (2009). Did Obama’s Ground Game Matter? The Influence of Local Field Offices during the 2008 Presidential Election, *Public Opinion Quarterly*, 73, 1023-1039.

Mazzoleni, G. (1998). La comunicazione politica. Bologna: Il Mulino, 138-148.

McCombs, M. & Shaw, D. (1972). The Agenda-Setting Function of the Mass Media. *Public Opinion Quarterly*, 36, 176-187.

Morozov, E. (2011). The Net Delusion. The Dark Side of Internet Freedom. New York: Public Affairs.

Mosca, L. & Vaccari, C. (2011). Nuovi media, nuova politica? Partecipazione e mobilitazione online da MoveOn al Movimento 5 Stelle. Milano: Franco Angeli.

Norris, P. (2000). A Virtuous Circle. Political Communications in Postindustrial Societies. Cambridge: Cambridge University Press.

Novelli, E. (2018). Le campagne elettorali in Italia. Protagonisti, strumenti, teorie. Bari-Roma: Laterza Editore.

Pariser, E. (2011). The Filter Bubble. How the New Personalized Web is Changing What We Read and What We Think. New York: The Penguin Press.

Plouffe, D. (2010). The Audacity to Win: How Obama Won and How We Can Beat the Party of Limbaugh, Beck, and Palin. New York: Penguin Books.

Weimann, G. (1991). The Influentials: Back to the Concept of Opinion Leader?, *Public Opinion Quarterly*, 55, 267-279.

http://www.audiweb.it/news/total-digital-audience-giugno-2017 (Retrieved 03-05-2018).

http://ec.europa.eu/digital-single-market/en/news/media-pluralism-and-democracy-special-eurobarometer-452 (Retrieved 03-05-2018)

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8310

Inferring Social-Demographics of Travellers based on Smart Card Data

Zhang, Yang and Cheng, Tao
SpaceTimeLab, Department of Civil, Environmental and Geomatic Engineering, University College London, UK

Abstract

With the wide application of the smart card technology in public transit system, traveller's daily travel behaviours can be possibly obtained. This study devotes to investigating the pattern of individual mobility patterns and its relationship with social-demographics. We first extract travel features from the raw smart card data, including spatial, temporal and travel mode features, which capture the travel variability of travellers. Then, travel features are fed to various supervised machine learning models to predict individual's demographic attributes, such as age group, gender, income level and car ownership. Finally, a case study based on London's Oyster Card data is presented and results show it is a promising opportunity for demographic study based on people's mobility behaviour.*

Keywords: *social-demographics; smart card data; travel variability.*

1. Introduction

Recently, social-demographic prediction based on individual's behaviour has become an emerging topic both in industry and academia. However, researchers mainly concentrate on people's online behaviour, such as web browsing (Hu, Zeng, Li, Niu, & Chen, 2007; Saste, Bedekar, & Kosamkar, 2017) and social network (Rao, Yarowsky, Shreevats, & Gupta, 2010; Vijayaraghavan, Vosoughi, & Roy, 2017). The discriminative power of people's mobility in the physical world has been overlooked, especially the travel behaviour via public transit.

Nowadays, public transit (PT) system plays a significant role in people's daily life. The modern PT network has widely equipped with the automatic ticketing system, called Automated Fare Collection (AFC) systems. As the popularity of AFC system, big data collecting through AFC records provide an opportunity to reveal hidden travel patterns by segmenting users to improve public transportation service quality and provide information for customers (Goulet Langlois, Koutsopoulos, & Zhao, 2016; J. Zhao, Qu, Zhang, Xu, & Liu, 2017). Traffic smart card has amassed a large amount of data to profile users' travel pattern, including travel mode, travel routes and time, and so much more. However, it lacks the social-demographic attributes of passengers to further explore 'who are the card carriers' and 'why they behave differently', which are crucial to better understand the users' requirement and travel patterns, offering a full picture of travel in urban area, which can help government make better transport planning, supply passengers with more personalized PT services and enhance the PT experience.

In this work, we devote to developing a framework for social-demographic inference based on smart card data (SCD). We first establish a feature extraction process to profile passengers' travel behaviours by using SCD. Then, several supervised machine learning algorithms are adopted to infer individual social-demographics, such as age level, gender, income level and car ownership. This study can also help us better understand the relationships between passengers' travel behaviours and social-demographics.

This paper is organized as follows. Section 2 briefly reviews the related works. Section 3 illustrates the framework and methodologies. Then, a case study based on London's Oyster Card dataset is carried out in Section 4 and results are presented and discussed. Finally, we summarise the conclusions, limitations, and future work in Section 5.

2. Related Works

A considerable number of existing works have demonstrated the impact of demographic factors on passengers' travel patterns. Early corresponding studies (Hanson & Hanson, 1981) suggest that the attribute of the individual or household are instrumental in shaping daily travel decisions. As suggested in the recent literature (Yang et al., 2017), generally, an in-depth understanding of mobility patterns can be obtained by clustering people into distinct groups according to their demographic characteristics.

With regard to the analysis of the association between travel patterns and social-demographics, some works statistically describe the social-demographic features among diverse travel patterns (Goulet Langlois et al., 2016; Ortega-Tong, 2013), others illuminated the travel behaviour across social-demographic groups (Shobeiri Nejad, Sipe, & Burke, 2013). In recent year, some researchers have sought to illustrate the linkages between travel patterns and social-demographic variables on specific groups of travellers. For example, Siren and Hakamies-Blomqvist (2004) examine the association between selected demographics variables and mobility of elderly citizens in Finland. Results show that older persons experience reduced mobility mainly for leisure-related trips and their mobility was strongly associated with driving behaviour, education level and home location. van den Berg, Arentze, and Timmermans (2013) analysed the relationship between socio-demographic and social activity-travel patterns.

Most of the existing studies focus on the qualitative analysis of the relationship between the social-demographics and the individual travel behaviours. Alternatively, social-demographic inference or prediction has attracted increasing attention in the big data era. Social-demographic inference aims to characterise travellers, which can aid in transport planning, land use improvement and business settlement. To the best of our knowledge, there is no existing literature investigating the predictability of traffic smart card data for passengers' social-demographics inference. To fill this research gap, in this paper, we further study on what extent travellers' social-demographics can be inferred from their PT transaction records.

3. Dataset

3.1. London's Oyster Card Data

The dataset used in this study is a compilation of Oyster Card transaction records in London, UK, during the full year of 2013. There are two types of SCD, one from the tube system and the other from the bus system. Each transaction is recorded automatically when a passenger taps in/out at a tube station or boards at a bus stop. Summarily, the entire dataset contains around 2.18 million journeys made by 9708 passengers, made up of 33.7% tube journeys and 66.3% bus journeys. Each transaction record contains the following fields: (1) unique ID, (2) boarding time, (3) alighting time (tube journey only), (4) boarding station, (5) alighting station (tube station only), (6) journey mode (bus or tube).

3.2. London Travel Survey Data

Transport for London (TfL) carried out the London Travel Demand Survey (LTDS), a continuous household survey of the London area, covering all London boroughs and the City of London. The LTDS is conducted based on the household for collecting individual or household demographic, social-economic and travel-related information. Around 8000 randomly selected households undertake the LTDS annually. All household members aged 5 and over need to complete the questionnaire. The unique Oyster card ID voluntarily provided by interviewed individuals in households for linking LTDS to Oyster card transaction records. The social-demographics data used in this study are provided in Table 1.

Table 1. Demographic attributes and corresponding categories

Attribute	Num. of labels	Categories and fraction
Age	3	Young (<30): 20.79% Adults (30 – 65): 58.04% Elder (>65): 21.17%
Gender	2	Male: 42.95% Female: 57.05%
Car ownership	3	Have no cars: 44.98% Have one car: 40.40% Have more than one car: 14.62%
Income level	3	Low income: 31.26% Middle income: 39.82% High income: 29.93%

4. Framework and Methodologies

The framework of social-demographic prediction is shown in Figure 1. The framework includes four main steps: (1) raw SCD preprocessing, (2) travel feature extraction, (3) social-demographic prediction, (4) performance evaluation. Details are given below.

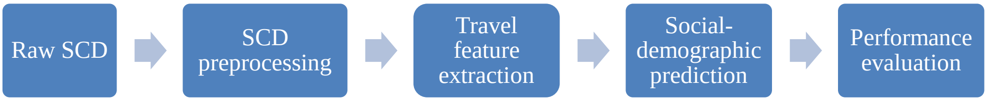


Figure 1. Methodology framework

4.1. Data preprocessing

The first step in the analysis was linking the two datasets by using smart card ID. After that, We take 6354 frequent users' SCD as the primary sample. To deal with the missing alighting time and station of the bus journey, we refer to the method proposed by Jinhua Zhao, Rahbee, and Wilson (2007).

4.2. Feature extraction

A key issue in passenger segmentation is to construct accurate and comprehensive passenger profiles from SCD. In this study, various travel features are defined as to calibrate passenger profiles in order to differentiate the individual travel patterns. All features are categorized into four types, related to temporal variability (When), spatial variability (Where) and travel mode preference (How), respectively. The feature extraction process and explanation has been demonstrated in our previous work (Zhang & Cheng, 2017). Here, we just simply list the features generated from SCD in Table *2*.

Table 2. Travel features for passenger profiling

Subgroup	Feature	Description
Temporal features	AFTI	The average first start time on weekdays
	LFTI	The average last start time on weekdays
	MPT_NUM	the number of trips during morning peak (7:00am-10:00am)
	EPT_ NUM	the number of trips during evening peak (4:00pm-7:00pm)
	AVG_TRIP	The average number of trips per day
	ACTI_DAY	Active days in the whole year
	ACTI_DUR	Active duration in the whole year
Spatial features	AVG_ TIME	The average of tube trip time
	VAR_ TIME	The variance of tube trip time
	MAX_TD	The average radius travelled by tube per day
	AVG_TS	The average of the number of different tube stations used per day
	VAR_TS	The variance of the number of different tube stations used per day
	AVG_BL	The average of the number of different bus lines used per day
	VAR_BL	The variance of the number of different bus lines used per day
	AVG_INNER	The mean value of the inner zone number
	AVG_OUTER	The mean value of the outer zone number
Mode preference	TUBE_NUM	The total number of the tube journeys
	BUS_NUM	The total number of the bus journey
	MODE_T	How often a passenger changes the transport mode per day? (average)

4.3. Demographic prediction

After generating individual’s travel features, with the social-demographics as the ground truth data, we utilize several popular supervised machine learning approaches to predict demographics. First, logistic regression (LR) is a natural choice for this type of a task. We also adopt random forest (RF), naïve Bayesian (NB) and multi-layer perceptron (MLP). We formulate the gender inference problem as a binary classification task and the others as the multi-class classification problems. The details of all these algorithms will not be discussed here.

5. Experiment Results

We report the social-demographic attributes inference for different classification methods, containing LR, RF, NB and MLP. The performance of prediction is evaluated by Accuracy (*Acc*), Precision (*Prec*), Recall (*Rec*) and F1 value (*F*1) (Qin et al., 2017). We conduct a 5-fold cross-validation and calculate the four performance metrics. Performance comparison results are shown in Table 3. Results show the best prediction accuracy of ‘Age group’, ‘Gender’, ‘Income level’ and ‘Car ownership’ can achieve 66.68%, 61.33%, 55.76% and 61.28%, respectively. Obviously, the prediction results can achieve a relatively high accuracy, but in some tasks, the scores of *Prec* and *Rec* are not very satisfied, since the class-imbalance problem has not been considered in these standard models.

Table 3. social-demographic inference performance comparison by four evaluation metrics

Model	Age				Gender			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
RF	0.6331	0.6317	0.6730	0.6661	**0.6133**	0.6242	0.5832	0.6030
LR	0.6647	0.5645	0.5598	0.6325	0.5950	0.5526	0.5630	0.5797
NB	0.4578	0.4432	0.5292	0.4915	0.5703	0.5660	0.5686	0.5677
MLP	**0.6668**	0.5743	0.5614	0.5936	0.6029	0.5456	0.5642	0.5936
Model	Income level				Car ownership			
	Accuracy	Precision	Recall	F1	Accuracy	Precision	Recall	F1
RF	0.5444	0.5502	0.5478	0.5536	0.6033	0.5753	0.6241	0.6439
LR	**0.5576**	0.5630	0.5633	0.5638	0.6001	0.4882	0.4924	0.5716
NB	0.5035	0.4683	0.5435	0.5133	0.3942	0.3808	0.4666	0.4346
MLP	0.5535	0.5593	0.5609	0.5586	**0.6128**	0.4825	0.4940	0.5864

6. Conclusion and Discussion

This study mainly explores the possibility of using smart card data to predict an individual's social-demographics. This framework helps to obtain people's social-demographic data without traditional travel surveys. The results presented in this paper can be baselines for further research.

However, there is still large room for improvement. First, for many demographic inference tasks, the dataset is category-imbalanced. More advanced techniques should be used to solve the class-imbalance classification problem. Second, some inference tasks are correlated. For example, the average income level of the elder is usually lower than that of the middle age people. Thus, it is necessary to investigate the correlations and use the multi-task learning method in the modelling to improve the prediction accuracy.

References

Goulet Langlois, G., Koutsopoulos, H. N., & Zhao, J. (2016). Inferring patterns in the multi-week activity sequences of public transport users. *Transportation Research Part C: Emerging Technologies, 64*, 1-16. doi:http://dx.doi.org/10.1016/j.trc.2015.12.012

Hanson, S., & Hanson, P. (1981). The Travel-Activity Patterns of Urban Residents: Dimensions and Relationships to Sociodemographic Characteristics. *Economic Geography, 57*(4), 332-347. doi:10.2307/144213

Hu, J., Zeng, H.-J., Li, H., Niu, C., & Chen, Z. (2007). *Demographic prediction based on user's browsing behavior*. Paper presented at the Proceedings of the 16th international conference on World Wide Web.

Ortega-Tong, M. A. (2013). *Classification of London's public transport users using smart card data*. Massachusetts Institute of Technology.

Qin, Z., Wang, Y., Cheng, H., Zhou, Y., Sheng, Z., & Leung, V. (2017). Demographic Information Prediction: A Portrait of Smartphone Application Users. *IEEE Transactions on Emerging Topics in Computing, PP*(99), 1-1. doi:10.1109/TETC.2016.2570603

Rao, D., Yarowsky, D., Shreevats, A., & Gupta, M. (2010). *Classifying latent user attributes in twitter*. Paper presented at the Proceedings of the 2nd international workshop on Search and mining user-generated contents.

Saste, A., Bedekar, M., & Kosamkar, P. (2017, 10-11 Feb. 2017). *Predicting demographic attributes from web usage: Purpose and methodologies*. Paper presented at the 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC).

Shobeiri Nejad, S. M., Sipe, N. G., & Burke, M. I. (2013). *Retail travel behavior across socio-economic groups: a cluster analysis of Brisbane household travel survey data*.

Siren, A., & Hakamies-Blomqvist, L. (2004). Private car as the grand equaliser? Demographic factors and mobility in Finnish men and women aged 65+. *Transportation*

Research Part F: Traffic Psychology and Behaviour, 7(2), 107-118. doi:http://dx.doi.org/10.1016/j.trf.2004.02.003

van den Berg, P., Arentze, T., & Timmermans, H. (2013). A path analysis of social networks, telecommunication and social activity–travel patterns. *Transportation Research Part C: Emerging Technologies, 26*, 256-268. doi:http://dx.doi.org/10.1016/j.trc.2012.10.002

Vijayaraghavan, P., Vosoughi, S., & Roy, D. (2017). *Twitter Demographic Classification Using Deep Multi-modal Multi-task Learning.* Paper presented at the Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers).

Yang, Z., Lian, D., Yuan, N. J., Xie, X., Rui, Y., & Zhou, T. (2017). Indigenization of urban mobility. *Physica A: Statistical Mechanics and its Applications, 469*, 232-243. doi:https://doi.org/10.1016/j.physa.2016.11.101

Zhang, Y., & Cheng, T. (2017). *Feature Extraction for Long-term Travel Pattern Analysis*. Paper presented at the GISRUK 2017.

Zhao, J., Qu, Q., Zhang, F., Xu, C., & Liu, S. (2017). Spatio-Temporal Analysis of Passenger Travel Patterns in Massive Smart Card Data. *IEEE Transactions on Intelligent Transportation Systems, PP*(99), 1-12. doi:10.1109/TITS.2017.2679179

Zhao, J., Rahbee, A., & Wilson, N. H. (2007). Estimating a Rail Passenger Trip Origin-Destination Matrix Using Automatic Data Collection Systems. *Computer-Aided Civil and Infrastructure Engineering, 22*(5), 376-387.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8311

Relevance as an enhancer of votes on Twitter[1]

Arroba Rimassa, Jorge [a]; Llopis, Fernando [b] and Muñoz Guillena, Rafael [b]
[a] Universidad Central del Ecuador, Ecuador, [b] Department of Languages and Computer Systems, University of Alicante, Spain.

Abstract

The concept of the influence of Katz and Lazarfeld given in the last century has evolved thanks to the appearance of Social Networks and especially Twitter. Because this microblogging has allowed candidates for any election process to be closer to their electors and also allows an analysis of the contents of the messages to determine their polarity.

The relevance of the messages that measure the level of influence that can be had in the voters, incorporated into the traditional analysis of the Social Networks allow to have a greater degree of precision in the electoral predictions that are made using natural language processing, NLP.

We have introduced in the methodology that we propose a mechanism to enhance the votes of those messages that have a greater relevance and turn them into votes in order to improve the predictability of the electoral results.

The proposed methodology was applied in the election for President of the Republic of Ecuador that was held on February 19, 2017, obtaining a Mean Average Error, MAE = 1.4 that demonstrates the relevance of incorporating the variable Relevance as an enhancer of votes.

Keywords: *Relevance, Twitter, Election process.*

[1] This research work has been partially funded by the University of Alicante, Generalitat Valenciana , Spanish Government, Ministerio de Educación, Cultura y Deporte and ASAP - Ayudas Fundación BBVA a equipos de investigación científica 2016(FUNDACIONBBVA2-16PREMIO) through the projects, TIN2015- 65100-R, TIN2015-65136-C2-2-R, PROMETEOII/2014/001, GRE16- 01: “Plataforma inteligente para recuperación, análisis y representación de la información generada por usuarios en Internet” and “Análisis de Sentimientos Aplicado a la Prevención del Suicidio en las Redes Sociales” (PR16_SOC_0013).

1. Introduction

When Katz and Lazarsfeld (1955) formulate the notion of influence in *Personal Influence* in which they state that what matters is knowing three elements: *The audience*: knowing how many and how are the people who attend a message; *Content Analysis*: comprising the concept of the messages issued and the *Effect Analysis*: the impact of the mass media used; to understand the context and the conditions in which the "campaigns" were carried out in the media to modify the opinions and behaviors, they would never suppose that the voters, nowadays, would be totally communicated with their candidates, their messages are personal.

How this was achieved, by the appearance of Social Networks; politicians begin to interact with voters directly, receiving positions of unconditional acceptance and also rejection and repudiation of others. Is that Social Networks allow everyone to comment and especially as I would say (Eco, 2015). For the ease that the Twitter gives we will use this microblogging to make electoral predictions.

2. Related jobs

Since the appearance of Twitter in 2006, some authors have made predictions or post-processing of the results of electoral processes in the world using the information they can download from it. We have selected a sample of various electoral processes in the world in which using mentions and sentiments analysis, (Ceron, 2015) and (Singh, 2018) were used and in 45% of these works they used only the mentions method, which consists of counting the total number of downloads for one or the other candidate and 55%, used some classification method to determine the polarity of the messages, the most used being Naive Bayes. In 2012 there was the greatest increase in the predictive use of Twitter, reaching 36%, as shown in Figure 1.

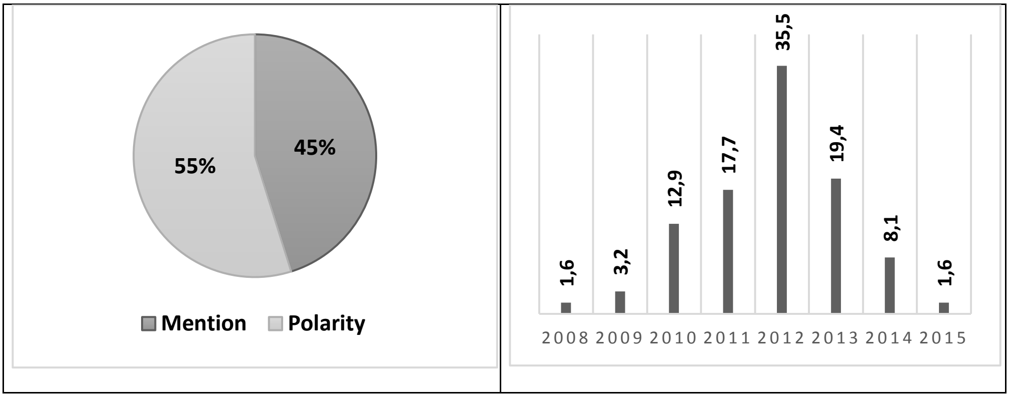


Figure 1. Method used to analyze and percentage of use for year. Source: Authors (2018).

3. Methodology

In order to predict the results of an electoral process using the analysis in the Social Networks, the opinion of the users in the network must be evaluated on both sides and the problem in question is focused on analyzing the position of each elector that the has manifested through a tweet and a way to solve this issue is to use natural language processing and text analysis to extract information. It must be analyzed how people interact (Blasquez & Domenech 2017).

The objective of the present investigation is to develop a methodology in which the concept and use of the Relevance as an enhancer of the votes of the messages that have a high influence are intended to be implemented.

As a practical case of application of this methodology, we used the Election for President of the Republic of Ecuador that was held on February 19, 2017.

The methodology developed is schematized in Figure 2. We will develop the steps of the methodology.

3.1. Defining accounts to follow

The accounts from which the documents were to be downloaded were first defined, the candidates Moreno, Lasso, Viteri and Moncayo were competing with some option for the first places for the presidency and a small percentage of the electorate would be for any of these others four candidates: Bucaram, Espinel, Zuquilanda and Pesántez; to these four candidates, for practical purposes they will be referred to as "others". In this sense, the official Twitter accounts of the candidates and their party were defined to extract the tweets from the users who follow them.

3.2. Twitter APP

Using the Twitter APP tool, the download was started from December 2016 until February 14, 2017; following the observations of (Tumasjan, 2010) on the periodicity of data collection. The elections would be on February 19, 2017 and a total of 823,135 tweets were downloaded corresponding to the users who follow these official accounts.

3.3. Collection of documents for each candidate

By gathering all these documents, the collection of documents was obtained for each of the candidates to be evaluated.

3.4. Pre-processing of Messages

This phase has the basic objective of reducing the dimensionality that usually presents problems later when the supervised learning methods of text classification are applied. The problems are not only due to computational reasons but also to the overfitting that can be presented in the dataset.

The mechanisms used for all the methods were: elimination of stopwords that contribute nothing to the text, the stopwords that were used are those that come by default in R. The cleaning process of blank spaces, markups, was also performed the emoticons, the reference links and the tags present in each text. Those terms whose distribution of frequency of appearance in the documents was much reduced were also eliminated.

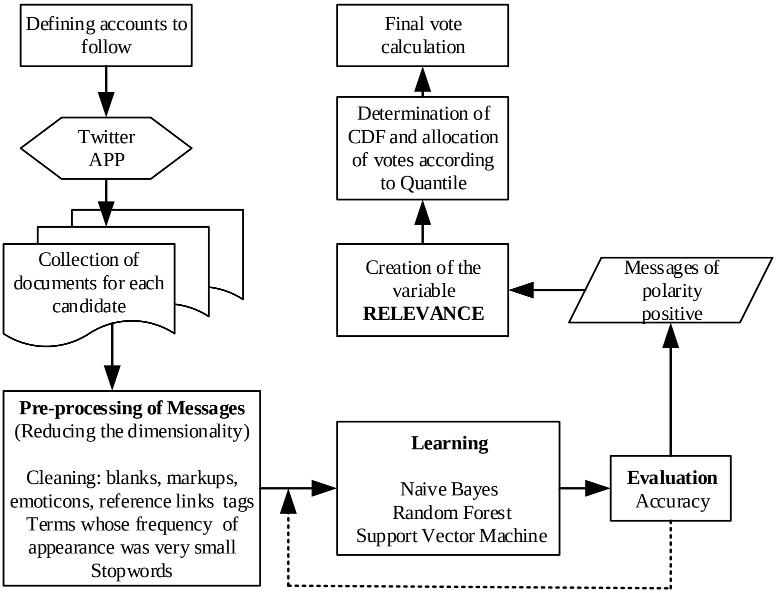

Defining accounts to follow
Twitter APP
Collection of documents for each candidate
Pre-processing of Messages
(Reducing the dimensionality)
Cleaning: blanks, markups, emoticons, reference links tags
Terms whose frequency of appearance was very small
Stopwords
Learning
Naive Bayes
Random Forest
Support Vector Machine
Evaluation
Accuracy
Messages of polarity positive
Creation of the variable
RELEVANCE
Determination of CDF and allocation of votes according to Quantile
Final vote calculation

Figure 2. Diagram of the methodology used. Source: Authors (2018).

3.5. Learning

Supervised classifiers require that the dataset be trained and then tested. The training of the dataset was done using the criteria of an expert manually, in which a sample of tweets was selected and manually classified according to the subjective value of each message in two categories, for or against of each of the candidates.

The training data are clean vectors, formed by the words of a tweet, to which they have been trained manually with a polarity value; and that function is validated with a set of test data. Once the classifier has the desired precision, this function is used to classify the rest of the data.

Next, with the training data set, the respective test of each dataset was carried out using each of the methods considered: Naive Bayes, Random Forest and Support Vector Machine.

3.6. Evaluation

For the evaluation, the Accuracy metric was used, which in general terms accounts for the overall presition in the performance of the classification method.

3.7. Messages of polarity positive

Once the respective dataset of each candidate was evaluated, only those messages that had positive polarity were considered.

3.8. Creation of the Relevance variable

Relevance is an indicator that determines the degree of importance that an object may have, in terms of communication, it can be said that one message is more relevant than another because of the level and influence it may have on the voters. Within the opinion leaders this influence is due to two factors fundamentally, its popularity and its prestige.

Our proposal to measure relevance is given by a function that depends on the followers you have, the degree of acceptance and the dissemination of the message.

But what can affect the relevance of a message in the political work ?. Basically in the adepts and in the potential followers that this can generate.

The makers of opinion are not necessarily the mass media, nor the politicians themselves, nor the opinion leaders; but they are ordinary people. Take the example of a father of a family, who with his authority affects the political decision of his children; Take also the example of a message that has a high degree of relevance, the followers of it make it their own and relay it to another, becoming a channel of dissemination. It is logical to suppose then that to a greater degree of relevance of an adept and of a message this one must have a reward, that becomes to gain adepts of such message and therefore in votes by a political option.

We have quantified this vote gain, or conversion given by the relevance in potential votes that can be generated. Those messages that are more relevant will have a greater impact on the electorate than those that do not have as much relevance. This way of quantifying or converting the votes is assumed to apply to those voters who have not yet decided who to vote for, the undecided ones, who are usually in a large majority in electoral campaigns and as the elections approach they diminish to the extent that they are aligning themselves with one or another position. This way of going ascribing to one or another position is given for many reasons, the one that interests us is the one that is based on relevant information that these undecided have at their disposal and that affects them in such a way that makes them adopt a position defined.

For Twitter several metrics have been defined on the relevance of the messages and they are used to measure the importance of a Twitter account, the most used is the one that measures the number of followers or favoriteCount, but we also want to measure the diffusion of the political messages as a mechanism of persuasion in other voters. As Twitter is increasing the amount of information continuously, then it is necessary to measure this expansion using another metric, that of the messages that are forwarded, retweetCount. In the present investigation we have defined the Relevance in function of these described parameters, giving greater importance or weight to the forwarding of the messages, to the extent that they serve as diffusers of the message. In equation 1 this metric is defined.

$$Relevance = \frac{favoriteCount + 3retweetCount}{Ntweets} \tag{1}$$

This metric that gives a higher score to those messages that have greater diffusion and that also come from accounts that have a large number of followers applies only to those tweets with positive polarity, which are in favor of a certain candidate.

3.9. Determination of CDF and allocation of votes according to Quantile

Once the calculation of the Relevance variable is determined, we evaluate its density function, given that this measure is a random variable, given that it comes from a biased sample of voters, made up of those citizens who send their messages on Twitter. The density function that best fits the data of the Relevance variable is a Johnson type Sl, whose parameters are: (Katz, 2009)

Table 1. Parameters of Johnson Sl Distribution of variable Relevance of messages.

Type	Parameter	Estimation
Shape	γ	1,143
Shape	δ	0,012
Localitation	θ	0,000
Scale	σ	1

Source: Authors (2018)

We used the Kolmogorov-Smirnov-Lilliefors test, K-S-L, (Pedrosa, 2015) for the goodness of fit and we accepted the null hypothesis H_0 = *the data come from a Johnson Sl distribution.*

With this density function, then, it's respective CDF cumulative distribution function was used to determine the thresholds for which we would give the conversion of the votes given by the value of the relevance. We use the quantile 97,5%, $Q_{97,5\%}$, as a threshold, over which if the Relevance is greater, that voter would be given a total of three votes, since his message could affect two other people, if the Relevance is greater than the 95% quantile, $Q_{95\%}$, but less than $Q_{97,5\%}$ will be assigned two votes, the one may affect one more voter. For values lower than the $Q_{95\%}$ we assumed that it would not have an impact on other voters. In Figure 3, the distribution of the Relevance variable and the vote allocation scheme are shown.

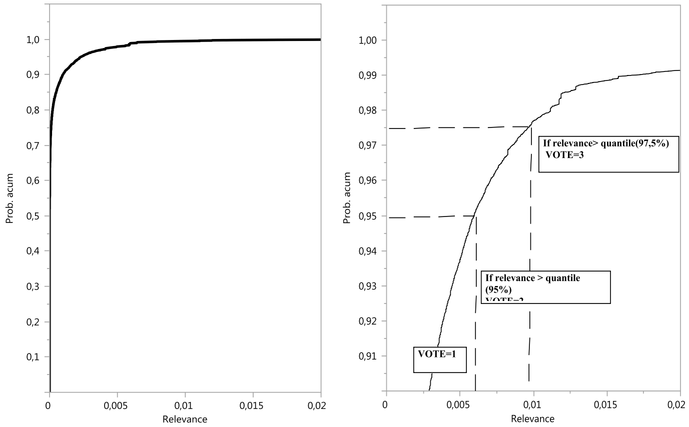


Figure 3. Johnson SI Distribution the variable Relevance and assignment of votes. Source: Authors (2018).

3.10. Final vote calculation

The final calculation of the votes is presented in the next section.

4. Results

Once for each Twitter message has been determined its polarity and its assessment in votes has proceeded to obtain the results; As shown in Table 2, the comparison between the official results given by the National Electoral Council of Ecuador, CNE, and those obtained using the defined Relevance methodology is presented. An MAE = 1.4 compared to the prediction made analyzing the Twitter messages with the proposed methodology using the valuation of votes given by the variable Relevance make this methodology an alternative in the electoral prediction.

Table 2. Comparison of results between Official results and the method using Relevance factor.

Candidate	Official Result CNE	Using the Relevance factor
Moreno	39,4	38,3
Lasso	28,1	29,7
Viteri	16,3	14,5
Moncayo	6,7	6,0
Others	9,5	11,5
MAE		1,4

5. Conclusions

The present study tries to demonstrate that the analysis of the tweets emitted by the users about their electoral preferences is as reliable as the results issued by different surveys.

The contribution of this research is the incorporation of the variable Relevance to enhance the electoral vote.

Additionally, the cost involved in the application of this methodology and the survey is incomparable; to more than the speed in the delivery of results.

References

Blasquez, D. & Domenech, J., (2017). Big Data sources and methods for social and economics analyses. *Technological Forecasting and Social Change*.

Ceron, A., Curini, L., & Iacus, S., (2015). Using social media to forecast electoral. *Statistica Applicata Italian Journal Of Applied Statistics*, 239-261.

Eco, U. (10 de 06 de 2015). *lastampa*.

Katz, E. & Lazarfeld, P., (2009). *Personal Influence, the Part Played by People in the Flow of Mass Communications.* New Jersey: Transaction Publishers.

Pedrosa, I. J.-B.-F.-C. (2015). Pruebas de bondad de ajuste en distribuciones simétricas. *Universitas Psychologica*, 245-254.

Singh, P., & Sawhney, R., (2018). *Progress in Advance Computing and Intelligent Engineering.* Singapore: Springer Nature Singapore.

Tumasjan, A. S. (2010). Predicting Elections with Twitter. What 140 Characters Reveal about Political Sentiment. *Fourth International AAAI Conference Weblogs and Social Media*, (págs. 178-185).

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8314

Big Data and Data-Driven Marketing in Brazil

Finger, Vítor[a]; Reichelt, Valesca[b] and Capelli, João[c]
[a]Department of Business Administration, Escola Superior de Propaganda e Marketing, Brazil, [b]Department of Business Administration, Universidade Luterana do Brasil, Brazil [c]Department of Business Administration, Escola Superior de Propaganda e Marketing, Brazil.

Abstract

The main purpose of this article is the understanding of which marketing strategies related to big data are being implemented by Brazilian companies in different sectors, in addition to assessing these actions within an already established construct. To reach the proposed objectives, an exploratory, qualitative research was conducted using the multiple case study method. Thus, data were collected through bibliographical, documentary and semi-structured interviews, with the intent of formulating the construct by which the companies are studied. The study unit interviewed consisted of market professionals and big data specialists. As the main result, it was widely noticed the application of strategies related to big data by the companies surveyed. The classification of these actions within an already established construct, however, was not possible, since it was understood the existence of distinct stages of adoption for this technology, and it was not possible to label these companies as users of big data.

Keywords: *Big Data; Digital Marketing; Data-Driven Marketing.*

1. Introduction

As time progressed and online interactions evolved, the amount of data and information over the network has grown exponentially. There are currently more than three billion people connected to the Internet (Emarketer, 2017), generating e-mails, having financial transactions and buying products on a daily basis. If analyzing data was a competitive advantage, it is now a matter of survival.

This increasingly extensive and multifaceted set of data that grows progressively, coming not only from new sources but also in new shapes, characterizes a new stage of data analysis that the marketing world is experiencing: the big data (Lohr, 2012). More than this, the modern marketing professional must follow the dispersive evolution of existing channels: the internet that was previously accessed through the computer, can now be reached on cell-phones, watches and even the consumer's glasses. Terms such as Internet of Things and Omnichannel will be increasingly present in the life of these professionals (Ashton, 2009). The complexity created by the amount of information on the network, and the different channels of access to it, makes working with it a barely superhuman task - and that is where programmatic marketing suits perfectly. Marketing automation provides technologies that leverage consumer intelligence systems, optimize interactions across channels, and monitor changes in customer behavior.

Thus, the present article aims to present the concept of big data, highlighting its main characteristics and utilities, in order to analyze the application of data-driven digital marketing. In this way, it is intended to identify the practices of big data adoption, within the cases studied, and to compare them with those found in the literature, pointing out similarities, contradictions and possible actions to circumvent such gaps. However, the literature on the subject is still limited, so this work contributes to the construction of scientific knowledge in the area.

2. Big Data Adoption

In partnership with Saïd Business School, from Oxford University, IBM Institute for Business Value developed in mid-2012 a study with 1144 professionals from 95 countries and 26 industries that aimed to identify how innovative companies have effectively used big data, seeking results focused on the client and taking advantage of the data present in its information ecosystem (Schroeck, Schockley, Smart, Romero-Morales, & Tufano, 2012). In order to avoid biased results, it was sought to have a sample that is active in different areas, among entrepreneurs and information technology professionals.

Their study understands big data as the convergence of four of its dimensions: volume, variety, velocity and veracity. In light of this concept, it is also established the perspective of "big data adoption", a term used to represent the natural evolution of elements needed to create competitive advantage in the current global marketplace - such as data, sources, technologies and skills (Schroeck et al., 2012).

Based on these concepts and the study carried out, five common trends were found: a) usage of big data focusing on consumer centricity; b) demand of an extensible and scalable information management system to develop big data; c) exploitation of internal data for the first efforts in the area; d) the need for great analytical capacity in order to obtain greater data value; e) identification of four stages of big data adoption (Schroeck et al., 2012).

Regarding the four described stages, a construct is born by which the evaluation of corporations within their level of involvement with big data is possible, identifying where the company is in the trajectory of full-scale implementation. Based on the description of the level of big data activities developed in their organizations, four stages of adoption and progression were suggested: Education, Exploration, Engagement and Execution (Schroeck et al., 2012). The first stage of Education has its main focus on the dissemination and development of knowledge within organizations. According to the authors, at this time companies study the potential benefits of big data analysis and technologies. In the next stage of Exploration, the organization structures a roadmap to develop big data. As explained by Schroeck et al. (2012), here begins the discussions about how to use big data to solve important business challenges.

Therefore, the third stage concerns the organization's effective Engagement with big data, initially proving its real business value. At this moment there is work to understand, technologies to test and skills needed to capitalize on new data sources (Schroeck et al., 2012). Finally, the Execution stage emphasizes the vast operationalization and implementation of analytical capabilities and big data within the company. Having the smallest share observed in the IBM study, this stage embraces big data’s market leaders, which are the first ones to implement big data as a way of transforming its business and extract the highest value possible from the information obtained (Schroeck et al., 2012).

3. Methods

For this study, it was conducted an exploratory research, ideal for problems that seek to foster ideas or clarification (Malhorta, 2005). The line of research was qualitative, commonly linked to the exploratory research. Qualitative analysis is important for the interpretation and understanding of how big data is being introduced and adopted by companies in the Brazilian market. The study was conducted in two stages: the first consisted of interviews with specialists and the second in a study of multiple cases.

The first stage of the research had interviews conducted with seven specialists: three managers of companies related to digital marketing, identified with the letter "G", and four big data specialists, identified with the letter "E". The latter are composed by market professionals, data scientists and university professors experts in the subject under study. The interviews were conducted in each professional’s own workplace. The techniques used for analyzing the collected data were category analysis and content analysis.

In the second stage of the research, four business cases of companies from different sectors that use data-driven marketing techniques were studied, in order to identify if big data practices were introduced - this assessment was conducted through a construct already validated by Schroeck et al. (2012). Multiple case studies fit on this study due to its understanding of theoretical and literal replications (Yin, 2015), which turn it into a tool for checking the results obtained in previous cases. Four companies that excel in the use and analysis of data to obtain information about their customers were selected and analyzed, being a retail store (company 1), an e-marketplace (company 2), a wine e-commerce (company 3) and an airline company (company 4).

This study also used bibliographical and documentary research to investigate and increase the quantity and quality of the data about the subject. Although it is a business sector somewhat debated by the bibliography, the most specific object in study - big data - is not yet widespread in the academic world. Thus, reports and other documents, from recognized organizations or entities, were necessary, which helped to enrich the repertoire under study, in order to better support the research and, consequently, its results.

4. Results

Through the interviews, it was realized that big data still has countless meanings, varying between huge amounts of data, generation of useful information and advanced analytical capacity. Although there are small differences, much of what was seen at the academy also appeared in the interviews conducted for the study - interviewees E02 and E03, for example, quoted Laney's (2001) view of big data’s three dimensions. In an attempt to unify the different concepts among the respondents into a single one, and based on what was predominant in the answers, it is possible to understand big data as the ability to deeply analyze a large amount of data, these being transformed into useful information.

As Taurion (2012), who defines big data as the understanding of an immense amount of data responsible for emphasizes the obsolescence of current technology, most interviewees mentioned “volume” within their definitions as another particularity of the concept. Although this is the biggest appeal of big data, since data increment becomes more important than the efficacy of a predictive model (Dumbill, 2012), volume was often viewed as the differential, but not the purpose. The concept of Zikopoulos, Deross, Bienko, Buglio, and Andrews (2015) it’s supported by interviewee E04, when both evidence the big data's ability to perform complex analyzes never seen before. In other words, what can be done with such an immense amount of data worth more than the fact that it is voluminous.

Besides its definition, it is interesting to analyze the thought of professionals and experts interviewed about the usefulness of big data. Not only it’s been talking about large-scale jobs that can not be done on a smaller scale, the purpose of these is to extract new ideas and create unique forms of value, changing markets, organizations and even the relationship between citizens and governments (Cuckier & Mayer-Schönberger, 2013). Among the answers provided during the interviews, the two main points commented were the personalization of user experience and predictive actions regarding demands and behaviors.

Thus, big data is strongly associated with relationship marketing in the stage of mass customization, advocated by Costa (2013). According to Peppers and Rogers (2014), the advantage of obtaining information about the company's public is to create the possibility of treating differently different consumers, as exemplified by the answers provided during the interviews. Likewise, the brand seeks to understand its client and his life moment to adapt to its needs (Peppers & Rogers, 2004), which supports the second big data utility. In short, as identified in the interviews, the characteristics to define big data are: the three dimensions (volume, velocity, and variety), the ability to analyze this immensity of data and the generation of useful information from them. On the other hand, its usefulness is to personalize the experience that users will have and predict future behaviors.

Seen that, the cases studied are analyzed through a broad bias, which instead of seeking a full adoption of big data, called by Schroeck et al. (2012) as the stage of Execution and which contains the smallest share of companies in this movement, seeks to identify practices used to understand the concept development within organizations. In other words, before turning big data into a single concept that involves a binary “yes or no” response to its adoption, it must be understood that it encompasses a number of strategies and actions that may be used in whole or in part by companies. Like Schroeck et al. (2012) clearly demonstrated, there are several stages of engagement with big data practices, and their adoption is progressive.

Regarding the volume needed to consider data analysis as big data, the cases studied showed different results. Schroeck et al. (2012) points out that the definition of what constitutes a large amount of data varies according to the industry studied, and the present cases are from four different ones. All cases, however, have volumes starting from millions of users, with Company 2 having a database with more than 132 million contacts. If for Dumbill (2012) the data increment becomes more important than the predictive model, and Cavoukian and Jonas (2012) explain that the great change is in the maximization of information, sources and competitiveness, it is understood that the volume used by the studied firms fits into such descriptions, since results of revenue and performance evidence the advantages gained from their work.

On the subject of variety of data and its sources, especially on unstructured data (Dumbill, 2012), the studied companies still have a small development. Although Company 2 uses different data sources to control its results, and Company 4 crosses data from behavioral and transactional sources, these data are still possible to work with traditional analysis tools. The big data differential, according to the interviewee E04, is the possibility of crossing unstructured structured information such as messages, texts, voices and the Internet of Things.

Even though it does not fit as a proof of data diversity, the analysis and crossover work carried out by Company 4 shows another important side of the big data: the need for large analytical capacity to obtain higher data values (Schroeck et al., 2012). By crossing behavioral data, such as personal preferences and lifestyle, with transactional, like flight history and main destinations, the airline reaches a unique understanding of its consumers, applying on them correct strategies and meeting their individual expectations. If big data's usefulness is based on thousands of variables and information that are cross-referenced and analyzed to generate one-to-one communication (E01), we have an example of how this is being done.

Schroeck et al. (2012) notes in his research the trend of using internal data as the first effort in the area, aligning with works found in Company 2. The execution of analyzes based on users origin allowed segmentation according to the customer’s country, creating specific approaches for each nationality and culture. Another interesting use of internal analysis is the cross of new names with those already in the company’s database, making it possible to identify the gender of a person and the kind of communication from that.

A final point advocated by Schroeck et al. (2012), and that is seen in the studied cases, relates to customer centricity. According to the interviewee G01, the use of data to customize and optimize user experience is the main aspect to be worked with big data, and Company 1 understands and applies this concept in its communication strategies. Through automated programs, and according to consumer interactions, the brand manages to work the customer's lifecycle and deliver the right message at the right time and to the right person. This work is so greatly done that, in 2015, the company received a Markie Awards in the Consumer Centricity category for customer-focused decision making.

The comparisons made in this chapter highlight the work done by the studied companies in relation to the concepts defended by experts and scholars on the subject. In this way, it became possible to approach the solution of the proposed research problem. The final considerations touch the subject so that the reader has an understanding of the result obtained with the present study.

5. Discussion

The characterization of big data, as evidenced both in literature and interviews, takes place in different ways. Seen this, it was not created a unique concept which, taken as truth, defines how within the universe of big data is certain action. However, it was possible to identify relevant aspects for its characterization.

Through the alignment between theoretical reference and data collected, it was noted that the three dimensions used as the definition for big data in its conception remain in vogue, since interviewees and authors have listed volume, velocity, and variety of data as essential aspects for the characterization of the term. Also, it was understood that not only the storage, but also the ability to analyze this immensity of data is vital for the work to have meaning in its realization, generating information useful to those who develop it. Synthesizing, the characterization of big data takes place both in the technical aspects and in the real use of data, through the personalization of experiences that the user will have and prediction of future behaviors of the same.

The four companies studied, each with its own differentiating aspects, evidenced to work its data in a continuous and progressive way, from the collection to the extraction of value. Understanding the customer lifecycle and customizing communications based on consumer preferences and history are some of the techniques used by companies during the study.

It was also found initiatives related to databases crossing, in order to enrich analyzes carried out and the value of information obtained. In this matter, Company 2 uses this technique to cross new contacts with those already existing in its base and Company 4 identifies the correct platform to get in touch with the client.

Although a certain part of big data conceptualization has been covered by the companies' actions in their data and communication works, there are points defined as essential by interviewees and scholars who were not mentioned. The use of semi-structured and unstructured data, such as voice messages, photographs and Internet of Things, constitutes an important gap for the full acceptance of companies in the big data universe. Likewise, the ability to forecast future trends and behaviors was little touched by companies, evidencing the lack of action in this regard. In short, it is possible to recognize the existence of strategies geared towards big data - advanced data extraction and analysis - in Brazilian companies, but they cannot be classified as big data users.

References

Ashton, K. (2009). That 'internet of things' thing. *RFiD Journal, 22* (7), 97-114.

Cavoukian, A., & Jonas, J. (2012). *Privacy by Design in the age of big data*. Retrieved September 08, 2015, from https://privacybydesign.ca/content/uploads/2012/06/pbd-big_data.pdf

Costa, G. C. G. (2013). *Negócios Eletrônicos: uma abordagem estratégica e gerencial*. Curitiba: InterSaberes.

Cuckier, K., & Mayer-Schönberger, V. (2013). *Big data: como extrair volume, variedade, velocidade e valor da avalanche de informação cotidiana.* Rio de Janeiro: Elsevier.

Dumbill, E. (2012). Big data now: 2012 edition. Sebastopol: O'Reilley Media.

Emarketer. (2017). eMarketer Updates Worldwide Internet and Mobile User Figures. Retrieved May 23, 2017, from https://www.emarketer.com/Article/eMarketer-Updates-Worldwide-Internet-Mobile-User-Figures/1015770

Laney, D. (2001). *3D Data Management: Controlling Data Volume, Velocity and Variety*. Stamford, CT: META Group. Retrieved February 20, 2001, from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf

Lohr, S. (2012, February 12). The Age of big data. *The New York Times*.

Peppers, D., & Rogers, M. (2004). *Managing customer relationships: A strategic framework*. New Jersey: John Wiley & Sons.

Malhorta, N. K. (2005). *Introdução à pesquisa de marketing*. São Paulo: Prentice Hall.

Schroeck, M., Schockley, R., Smart, J., Romero-Morales, D., & Tufano, P. (2012). *Analytics: The real-world use of big data. IBM Institute for Business Value*. Retrieved November, 03, 2015, from https://www.ibm.com/smarterplanet/global/files/se__sv_se__intelligence__Analytics_-_The_real-world_use_of_big_data.pdf

Taurion, C. (2012). *Você realmente sabe o que é big data?* IBM Developer Works. Retrieved September 10, 2015, from https://www.ibm.com/developerworks/community/blogs/ctaurion/entry/voce_realmente_sabe_o_que_e_big_data?lang=en

Zikopoulos, P., Deross, D., Bienko, C., Buglio, R., & Andrews, M. (2015). *Big data Beyond the Hyper: a guide to conversations for today's data center.* Mc Graw-Hill, Retrieved November 20, 2015, from https://www-01.ibm.com/marketing/iwm/iwm/web/signup.do?source=sw-infomgt&S_PKG=ov28197&dynform=11707

Harness the power of big data - The IBM big data Platform. (2013). Mc Graw-Hill. Retrieved September 10, 2015, from http://www-01.ibm.com/software/de/big-data/pdf/assets/Harness.PDF

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8315

Do People Pay More Attention to Earthquakes in Western Countries?

Habibi, Hanna [a] and Feld, Jan [b]

[a]School of Economics and Finance, Victoria University of Wellington, New Zealand, [b]School of Economics and Finance, Victoria University of Wellington, New Zealand.

Abstract

This paper investigates whether people from Western countries pay more attention to earthquakes in Western countries than those in non-Western countries. Using Google Trends data, we examine the proportion of Google searches from the United States, the United Kingdom, Canada, Australia, and New Zealand for 610 earthquakes across the world over the period of 2006-2016. Our results suggest that people in these countries pay around 44 percent more attention to earthquakes in Western countries, holding constant earthquake magnitude and number of casualties. Our results remain significant and similar in magnitude after controlling for geographical and social characteristics, but reduce in magnitude to almost zero and become insignificant after controlling for GDP per capita of the countries where the earthquake struck. Our results suggest that there is a developed country bias, rather than a Western country bias, in people's attention. This bias might lead to a lower flow of international relief to economically less developed countries, which are less able to deal with disasters.

***Keywords**: Public attention, natural disasters, internet search volume*

1. Introduction

Public attention to critical events is important because it leads to action from non-profit organisations and governments (Newig, 2004; Newell, 2006). For example, the rise in media attention to the issue of climate change during 2006-2007 led to a considerable increase in national-level climate legislations in 2007-2008 (Schmidt et al., 2013). Yet, we know little about what generates public attention. There is anecdotal evidence that people from Western countries pay more attention to critical events in Western countries. Take, for example, natural disasters that hit a country unexpectedly. Franks (2006) compares the media coverage of Hurricane Katerina in the United States and Hurricane Stanley in Guatemala that struck within weeks from each other in 2005. By the end of January 2006, the newspapers in the United Kingdom referred to Hurricane Katerina 3,105 times, while there were only 34 mentions of Hurricane Stanley. However, this comparison only relies on two natural disasters and furthermore, Frank looks at media coverage and not directly at public attention. Media coverage might have other drivers. We do not know whether people are biased towards natural disasters in Western countries as well.

In this paper, we test whether people from Western countries pay more attention to earthquakes in Western countries holding constant earthquake characteristics such as, the magnitude of the earthquake, the number of deaths, and whether the earthquake generated a tsunami. We measure public attention to 610 significant earthquakes across the world from 2006-2016, using the proportion of Google searches on the keyword “earthquake + country name” from internet users in the United States, the United Kingdom, Canada, Australia, and New Zealand. We broadly follow Samuel Huntington (1993) and categorize Western countries as Western Europe, the United States, Canada, Australia, and New Zealand.[1]

A number of studies have investigated the determinants of media attention to natural disasters (Eisensee and Strömberg, 2007; Van Belle, 2000; Koopmans and Vliegenthart, 2010). Koopmans and Vliegenthart (2010) who have the most similar variables to our study, investigate the determinants of media attention from the United States, the United Kingdom and the Netherlands and find insignificant results for the earthquakes in Western countries. While these studies focus on media attention, we use Google Trends to measure public attention directly. Factors such as the limited number of reporters in the country of the earthquake might have an influence on the media attention to such earthquakes but, nowadays, the events that grab people’s attention are not limited to those that are covered in the media. In sum, we are the first to examine explicitly whether people, as opposed to the media, in Western countries pay more attention to earthquakes that strike in Western countries than those in non-Western countries.

[1] Beside these countries, Huntington recognizes Papua New Guinea and French Guiana as Western countries. We have decided to exclude them because they are not generally considered as Western countries.

2. Data and Empirical Strategy

2.1 Data

We obtain our data on earthquakes from the Global Significant Earthquake Database provided by National Oceanic and Atmospheric Administration (National Geophysical Data Center, 2017).[2] In our analysis, we control for earthquake characteristics including magnitude, death toll, and whether the earthquake generated a tsunami. Table (1) shows summary statistics for our estimation sample.

Table 1. Descriptive Statistics

	(1) **N**	(2) **Mean**	(3) **SD**	(4) **Min**	(5) **Max**
Public Attention	2,950	1.42	8.38	0	100
Western	2,950	0.15	0.36	0	1
Magnitude	2,950	5.98	1.09	1.60	9.10
Tsunami	2,950	0.20	0.40	0	1
Number of Deaths	2,950	728.3	13,490	0	316,000
Distance (in 10,000 kms)	2,950	0.994	0.424	0.019	1.959
Common Border	2,950	0.02	0.12	0	1
Share of Migrants in the Country of Earthquake	2,950	0.25	0.44	0	5.75
Colony	2,950	0.07	0.25	0	1
Share of Christians	2,950	0.43	0.41	0.00	0.99
Common Official First Language	2,950	0.29	0.45	0	1
GDP per capita (in 10,000s USD)	2,950	1.224	1.572	0	6.221

NOTE. — All numbers are based on our estimation sample. 'SD' refers to the standard deviation of the respective variable.

We use monthly Google Trends data to measure public attention. Google Trends is an analytical tool that provides data on the proportion of Google searches on a particular topic, reflecting Google users' interest in that topic. This analytical tool allows for comparison between different topics adjusting for time and location. It takes a random sample of Google search data as representative of all Google searches and provides a proportionate measure, scaled from 0-100, that shows the amount of Google searches on a particular topic in a given time and location. This scaling means that Google Trends data adjusts for differences in the number of internet users in different locations (Google, 2017). We collect the data for Google searches from the United States, the United Kingdom, Canada,

[2] This database contains information on earthquakes that meet at least one of the following criteria; 10 or more deaths, approximately 1 million USD or more damage, a magnitude of 7.5 on the Richter scale or greater, the Modified Mercalli Intensity (MMI) of X or greater [2], or whether the earthquake generated a tsunami.

Australia, and New Zealand on the search term “earthquake + country name” in the month in which the earthquake occurred.[3] We measure attention by exploring Google Trends data for “earthquake + country name” and adjusting the search for geolocation and time period of interest. We obtain the monthly Google Trends score of our keywords for the period 2006-2016 from specific origin of search – limited to the five countries that we measure their attention -, one at a time. Since Google Trends allows for the comparison between different keywords, we compare the keywords for all the earthquakes to identify the earthquake that received the highest amount of attention in the sample period in a given country. For example, Google users in the United States paid the highest attention to “Earthquake Japan” in March of 2011, and for this reason, its Google Trends score is 100. Following this approach, we obtain the data on the proportion of Google searches on our 610 earthquakes of interest from the mentioned five countries for the period of 2006-2016. As shown in table (1), the mean of public attention equals 1.42 percent. The skewed distribution and considerably low mean are the result of the fact that people pay so much attention to a few earthquakes while many earthquakes barely receive any attention.

Our explanatory variable of interest is a Western dummy. Relying broadly on Huntington (1993), we categorize Western countries as countries in Western Europe, the United States, Canada, Australia, and New Zealand. Almost 15% of earthquakes in our database struck in Western countries.

We categorize our control variables into three categories, geographical, social, and economic characteristics. Geographical characteristics includes the distance in kilometres and whether the countries are neighbour. The data on these two variables is available on GeoDist database (CEPII, 2017). Social characteristics include common official first language, colony, share of migrants, and share of Christians. The data on common official first language and colonial ties is also available on the GeoDist database (CEPII, 2017). We use the data on bilateral migration from the World Bank (Özden et al., 2011) to measure the share of migrants from the country of the earthquake in the country where we measure public attention. Finally, we obtain the data on the share of Christians in the countries where the earthquake struck in 2015 from the World Christian Database (Johnson and Zurlo, 2007). Lastly, we use GDP per capita in USD as our measure of economic characteristics and obtain the data on countries GDP per capita from the World Bank.

[3] Google is by far the most popular search engine in the world as well as the five countries that we measure their public attention. Google captured almost 87%, 90%, 94%, 95%, and 91% of the market share of search engine users in the United States, the United Kingdom, Australia, New Zealand, and Canada respectively in 2017. This share on average is almost 17 times higher than the share of second popular search engine (Bing) in these countries (Stats, 2017).

2.2 Empirical Strategy

In order to understand the role of Western country status in the public attention paid to earthquakes, we estimate four specifications using the following empirical model:

$$Ln(Attention_{ic}) = \beta_1 Western_i + \delta X'_{ic} + u_{ic} \qquad (1)$$

where $Ln(Attention_{ic})$ is the natural logarithm of the Google Trends score for earthquake i in country c (country of attention) which is our measure of public attention. Because we have many values between zero and one, we add one to the Google Trends score and take its log. $Western_i$ is a dummy variable that is equal to one if the earthquake struck in a Western country. The coefficient of interest is β_1 which shows the increase in public attention when the earthquake occurred in a Western country. To allow for the number of deaths and the magnitude of the earthquake to have non-linear effects on attention, we include cubic polynomials of magnitude and the number of deaths. In all specifications, we control for these earthquake characteristics. We also include country of attention fixed effects, and to account for countries paying more attention to earthquakes in their own country we include a domestic dummy. Because we observe the attention to the same earthquake from five countries, we cluster standard errors at the earthquake level.

The vector X'_{ic} contains our three sets of control variables, geographical, social, and economic characteristics that we include in some of our specifications. Geographical characteristics contains bilateral distance and neighbour - a dummy variable equalling one when the two countries are contiguous. Social characteristics contains four control variables: share of migrants that is the share of migrants from the country of earthquake in country of attention, share of Christians in the country of the earthquake, a dummy for whether the country of earthquake and country of attention had colonial ties, and a dummy if the two countries share an official first language. Our measure of economic characteristics is GDP per capita of the country of earthquake for each year.

3. Results

Table 2 shows how Western country status predicts public attention in the United States, the United Kingdom, Canada, Australia, and New Zealand. We find that people in these countries pay more attention to earthquakes in Western countries. Our results suggest that earthquakes in Western countries receive 0.37 log points - about 44 percent - more public attention than earthquakes in non-Western countries with the same magnitude and number of casualties. This result stays consistent after additionally controlling for geographical and social characteristics (columns 1, 2, and 3). However, after controlling for economic characteristics we get insignificant results for Western country status (column 4). Results of this specification suggest that a 10,000 USD increase in GDP per capita increases the public attention to earthquakes by 0.15 log point - around 16 percent. The Western bias seems to be driven by people paying more attention to earthquakes in economically developed countries. The results of our last specification are consistent with Koopmans and Vliegenthart (2010) who found insignificant results for Western country status when controlling for GDP per capita.

Table 2. Determinants of Public Attention to Earthquakes

	(1)	(2)	(3)	(4)
Dependent Variable:	Log Public Attention	Log Public Attention	Log Public Attention	Log Public Attention
Western	0.365***	0.376***	0.454***	0.024
	(0.083)	(0.085)	(0.083)	(0.146)
Distance (in 10,000 kms)		-0.140**	-0.127**	-0.125**
		(0.021)	(0.022)	(0.021)
Neighbour		-0.055	-0.166	-0.350**
		(0.096)	(0.112)	(0.119)
Colony			-0.075	-0.092
			(0.032)	(0.034)
Common Official First Language			-0.136***	-0.089**
			(0.041)	(0.039)
Share of Migrants from Country of Earthquake			0.167***	0.160***
			(0.043)	(0.049)
Share of Christians			-0.189***	-0.187***
			(0.070)	(0.058)
GDP per capita (in 10,000s USD)				0.146***
				(0.041)
Earthquake Characteristics	YES	YES	YES	YES
R-squared	0.277	0.284	0.320	0.381
Observations	2,950	2,950	2,950	2,950

NOTE. — The dependent variables in all Columns are the log of public attention, which is a proportionate measure scaled from 0-100 calculated by Google Trends. All columns are estimated with OLS regressions that include country-of-attention fixed effect, magnitude of the earthquakes, number of deaths and a dummy variable, which is equal to 1 if the earthquake generated tsunami. We include cubic polynomials of magnitude and number of deaths to control for their non-linear effects. The regressions also include a dummy variable, which is equal to 1 if the earthquake is stricken in the same country that attention is captured from. Robust standard errors in parentheses are clustered at earthquake level. * p<0.1, ** p<0.05, *** p<0.01.

4. Conclusion

We have estimated the role of Western country status on the public attention paid to 610 earthquakes across the world from 2006 to 2016. Our findings show that people from the United States, the United Kingdom, Canada, Australia, and New Zealand pay on average 50 percent more attention to earthquakes in Western countries than to earthquakes in non-Western countries. This result disappears after controlling for GDP per capita of the country in which an earthquake struck, suggesting that the bias in attention is mainly towards more economically developed countries rather than Western countries. A potential consequence of this bias is that it may make it difficult to motivate governments to provide relief for less developed countries who may need the help more urgently.

References

Balla, S. J., Lodge, M. & Page, E. C. (Eds.). (2015). The Oxford handbook of classics in public policy and administration. OUP Oxford.

CEPII - Centre d'Études Prospectives et d'Informations Internationales (2017) GeoDist database. [Online]. Available: http://www.cepii.fr/anglaisgraph/bdd/distances.htm [Accessed: 2017 September].

Eisensee, T., & Strömberg, D. (2007). News droughts, news floods, and US disaster relief. The Quarterly Journal of Economics, 122(2), 693-728.

Franks, S. (2006). The CARMA report: western media coverage of humanitarian disasters. The Political Quarterly, 77(2), 281-284.

Global Bilateral Migration Database, World Bank Group Özden, Ç., Parsons, C. R., Schiff, M. & Walmsley, T. L. (2011). Where on earth is everybody? The evolution of global bilateral migration 1960–2000. The World Bank Economic Review, 25(1), 12-56.

Google (2017) Google Trends. [Online]. Available: http://www.google.com/trends/. [Accessed 2017 June]

Huntington, S. P. (1997). The clash of civilizations and the remaking of world order. Penguin Books India.

Johnson, T. M. & Zurlo, G. A. (2007). World christian database. Leiden/Boston: Brill.

Koopmans, R. & Vliegenthart, R. (2010). Media attention as the outcome of a diffusion process—A theoretical framework and cross-national evidence on earthquake coverage. *European Sociological Review*, *27*(5), 636-653.

National Geophysical Data Center / World Data Service (NGDC/WDS): Global Significant Earthquake Database. National Geophysical Data Center, NOAA. doi:10.7289/V5TD9V7K [Accessed 2017 June]

Newell, P. (2006). *Climate for change: Non-state actors and the global politics of the greenhouse*. Cambridge University Press.

Newig, J. (2004). Public attention, political action: the example of environmental regulation. *Rationality and Society*, *16*(2), 149-190.

Özden, Ç., Parsons, C. R., Schiff, M. & Walmsley, T. L. (2011). Where on earth is everybody? The evolution of global bilateral migration 1960–2000. The World Bank Economic Review, 25(1), 12-56.

Schmidt, A., Ivanova, A. & Schäfer, M. S. (2013). Media attention for climate change around the world: A comparative analysis of newspaper coverage in 27 countries. *Global Environmental Change*, *23*(5), 1233-1248.

Stats, S. G. (2017). StatCounter Global Stats.

Van Belle, D. A. (2000). New York Times and network TV news coverage of foreign disasters: The significance of the insignificant variables. *Journalism & Mass Communication Quarterly*, *77*(1), 50-70.

WCD - World Christian Database (2017) Country / Religion Database. [Online]. Available: http://www.worldchristiandatabase.org/wcd/ [Accessed: 2017 October].

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8316

From Twitter to GDP: Estimating Economic Activity From Social Media

Indaco, Agustín
The Graduate Center CUNY, New York, U.S.A

Abstract

This paper shows how the use of data derived from Twitter can be used as a proxy for measuring GDP at the country level. Using a dataset of 270 million geo-located image tweets shared on Twitter in 2012 and 2013, I find that: (i) Twitter data can be used as a proxy for estimating GDP at the country level and can explain 94 percent of the variation in GDP; and (ii) that the residuals from my preferred model are negatively correlated to a data quality index which assesses the capacity of a country's statistical system. This suggests that my estimates for GDP are more accurate for countries which are considered to have more reliable GDP data. Taken together, these findings show that institutions and individuals could use social media data to corroborate official GDP estimates; or alternatively for government statistic agencies to incorporate social media data to complement and further reduce measurement errors.

Keywords: *National Accounts, Big Data.*

1. Introduction

Despite incessant debate about its ability to accurately measure the state of the economy, the gross domestic product (GDP) is still the most widely used indicator to gauge the economic performance of countries (Masood (2014)). One of the many problems with estimating GDP is that its measurement is often complicated and expensive to produce, particularly for developing countries. This could lead to measurement errors that in turn mislead policy evaluation and recommendations. Another concern is that given the importance surrounding official GDP estimates, both in terms of market fluctuations as well as public perception of politicians' performances, governments can find short-term benefits in manipulating these estimates. In light of this, a lot of research has been focused on alternative ways of measuring GDP other than the traditional sample survey method, both to corroborate as well as a control mechanism.

In this paper I will argue for the use of data derived from social media posts as a proxy for measuring GDP. By locating and analyzing the content of hundreds of million social media posts, I will show that one can estimate economic activity at the country level.

Social media can contribute greatly to economic research in this regard, as vast information can be extracted from the location and content of their posts. Measures taken from social media can serve both as substitutes as well as complements to traditional survey data. In particular, social media data has several properties that result beneficial when estimating economic measures. First, social media data is publicly available and has a low cost of obtaining and storing. Unlike survey data that are costly to recollect, social media data is organically being generated by users from all over the world and available to statistic agencies and the public at no cost (other than the necessary computing power and data storage). The public aspect of this data also allows for more transparency in official statistics, as the estimates could be replicated endlessly by individuals and institutions all over the world. Second, social media data is available in real time, which allows for economic estimates that are currently produced in annual or quarterly intervals to be produced at shorter time intervals. This would allow for clearer foresight for companies and individuals when making economic decisions. Third, given that geo-tagged social media posts can be geographically assigned to a precise location within approximately a 10 meter radius, one can aggregate social media posts at any sub-national geographical level one deems interesting. This includes aggregating data between areas that are not bound together politicaly and thus fabricate meaningful areas of study that are not possible with official datasets.

Recently, the use of visible light emanating from earth as captured by weather satellite images has been widely suggested as a good proxy for measuring economic activity in a series of papers. Different studies have shown that night lights can be used to measure GDP estimates at the country level (Pinkovskiy and Sala-i Martin (2016)), GDP growth at the country level (Henderson et al. (2012)) and GDP for sub-national regions (Doll et al. (2006), Henderson et al. (2012) and Sutton et al. (2007)). These studies have shown that the intensity of artificial night-lights highly correlates with GDP and thus can be used to estimate economic activity for different geographic regions.

In this paper I propose using posts from popular social media applications, in this case Twitter, as a measure that has all the same benefits as night-lights, but can be a more accurate estimator of GDP and has several other advantages.

For this paper I have all geo-located image tweets shared on Twitter for the years 2012 and 2013. I have two main findings: (i) that social media data can be used as a proxy for estimating GDP at the country level, as shown by the preferred model explaining 94 percent of the variation in GDP; and (ii), I find a negative correlation between the residuals of my model and a data quality score put together by the World Bank which suggests that my estimates for GDP are more accurate for countries which are considered to have more reliable GDP data. This is a strong result that suggests that social media data could be used as a complement to survey data to increase the accuracy of GDP estimates.

2. Data

2.1 Twitter data

Twitter is a social media application which allows users to post short messages of any subject of their choosing. These messages are known as tweets. Twitter emerged in 2006 and by 2012 it had 140 million global users which sent out 340 million tweets per day. Unless restricted by the user, tweets are publicly available and can be read via the application or on a web browser. Created as a text-only platform, Twitter initially did not allow users to share images, videos or other sorts of media in their tweets. This changed in August 2011, when Twitter rolled out a platform that allowed users to add images to their tweets.

The dataset used in this paper contains all geo-tagged image tweets posted on Twitter for years 2012 and 2013. This dataset was provided directly by Twitter, through a Twitter Data Grant submission in 2014 by the Cultural Analytics Lab. The total dataset contains 270 million tweets from all around the world. Each tweet contains information on: i) a unique identifier for each individual Twitter user; ii) the latitude and longitude (5 decimal points) of where the tweet was sent from; iii) the date and time in which the tweet was sent; iv) the image tweeted; and v) any accompanying text.

Table 1 summarizes this Twitter data by year and by country income groups (using the World Bank's classification). The breakdown of average tweets per income group shows that countries in higher income groups have more tweets.

Figure 1 shows that there is some clear visual patterns to the location and distribution of tweets worldwide that seem to represent economic activity and population density. The location from where each image tweet was sent is represented by a small light blue point. Clusters of light blue points can be found both in areas that are more densely populated as well as areas where we know have higher levels of per capita income. For example, in the United States, the largest concentration of image tweets seem to be centered in the coastal areas, but not so in the less-populated South West and Rocky Mountain States. South America has a cluster of tweets mainly surrounding big cities in Ecuador, Colombia and Venezuela in the north and Brazil, Argentina, Uruguay and Chile further south. In Africa, image tweets tend to be concentrated in richer countries: Morocco, Algeria and Egypt, and in Sub-Saharan Africa in South Africa, Nigeria and Kenya. Western Europe seems to be mostly lit up and the concentration of tweets becomes sparser as we move east into Ukraine, Belarus, Latvia, Estonia and ultimately into Russia.

Table 1. Twitter Data Summary Statistics: Mean and S.D.

	2012	**2013**
Tweets	109,678.1	528,694.9
	(354,724.1)	(2,397,003.8)
By Income Group		
High (122)	211,993.9	1,126,937.3
	(541,334.7)	(4,048,721.5)
Upper-middle (100)	89,123	406,004.9
	(201,367.9)	(836,887.4)
Lower-middle (86)	37,394.5	209,938.9
	(106,230.9)	(929,686.9)
Low (51)	735.9	3,602.5
	(898.9)	(4,370.9)

Note: number of countries per income group in brackets

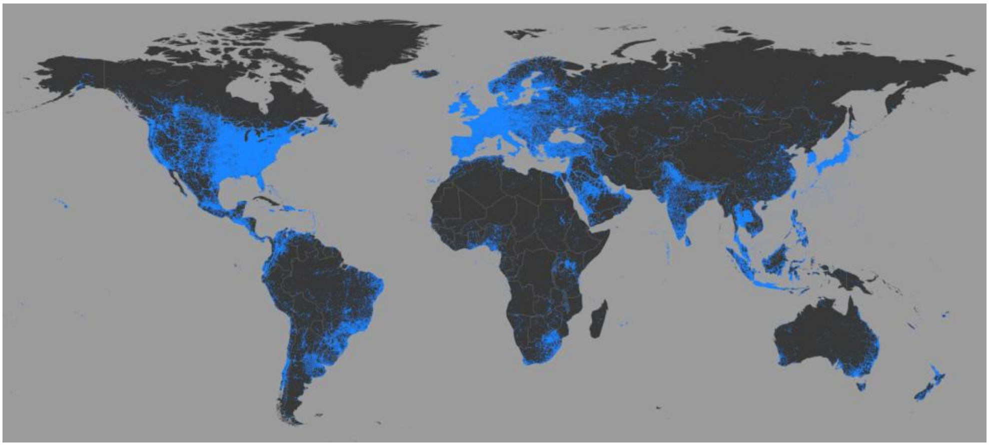


Figure 1. Map of image Tweets shared around the world in 2012 and 2013.

2.2 Socio-economic data

The World Bank provides freely and publicly available data on various relevant socio-economic indicators at the country level. Given that one of the main objectives of this paper is to provide an effective proxy for estimating GDP that allows for more transparency in official statistics, it is important that all the data used in this paper is publicly available and thus could be replicated by individuals and institutions. Besides from GDP, I also obtain total population for each country from the World Bank database.

Another indicator I obtain from the World Bank is the percent of the population that use the internet. Given that Twitter requires internet service access to establish a connection, the penetration of internet in a given country is a useful variable to include in our baseline regression.

The World Bank also produces a composite score assessing the capacity of a country's statistical office. In particular they focus on three specific areas: methodology, data sources, and periodicity and timeliness. The overall score is a simple average of all three area scores on a scale of 0-100, where higher values indicate higher quality data. In Section 3.1 I use these data quality scores to see if the discrepancies in our estimates are larger for countries with inferior data quality, as assessed by the World Bank.

3. Using Twitter to estimate GDP

The main goal of this paper is to see whether Twitter data is a valid proxy for estimating GDP. I will estimate GDP at the country level using tweets as the main variable of interest for panel data for years 2012 and 2013. For this I will estimate:

$$ln(GDP)_{i,t} = \beta_0 + \alpha_t + X'_{i,t} + \beta_1 \ln(Tweets)_{i,t} + \varepsilon_{i,t} \qquad (1)$$

where I estimate GDP for country i in year t. The vector Xi,t is composed of country characteristics including population, the share of the population with access to internet and continent to which it belongs. The coefficient of relevance to us is β1 which shows the relevance of the number of image tweets taken from that country in each of those years for estimating GDP. In Equation 1, year fixed effects (αt) control for any differences in use of Twitter from one year to the other. All of these are publicly available data.

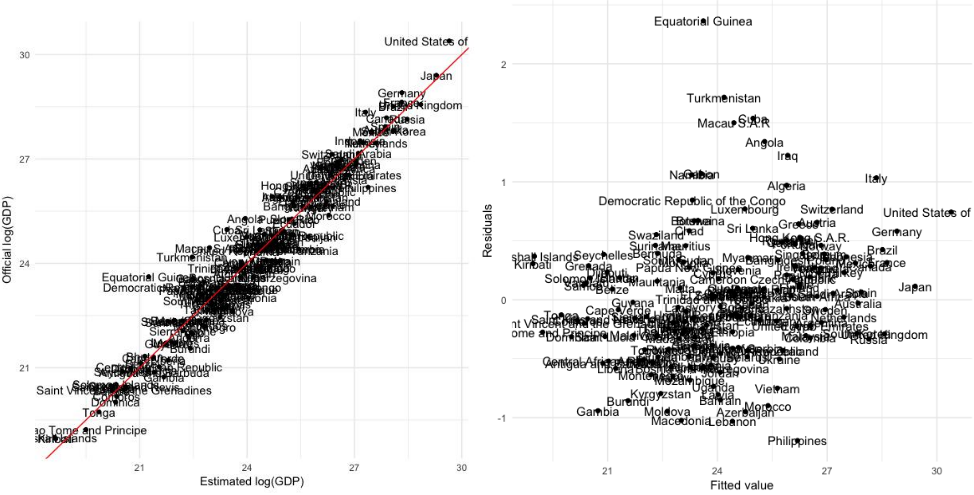
The corresponding estimates are reported in Table 2. There are 184 countries in our dataset for which I have data on GDP, Twitter and population, for both years. In column 1, I regress the natural log of GDP solely on the number of image tweets sent from each country. This is the baseline regression. The coefficient of interest on ln(Tweets) is highly significant and the R2 is 0.78. When the population of the country is included in column 2, the coefficient on ln(Tweets) is reduced, but remains statistically significantly different from zero at the 1 percent level, and R2 increases to 0.87. In column 3, I add categorical dummies for the continents in which each country is situated in. This captures the cultural differences in image sharing on social media platforms that exists between regions. Neither the coefficient of interest or the goodness of fit change greatly. Column 4 adds the share of the population that has access to internet. The number of observations are reduced to 180 countries per year because The World Bank does not have data on the share of the population with access to internet for six countries (these are: Libya, Kosovo, Curacao, Palau, South Sudan and San Marino). The coefficient of interest on ln(Tweets) is again reduced, but remains statistically significantly different from zero at the 1 percent level, and R2 increases to 0.94. These measurements are slightly larger than those obtained by similar studies using night-lights (Doll (2006) and Sutton (2007)).

Table 2 shows that the number of image tweets sent in a year is a pretty good measure for estimating GDP at the country level, being able to explain 78 percent of the variation in GDP on its own, and up to 94 percent when introducing other variables that are readily and publicly available. In all specifications, the coefficient on the number of image tweets is statistically significantly different from zero. Figure 2(a) is a visual representation of these estimates: the estimates lay pretty closely around the 45 degree line. There are a few exceptions that stand out; most notably Equatorial Guinea. Figure 2(b) plots the residuals of equation 1 against the fitted values, allowing us to study the distribution of the residuals; which seems to be randomly distributed around zero (i.e.: no clear pattern emerges).

Table 2. Estimating Country GDP

Dep. Var.: ln(GDP)	(1)	(2)	(3)	(4)
ln(Tweets)	0.66***	0.49***	0.45***	0.18***
	(0.02)	(0.02)	(0.02)	(0.02)
ln(Population)		✔	✔	✔
Continent			✔	✔
Internet				✔
R^2	0.78	0.87	0.89	0.94
Adj. R^2	0.78	0.87	0.88	0.94
Num. obs.	368	368	368	356

Note: ***p<0.01, **p<0.05, *p<0.10



Figures 2 (a) Estimated vs Actual GDP for 2013 and (b) Residual and Fitted Value for 2013 GDP Estimates

3.1 Data quality issues

While the previous section showed that image tweets could be used to estimate GDP at the country level, Jerven (2013) showed that GDP estimates have been criticized for being inaccurate, particularly in developing countries. If this is true, it could be the case that as I am trying to estimate the GDP reported by countries and not necessarily the true GDP and thus that the model's estimates are off because of measurement error on the official GDP estimates. If this is the case, data from tweets could be useful as an additional measure at the national level to produce more accurate estimates.

In order to analyze this, I will incorporate a measure of data quality put together by the World Bank. The World Banks Statistical Capacity Indicator is a composite score assessing the capacity of a country's statistical system. It is based on a diagnostic framework assessing the following areas: methodology, data sources, and periodicity and timeliness.

The overall score is a simple average of all three area scores on a scale of 0-100, where higher values indicate better data quality assessment.

Given that the World Bank works solely with Upper-middle income, Lower-middle income and Low-income countries, the data available for such measures are restricted to these countries. There are 140 countries for which there is an indicator on the quality of the data, as well as GDP, Twitter, population and percent of population with access to internet. Hence, I run the same regression in equation 1 for the subset of countries for which this data is available for 2012 and 2013. As can be seen in Columns 1-4 of Table 3, the overall estimates are very similar for this subset of countries as for our general model presented in Table 2, both in terms of the coefficient on ln(Tweets) as well as the R2. I then collect the residuals of equation 1 and run the following regression:

$$|Residuals|_{i,t} = \beta_0 + \beta_1 DataQuality_{i,t} + \beta_2 \ln(Tweets)_{i,t} + \beta_3 \ln(GDP)_{i,t} + \varepsilon_{i,t} \qquad (2)$$

where I regress the absolute value of the residuals for country i in year t on the data quality index, the number of tweets and GDP. The coefficient of interest is β1: a negative and statistically significant coefficient would indicate that our baseline model in Equation 1 more accurately estimates GDP for countries which have more reliable national account estimates. Column 5 of Table 3 shows that the data quality indicator coefficient is in fact negative and statistically significantly different than zero at the 1 percent confidence level. Column 6 includes countries' GDP estimate to control for the possibility that the model's estimates are more accurate for countries with larger economies. Adding this coefficient makes the coefficient on the data quality variable slightly more negative and still statistically significant. This shows that GDP estimates using our baseline model are more accurate for countries with high quality data, and vice versa. Given the long literature showing that official GDP estimates are inaccurate, it is important to acknowledge that it is possible that the GDP estimates we are trying to fit the model to are in fact inexact. The negative coefficient on the data quality index in equation 2 suggests that there is information to be captured from Twitter data that could help close the gap between estimated GDP and the true GDP. This is a strong result that suggests that social media data could be used as a complement to survey data to increase the accuracy of GDP estimates.

Table 3. Data Quality Issues

	(1)	**(2)**	**(3)**	**(4)**	**(5)**	**(6)**
Dep. Var.:	**ln(GDP)**	**ln(GDP)**	**ln(GDP)**	**ln(GDP)**	**Abs. Resid.**	**Abs. Resid.**
ln(Tweets)	0.60*** (0.02)	0.37*** (0.02)	0.39*** (0.03)	0.24*** (0.03)	✔	✔
Data Quality					-0.04** (<0.01)	-0.05*** (<0.01)
ln(Population)		✔	✔	✔		
Continent			✔	✔		
Internet				✔		
ln(GDP)						✔
R^2	0.76	0.90	0.91	0.93	0.03	0.07
Adj. R^2	0.76	0.90	0.91	0.93	0.03	0.06
Num. obs.	240	240	240	236	240	240

Note: ***p<0.01, **p<0.05, *p<0.10

4. Conclusion

The main goal of this paper was to investigate whether social media data from Twitter could be used as a proxy for estimating GDP at the country level.

For this, I have all geo-located image tweets shared on Twitter for the years 2012 and 2013. The first finding of this paper is that Twitter data can be used as a proxy for estimating GDP at the country level, and my preferred model can explain 94 percent of the variation in GDP. The coefficient on the number of tweets sent from each country is statistically significant.

I then go on to study the relationship between the residuals of my preferred model and a data quality score computed by the World Bank. Given the numerous concerns related to the accuracy of the official GDP estimates, particularly in less-developed countries, it could be the case that the figures we are trying to estimate are not in fact the true GDP. This could in turn cause our estimates to be imprecise because of the measurement error in the official GDP estimate we are trying to calibrate our model to in the first place. I find that the residuals from my model are in fact negatively correlated to the data quality index; which suggests that my estimates for GDP are more accurate for countries which are considered to have more reliable GDP data. These two findings taken together suggest that social media data could be used as a complement to survey data to increase the accuracy of GDP estimates.

These findings lead us to conclude that social media data contains useful information that can be used to estimate GDP at the country level. Potentially, this could be used for institutions and individuals to corroborate official GDP estimates, or alternatively for government statistic agencies to incorporate social media data to complement and further reduce measurement errors.

References

Doll, C. N. H., J.-P. Muller, and J. G. Morley (2006). Mapping regional economic activity from night-time light satellite imagery. *Ecological Economics*, 57(1):75–92.

Henderson, J. V., A. Storeygard, and D. N. Weil (2012). Measuring Economic Growth from Outer Space. *American Economic Review*, 102(2):994–1028.

Jerven, M. (2013). Poor Numbers: How We Are Misled by African Development Statistics and What to Do about It. Cornell University Press.

Masood, E. (2014). The Great Invention: The Story of GDP and the Making (and Unmaking) of the Modern World. Saqi Books.

Pinkovskiy, M. and X. Sala-i Martin (2016). Lights, camera income! illuminating the national accounts- household surveys debate. *The Quarterly Journal of Economics*, 131(2):579–631.

Sutton, P. C., C. D. Elvidge, and T. Ghosh (2007). Estimation of Gross Domestic Product at Sub-National Scales using Nighttime Satellite Imagery. *International Journal of Ecological Economics and Statistics*, 8(S07):5–21.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8320

Has Robert Parker lost his hegemony as a prescriptor in the wine World? A preliminar inquiry through Twitter

Compés-López, Raúl[a b]; Font-Julian, Cristina I.[b] and Orduna-Malea, Enrique[b]
[a]Departmento de Economía y Ciencias Sociales; [b]Universitat Politècnica de València, Spain

Abstract

The aim of this work is to determine to what extent Robert Parker has lost his influence as a prescriber in the world of wine through a webometric analysis based on a comparative analysis of Parker's web influence and that of a competitor who represents an anthitetical vision of the world of wine (Alice Feiring). To do this, we carried out a comparative analysis for Parker's (@wine_advocate) and Alice Feiring's (@alicefeiring) official Twitter accounts, including a broad set of metrics (productivity, age, Social Activity, number of followees, etc.), paying special attention to specific followers' features (age, gender, location, and bios text). The results show that Parker's twitter profile exhibits an overall higher impact, which denotes not only a different online strategy but also a high level of engagement and popularity. The low level of shared followers by Parker and Feiring (1,898 users) offer prima facie evidence of an online gap between these followers, which can indicate the existence of a divided group of supporters corresponding with the visions that Parker and Feiring represent. Finally, special features are notice for Feiring in gender (more women followers), language (more English-speaking followers) and country (more followers from the United States).

Keywords: *Robert Parker; Wine industry; Wine prescriptor; Webmetrics; Web data analysis; Twitter data.*

1. Introduction

The globalization of the wine world and the increase in the number of wineries and brands in recent decades has caused a great demand for information from consumers. The value of a product depends on the valuation of its different types of attributes, but not all of these attributes can be attained by the consumer for the same price (Compés, 2002). In the case of wine, the search attributes are the least expensive to know, since they consist of information that the consumer can find out directly before buying. The trusted attributes are the most complex, but the existence of Denominations of Origin and other third-party certifying entities allows them to be converted into search attributes. At that point, the attributes posing the greatest problems are the those of experience –taste, flavour–, which can be known directly, but only after having acquired and tested the product, which means that there is a risk, especially for wines with a higher price. A personal testing or the advice of a friend would eliminate or reduce this risk, but the consumer is not always able to taste the wines before buying them or have a trusted friend next to him at the time of purchase.

To avoid the market failures associated with this type of informational asymmetry, a powerful industry of information and specialized valuation has been created in the main wine markets of the world. This industry is formed by journalists, guides, apps, magazines and competitions. All of them, through their comments, ratings and distinctions, aim to guide the decisions of the consumers. When they are very influential, the wines come to bear their marks/labels (Orth & Krška, 2001), which convert the attributes of experience into attributes of search; wineries even try to elaborate wines that satisfy their tastes in order to obtain the best valuations and scores. In these cases, the administrators of the ratings become prescribers or gurus (Ali et al., 2008; Huang et al., 2009), and influence the choice of many consumers (Chocarro & Cortiñas, 2013, Hamerson, 2010).

Although there are several famous and respected critics including Jancis Robinson, Michael Broadbent, Steven Tanzer, Jean-Marc Quarin or James Suckling, the greatest prescriber in the world of wine is Robert Parker. He has been considered not only the most influential critic (Bowman, 2009) but even the Wine Emperor (McCoy, 2014). His influence in the world wine industry, exercised through his publication Robert Parker's Wine Advocate, has been extensive. It has influenced tasting notes (James, 2018), wine rhetoric (Hommerberg, 2011) and wine valuation criteria (Cardebat & Livat, 2016) including the famous "Parker points" (Shapin, 2005). His valuations have created a trend in wine production -parkerization- based on his tastes in favor for bold, fruity, and concentrated wines, especially in Bordeaux (Parker, 2003; Hay, 2010). Parker's vision and tastes have influence the wine industry, provoking a homogenization that goes against terroir and differentiation. Nonetheless, Parker is one of the reasons behind the great growth of the wine demand in the US (becoming the first world wine market to overcome France). The hegemony of Robert Parker and the threat of the consequent homogenization and industrialization of the wine world (Torrès, 2016) have provoked an opposite reaction, with one of its main critics being Alice Feiring (2008).

The objective of this paper is to determine to what extent Robert Parker has lost his influence as a prescriber in the world of wine. To do this, we investigated whether his Journal and other ways to disseminate his comments and points are being less followed, and if his critics have been found to be increasing their influence. As a preliminary attempt to carry out this goal, a webometrics approach (Thelwall, 2009; Orduna-Malea and Aguillo, 2015), in its business side (Orduna-Malea & Alonso-Alonso, 2017), is applied. In this specific case, Twitter data (Zimmer & Proferes, 2014) will be used to ascertain whether Parker's vision of wine is losing influence against his main critic Alice Feiring.

2. Method

In order to proceed with a comparison between the wine's world visions of Robert M. Parker Jr. (hereinafter, Parker) and Alice Feiring (hereinafter, Feiring), data from Twitter was gathered. The official Twitter accounts of Parker (@wine_advocate) and Feiring (@alicefeiring) were considered. Followerwonk (https://moz.com/followerwonk) data source was then used to extract the following metrics from each profile: Age (number of years since the creation of the Twitter account), Followers, Friends, Likes, and Social Authority (score from 0 to 100 reflecting the importance of each account). Additionally, a deep analysis of both Parker and Feiring's followers (17,128 and 14,927 followers respectively) was carried out, obtaining information about followers' language, precedence, and gender. Moreover, keywords included in the followers' Bios field were extracted and analysed by a text analyser (https://www.online-utility.org/text/analyzer.jsp) in order to find out the main common terms used by followers.

Additionally, queries in Google Trends (https://trends.google.com/trends) were performed to check the relevance of Parker's vision over time. The queries "The Wine Advocate" and "Robert M. Parker Jr." on the one hand, and "The Feiring Line" and "Alice Feiring") on the other were queried (location: Worldwide; source: Web search; timespan: since 2004).

All data gathering and queries were performed during the first week of March 2018.

3. Results

3.1. Web search trends

The relative search interest for "The Wine Advocate" has declined notably since 2004 (Figure 1). Furthermore, the results obtained for "Robert M. Parker Jr." reinforce a fall of interest in the "Parker's wine vision". However, this drop cannot be related with a rise of a "Feiring's wine vision". The query "Alice Feiring" cannot be included in Figure 1 since the relative search interest (if compared with Parker results) is below "1" for 51.5% (88) of the months measured, and just "0" for 25% (43) of the months. Supplementary queries such as "The Feiring line" provide few results as well.

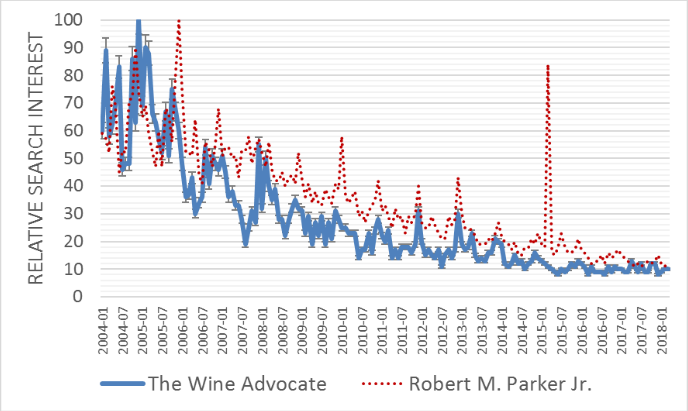


Figure 1. Search trends comparison over time.
Source: Google Trends (Worldwide; in the entire web; since 2004)

Regarding the geographical location of users, the query "Robert M Parker Jr." exhibits a greater breakthrough worldwide. Considering apart the special interest in English-speaking countries (United States, United Kingdom, Australia, South Africa), it's noteworthy the interest in Spain, France, Germany, and Latin America (Argentina and Chile). On the opposite, the query "Alice Feiring" exhibits a more restricted interest, focused mainly in the United States, France, and Spain.

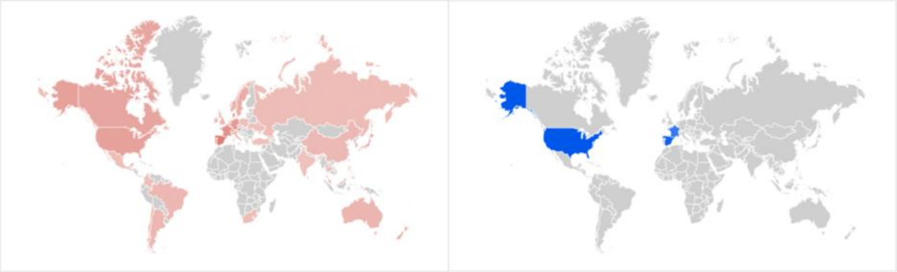


Figure 2. Geolocalized searches for "Robert M Parker Jr" (left) and "Alice Feiring" (right).
Source: Google Trends

3.2. Twitter profiles comparison

Parker obtains a higher number of followers (17,128), likes (4.523) and social authority (54) than Feiring (Table 1). This is quite remarkable since Parker's Twitter profile follows less friends (37), is younger (under 5 years) and is less productive (3,868 tweets), what does not prevent him from obtaining a greater percentage of ReTweets (32.5% against 21.5%). Conversely, Feiring's account is very productive (13,311 tweets published) and follows more users (1,203). Otherwise, the number of followers shared by the two accounts is low (1,898 followers). That is, only 11% of all Parker's followers follow Feiring.

Table 1. Principal metrics for Parker and Feiring's Twitter profiles.

Metrics	Parker	Feiring
Social Authority	54	46
Followers	17,128	14,927
Age	4.61 years	9.33 years
Tweets	3,869	13,311
Likes	4,523	3,105
Friends	37	1,203

Source: Followerwonk

The average social authority of Feiring's followers is slightly higher (17.23) than Parker's (16.93). The distribution of followers according to their social authority is uneven (Figure 3). Only 5.7% (968) of Parker's followers obtain a social authority score at least of 50. In a similar way, only 5.3% (790) of Feiring's followers achieve this score.

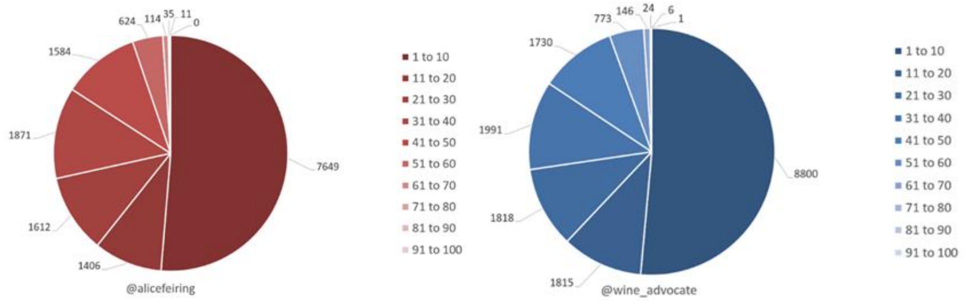


Figure 3. Distribution of followers' social authority.
Source: Followerwonk.

Gender

Despite the huge percentage of users to whom it was not possible to determine the gender (50.9% in the case of Parker, and 54.6% for Feiring), we observe a greater percentage of male followers in both profiles (Parker: 36.3%; Feiring: 27.2%). However, the percentage of female followers is greater in the case of Feiring (18.1%) than for Parker (12.7%).

Language

As regards the language used by followers, English predominates both for Parker (61.3% of followers) and for Feiring (83.9%), followed by Spanish (18.5% and 4.5% respectively), French (6.4% and 4.4%), and Italy (5.6% and 4%).

Bios keywords

The top 5 keywords included in the bios field of the followers are displayed in Table 2. Single words without enough meaning themselves (world, wine, food…), void terms (the, and, to…) and senseless composed terms (of the, the best, the wine...) were excluded. Likewise, variants were merged (eg. 'food and wine' and 'food & wine'). Although the most frequent terms are quite similar in both Twitter profiles, the higher frequency of Winery and Sommelier for Feiring, despite having a less number of followers is notewhorty.

Table 2. Keywords included in Bios of Twitter followers.

PARKER		**FEIRING**	
Keyword	**N**	**Keyword**	**N**
Winery	351	Winery	472
Sommelier	345	Sommelier	452
Food wine	231	Food wine	332
Fine wine	210	Wine food	282
Wine lover	186	Wine lover	157

Source: Followerwonk.

Geographical location

Two main zones in Europe (Paris and London) with a similar concentration of followers for both (Figure 4; top) are observed. Otherwise, a greater concentration of Parker's followers is observed in Spain. In the United States (Figure 4; bottom), the three main centres of interest correspond with the areas of New York, San Francisco and Los Angeles, where the concentration of followers is greater for Feiring. However, data should be taken cautiously since maps are plotted considering up to 5,000 Twitter users for each account (29.2% of all Parker's followers, and 33.5% of Feiring's).

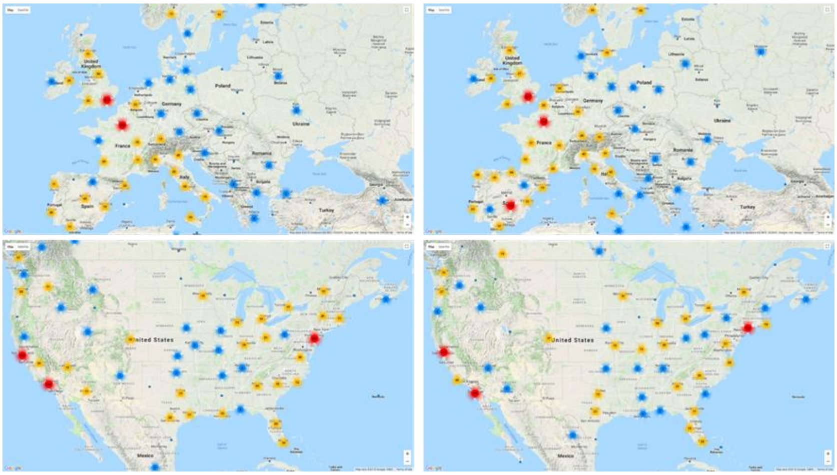


Figure 4. Geolocation of Feiring's followers (left) and Parker's (right) in Europe (up) and United States (below).

Raw data has been manually explored to contrast this approximation. In the case of Parker, geographical data is available only for the 67.4% (11,543) of followers, whereas for Feiring is 79.4% (11,849 followers). Even so, the percentage of useful data is still lower since users tend to full the geographical data field with invented terms ("Here", "somewhere", "At lunch"), imprecise locations or simply spam (i.e: "Contact us for promotions"). After clustering data at a country-level, Parker's followers come basically from 5 countries: USA (3,546), Spain (1,536), UK (1,092), France (832) and Italy (684). These same countries are the main places from which Feiring's followers come from, although with differences in their ranking positions: US (5,958) , UK (886), France (574), Italy (496), and Spain (410).

4. Discussion and conclusions

The results obtained find a decline in search interest for terms related both to "Robert Parker" and "The Wine Advocate". This may reflect a decline of 'Parker's vision of wine'. It can also be underlined that *The Wine Advocate* search goes down, during a long time, more than Robert Parker's. A feasible explanation may be due to the fact that 'Robert Parker' becomes a trademark most important than his mean. In any case, it is obvious that although Robert Parker has lost influence, he has created a methodology in the way of measuring and valuing a wine - the famous Parker points - that has been imitated by other critics and has educated thousands of consumers, and this is already part of his legacy.

However, this fall could be due also to the fact that users access alternative sources instead of personal websites. Precisely, the Twitter analysis does not find a Parker's alternative vision through Feiring. Parker's profile exhibits an overall higher impact, which denotes not only a different profile strategy but also a high level of engagement and popularity. The low level of shared followers could be indicative of different (almost antagonic) positions in the world of wine. However, a deeper analysis of followers' activities and preferences is still necessary. Yet, special features are notice for Feiring in gender (more women followers), language (stronger influence of English) and country (strong influence in the United States). Likewise, it seems that Feiring's followers are more related to wine professionals than Parker's, more pure consumers.

However, this results should be taken cautiously since the tools used may introduce a coverage bias. Other platforms (Facebook, Instagram, etc.) and techniques (opinion mining, link analysis, etc.) must complement this work equally. The identification of two positions (Parker and Anti-Parker) needs of more quantitative and qualitative methods as well. In this case, only Alice Feiring has been studied as a banner of the "anti-Parker" wine vision, it would be also useful to compare Robert Parker's influence with the other great wine critics, some of them editing also their own journals. Finally, it would be necessary to analyse Parker's tasting notes and the terms used to categorize the highest scoring wines, and compare them with the competition and its critics in order to get a deeper understanding of the movements - in terms of wine prescriptors, trends and fashions - in the market of wine.

References

Ali, H. H., Lecocq, S., & Visser, M. (2008). The impact of gurus: Parker grades and en primeur wine prices. *The Economic Journal*, 118(529), 158-173.

Bowman, S. (2009). Uprooting Robert Parker. *Gastronomica: The Journal of Critical Food Studies*, 9(1), 98-101.

Cardebat, J. M., & Livat, F. (2016). Wine experts' rating: a matter of taste?. *International Journal of Wine Business Research*, 28(1), 43-58.

Chocarro, R., & Cortiñas, M. (2013). The impact of expert opinion in consumer perception of wines. *International Journal of Wine Business Research*, 25(3), 227-248.

Compés, R. (2002). Atributos de confianza, normas y certificación: comparación de estándares para hortalizas. *Economía agraria y recursos naturales*, 2(1), 115-130.

Feiring, A. (2008). *The Battle for Wine and Love: Or How I Saved the World from Parkerization*. Orlando (FL): *Houghton Mifflin Harcourt*.

Hapin, S. (2005). Hedonistic fruit bombs. *London Review of Books*, 27(3), 30-32.

Hay, C. (2010). The political economy of price and status formation in the Bordeaux en primeur market: the role of wine critics as rating agencies. *Socio-economic review*, 8(4), 685-707.

Hommerberg, C. (2011). *Persuasiveness in the discourse of wine: The rhetoric of Robert Parker* (Doctoral dissertation). Växjö (Sweden): *Linnaeus University Press*.

Huang, P., Lurie, N. H., & Mitra, S. (2009). Searching for experience on the web: an empirical examination of consumer behavior for search and experience goods. *Journal of marketing*, 73(2), 55-69.

Jamerson, H. M. (2010). *Wine Tastes: The Production of Culture among Service Workers and Consumers in Napa Valley Wineries* (doctoral dissertation). Atlanta: *Emory University*.

James, A. (2018). How Robert Parker's 90+ and Ann Noble's Aroma Wheel Changed the Discourse of Wine Tasting Notes. ILCEA. *Revue de l'Institut des langues et cultures d'Europe, Amérique, Afrique, Asie et Australie*, (31), Online.

McCoy, E. (2014). *The Emperor of Wine: The Rise of Robert M. Parker, Jr., and the Reign of American Taste*. New York: *Harper Collins*.

Orth, U. R., & Krška, P. (2001). Quality signals in wine marketing: the role of exhibition awards. *The International Food and Agribusiness Management Review*, 4(4), 385-397.

Parker, R. M., & Rovani, P. A. (2002). *Parker's wine buyer's guide*. New York: *Simon and Schuster*.

Parker, R. M. (2003). *Bordeaux: a consumer's guide to the world's finest wines*. New York: *Simon and Schuster*.

Thelwall, M. (2009). *Introduction to webometrics: Quantitative web research for the social sciences*. San Rafael (CA): Morgan & Claypool.

Torrès, O. (2006). *Introduction: The McDonaldization of Wine*. London: *Palgrave Macmillan*.

Orduña-Malea, E., & Aguillo, I. F. (2015). Cibermetría. *Midiendo el espacio red*. Barcelona: UOC Publishing.

Orduña-Malea, E., & Alonso-Arroyo, A. (2017). Cybermetric Techniques to Evaluate Organizations Using Web-based Data. Oxford: *Chandos Publishing*.

Zimmer , M. & Proferes, N. J. (2014). A topology of Twitter research: disciplines, methods, and ethics". *Aslib Journal of Information Management*, 66(3), 250–261.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8327

Digital Vapor Trails: Using Website Behavior to Nowcast Entrepreneurial Activity

Slaper, Timothy F.[a]; Bianco, Alyssa[b] and Lenz, Peter E.[c]
[a]Indiana Business Research Center, Indiana University, United States, [b]Dstillery, United States, [c]Dstillery, United States.

Abstract

Following recent research, we explore virtually contemporaneous, and geographically granular, user online activity related to entrepreneurship. In this paper, we present evidence that data harvested by Dstillery can complement efforts of, and data collected by, government agencies and organizations advocating for entrepreneurship, business formation and economic growth, e.g., the Kauffman Foundation. Our website-based behavior data is close to real time and at a geographically granular level. We find that the concentration of a region's visits to website resources for entrepreneurship and business development are statistically related to business start-up and, particularly, growth activity. Visits to websites related to entrepreneurship are more strongly associated with growth entrepreneurship, in contrast to start-up entrepreneurship. While data capture and analysis related to entrepreneurship website activity is in its infancy, this analysis points to the potential of this data source to nowcast business formation and growth at a regional level.

Keywords: *Nowcasting; entrepreneurship; start-ups; business formation; website behavior.*

1. Introduction

Data based on website activity—or, in this case, the behavior of economic agents—raise the possibility of enabling researchers and policymakers with a means to augment, and make more timely, official government statistics at a geographically granular level. Recently, Google Trends has gained increasing popularity in the practice of "nowcasting" economic variables, as well as other social and health outcomes. Google Trends data have been used to predict the outbreaks and spread of disease (Carneiro & Mylonakis, 2009), tourist flows (Siliverstovs & Wochner, 2018), consumer behaviors (Vosen & Schmidt, 2012), and unemployment (Askitas & Zimmermann, 2009; Pavlicek & Kristoufek, 2015; Naccarato, Pierini, & Falorsi, 2015; Vicente, López-Menéndez, & Pérez, 2015; D'Amuri & Marcucci, 2017).

In short, research on the potential of using digital vapor trails to predict economic outcomes, or nowcast economic activity, is growing (e.g., Choi & Varian, 2012; Wu & Brynjolfsson, 2015; Goel, Hofman, Lahaie, Pennock, & Watts, 2010). As noted by Glaeser, Kim, and Luca (2017), online data sources make it possible to measure new activities and outcomes that, heretofore, were outside the scope of official, traditional data sources.

In this paper, we explore virtually contemporaneous and geographically granular user online activity related to entrepreneurship. We present evidence that data harvested by Dstillery can complement efforts of, and data collected by, government agencies and organizations advocating for entrepreneurship, business formation and economic growth—such as the Kauffman Foundation—using a proxy measure of entrepreneurial activity. We extend the existing nowcasting literature based on digital vapor trails by aligning data new to economic research, namely from Dstillery, with data developed by an entrepreneurship advocacy organization—the Kauffman Foundation. Our key contribution is to evaluate the potential (potential because capturing the data is still emerging) of whether website behavior data tracks with time-tested and traditional measures of business formation activity at a regional geographic scale, based on metropolitan statistical areas (MSAs).

2. Data and Method

For over two decades, the Kauffman Foundation has published the Kauffman Index on Entrepreneurship. In many ways, it is the state-of-the-art in measuring business start-ups, formation and growth (Fairlie, Morelix, Reedy, & Russell, 2015; Kauffman Foundation, 2017a, 2017b). Over time, and with much research and analysis, the index differentiated start-up activity and early business growth activity. In other words, what may signal the creation of a new business is not the same as the signals that a new business is growing and prospering. Given the time and research focus devoted to developing the three measures for the two entrepreneurial categories—start-up and growth—we embrace the six measures that comprise the two indexes as the best available. These data are available for the top 40 metro areas in the country. The Kauffman Foundation does not publish an index for other metro areas.

While Dstillery is well established in the digital marketing ecosystem, the application of their data linking individual website behavior to economic activity is novel. Dstillery is a predictive marketing intelligence firm that anonymously collects, classifies, and disseminates behavioral data. Dstillery's digital data is collected as a proxy for real-world behaviors. Marketers use these data to optimize which devices see ads at times relevant to a consumer (Raeder, Stitelman, Dalessandro, Perlich, & Provost, 2012).

Dstillery can create an analytical category based on the websites a marketer or researcher may think pertains to a particular constituency. Dstillery captures a representative sample of devices that visit these websites and finds the highest scoring features to create a model that scores devices based on their affinity to the behavior of interest (Ibarra & Lenz, 2016), in this case the E-ship audience. The score of that set of devices is compared to a random sample of devices across the internet. Device scores are aggregated at the ZIP code level, based on the predicted "home" location of each device according to a probabilistic model that takes into account time of day and frequency of visit to a discrete location.

The Indiana Business Research Center team provided a list of relevant organizations and websites from the Kauffman Foundation website, plus several entrepreneurial website guides, to generate a list of 100 websites for Dstillery to observe and measure traffic. These data for E-ship affinity was, in turn, translated or aggregated from ZIP code geographic units of analysis into MSAs. Concentration of activity values were scaled by the relative population of each ZIP code within an MSA. (Unscaled data yields results that were less robust.)

The first tranche of Dstillery data for entrepreneurship—E-ship—from the website list was captured in January 2018. These data are new and exploratory. The Kauffman Index data are from the latest iteration of the index (2017), but some of the data used to calculate the index are based on business formation relationships, or data, from 2014. We assume one important regional characteristic: A region’s internet and website behavior in the late months of 2017 is consistent with that region’s experience in 2016 and years previous. That is, people’s tastes and interests don’t change dramatically in a place/location over time unless there is some titanic event. Another way to view this assumption is that a region’s culture does not change quickly; thus, a region’s propensity to express interest in activities, entrepreneurial or not, will not change dramatically compared to other regions over the course of a few years.

One may view this as a corollary of the work of Obschonka et al. (2015): Personality profiles of a region can help explain entrepreneurial activity. One would not expect that the psychological cultural characteristics of a region would change significantly from one year to the next.

3. Results

Because these data are so new—Dstillery only recently started to capture E-ship data—we only have one snapshot of E-ship–related web behavior. (Over time, however, Dstillery is expected to collect these so that researchers will be able to track E-ship concentration changes over time and how they relate or possibly predict start-up and entrepreneurship growth.)

First, we performed a simple correlation between the three sub-measures of the start-up activity and the entrepreneurship growth indexes—six component measures in all (namely the rate, the share and the density) and the Dstillery E-ship measure for website traffic for the 40 metropolitan statistical areas covered by the Kauffman Index data. (Please refer to the Kauffman Foundation reports (2017a, 2017b) for more detail on what the rate, share and density measures capture.)

The correlations are not particularly strong, but there are interesting differences in the relationship between Dstillery E-ship data and the two types of entrepreneurial activity tracked by the Kauffman Foundation. For the start-up index, the correlations are 0.32, 0.00 and 0.19 for rate, share and density components, respectively. The correlations for the entrepreneurial growth index components are 0.49, 0.27 and 0.42 for rate, share and density, respectively. This would indicate that entrepreneurs in the growth phase of their new companies utilize web-based resources to a much greater extent to help them grow their businesses, in contrast to the utilization of entrepreneurs who are just getting started.

We then considered the relationship between Dstillery E-ship and the higher correlated index measures for both start-up rate, growth rate and growth density. We used a simple OLS model to assess the degree to which the variation in the index component values for these three concepts may be explained by the regional/MSA concentration of E-ship website traffic as captured by Dstillery. The explained variation—adjusted R-squares—for the three dependent variables were not particularly strong: Start-up Rate of New Entrepreneurs – 0.08; Rate of Startup Growth – 0.22; and High-Growth Company Density – 0.16. The coefficients are positive with p-values of 0.048, 0.001 and 0.007, respectively.

The modest association suggested by the statistical results is both good news and bad news. That E-ship web traffic can provide some explanatory power across metropolitan areas suggests that this concept and measure may be a good candidate to use as one piece of an entrepreneurship nowcasting data set and deserves further study and development. Part and parcel of that development would be to ensure that the E-ship website universe is complete and captures all relevant web-based resources. The bad news is that this simple exercise has compared differences in two snapshots of metro areas. It doesn't measure change over time. Moreover, the foundational characteristics of each region has not been explored sufficiently in this analysis. The Kauffman measures treat each start-up the same, whether a food truck or a medical testing clinic or a high-tech systems integration company. Arguably, those regions that dominate in the digital and technology space would also have greater connectivity and more devices (digital reach) and would likely have a greater proportion of their denizens who would access web-based resources to gain knowledge, know-how or seek enterprise funding. Regional industry characteristics and opportunities may explain many behaviors.

4. Conclusion

In this paper, we have used a novel data set to test the hypothesis that regional differences in E-ship–related web traffic may help to explain differences in regional business start-ups and new business growth. The data are so novel, that there is only one snapshot for the concentration of E-ship–related website usage. We compared two snapshots: the Kauffman Index of Entrepreneurship Activity and Dstillery E-ship data for 40 metropolitan areas in the United States. We found that E-ship web activity is more closely associated with growth entrepreneurship than with start-ups. While this difference may be attributed to differences in regional characteristics across metro areas, it may also signal that those in the start-up phase of creating a new business are not as inclined to utilize online resources to learn how to run a business or expand their knowledge base. If predominantly the latter, these findings may point to the need for a policy or resource response to better serve those in the very early stages of starting a business.

Given the one-off nature of the analysis, we cannot currently advocate for the Dstillery E-ship data to be considered a viable data source for nowcasting entrepreneurship and business formation. That said, as the data series is captured over time, these data may be valuable for policymakers, economic development practitioners and even government economic statisticians to watch in the future.

References

Askitas, N., & Zimmermann, K. (2009). *Google econometrics and unemployment forecasting* (DIW Berlin Discussion Paper 899). Berlin: German Institute for Economic Research. doi:10.2139/ssrn.1465341

Baker, S. R., & Fradkin, A. (2017). The impact of unemployment insurance on job search: Evidence from Google search data. *The Review of Economics and Statistics, 99*(5), 756–768.

Carneiro, H. A., & Mylonakis, E. (2009). Google Trends: A web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases, 49*(10), 1557–1564.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record, 88*(s1), 2–9.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting, 33*(4), 801–816.

Fairlie, R. W., Morelix, A., Reedy, E. J., & Russell, J. (2015). The Kauffman Index 2015: Startup activity national trends. Kansas City, MO. doi: 10.2139/ssrn.2613479

Glaeser, E. L., Kim, H., & Luca, M. (2017). *Nowcasting the local economy: Using Yelp data to measure economic activity* (NBER Working Paper 24010). Cambridge, MA: National Bureau of Economic Research.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences, 107*(41), 17486-17490.

Ibarra, P. & Lenz, P. (2016, March). Using digital signals to measure audience brand engagement at major sports events: The 2015 MLB season. Paper presented at *MIT Sloan Sports Analytics Conference*, Boston.

Kauffman Foundation (2017a). Kauffman Index 2017: Growth entrepreneurship metropolitan area and city trends. Kansas City, MO. doi:10.2139/ssrn.3080714

Kauffman Foundation (2017b). 2017 Kauffman Index of Startup Activity: Metropolitan area and city trends. Kansas City, MO. doi:10.2139/ssrn.2974544

Naccarato, A., Pierini, A., & Falorsi, S. (2015). *Using Google Trend data to predict the Italian unemployment rate* (Department of Economics Working Paper 203). Rome: University Roma Tre.

Obschonka, M., Stuetzer, M., Gosling, S. D., Rentfrow, P. J., Lamb, M. E., Potter, J., & Audretsch, D. B. (2015). Entrepreneurial regions: Do macro-psychological cultural characteristics of regions help solve the "knowledge paradox" of economics? *PLOS ONE*, *10*(6). doi:10.1371/journal.pone.0129332

Pavlicek, J., & Kristoufek, L. (2015). Nowcasting unemployment rates with Google searches: Evidence from the Visegrad Group countries. *PLOS ONE, 10*(5). doi:10.1371/journal.pone.0127084

Raeder, T., Stitelman, O., Dalessandro, B., Perlich, C., & Provost, F. (2012). Design principles of massive, robust prediction systems. *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD

'12), 1357-1365. New York: Association for Computing Machinery. doi:10.1145/2339530.2339740

Siliverstovs, B., & Wochner, D. S. (2018). Google Trends and reality: Do the proportions match? Appraising the informational value of online search behavior: Evidence from Swiss tourism regions. *Journal of Economic Behavior & Organization 145*, 1–23.

Vicente, M. R., López-Menéndez, A. J., & Pérez, R. (2015). Forecasting unemployment with internet search data: Does it help to improve predictions when job destruction is skyrocketing? *Technological Forecasting and Social Change, 92*(Supplement C), 132–139.

Vosen, S., & Schmidt, T. (2012). A monthly consumption indicator for Germany based on Internet search query data. *Applied Economics Letters, 19*(7), 683–687.

Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In A. Goldfarb, S. M. Greenstein, & C. E. Tucker (Eds.), *Economic Analysis of the Digital Economy* (pp. 89-118). Chicago: University of Chicago Press.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8329

Evolution and scientific visualization of Machine learning field

Río-Belver, Rosa[a]; Garechana, Gaizka[a] , Bildosola, Iñaki[a] and Zarrabeitia, Enara[a]
[a] Technology, Foresight and Management Research Group, Departamento de Organización de Empresas, Universidad del Pais Vasco UPV/EHU, Spain.

Abstract

This article provides a retrospective and understanding of the development of automatic learning methods. The beginnings are visualized as a discipline within Computer Sciences in the subcategory of Artificial Intelligence, its development and the current transfer of knowledge to other areas of Engineering and its industrial applications. Based on the publications about machine learning and its application contained in the Web of Science database, records from 1986 to 2017 are downloaded. After a description of the technological profile, a new approach is introduced to the classification of a discipline based on the year of appearance of those terms that define it. Mining of technological texts and network theory has been applied to extract the terms and interpret their evolution. They are the those that define the stages of emergence, development and maturation of the discipline Machine learning. The novelty of this approach lies in the technical nature of applied research in Machine Learning, which aims to be a guide for the development of future engineering applications and to make technology transfer to industry visible.

Keywords: *Machine Learning; Tech-Mining; Scientometrics; Social Network Analysis; Visualization; Bibliometrics*

1. Introduction

Industry 4.0 is generating an unprecedented revolution in the manufacturing sector, greatly favored by the Internet of Things (IoT), the growth of Big Data and the sensorization of machines. According to the OECD, Peña Lopez (2015), the new industrial revolution is understood as the incorporation of fundamentally digital technologies conducive to achieving the smart factory.

The application of automatic learning methods defined by Alpaydin (2014) to industrial production will be one of the pillars of the new revolution. Decision-making must be decentralized and productive systems should have the ability to make basic decisions and become as autonomous as possible.

Understanding the keys to the development of the machine learning discipline makes it possible to understand the transfer process from algorithm development in the laboratory to machine programming in industry. To study a scientific discipline we have to refer to methodologies developed by Alejo (2015), Garechana et al. (2014), Gore et al. (2016), Noyons (2009), Porter et al. (2011), which shows how text-mining methods, natural language processing and network theory collaborate to produce indicators. The algorithms developed within Machine Learning are also helping the development of bibliometrics, Ranei (2017).

This article is arranged in four sections. After the introduction, the second section describes both the methodology followed and the composition of the sample to be analyzed. Next, the analyses leading to the elaboration of the technological profile, the analysis of the field topics and the visualization of the networks are carried out, finishing with the conclusions and future lines of work.

2. Method

The method followed in the preparation of this study is explained through the block diagram shown in figure 1.

The data for the analysis of the scientific field of Machine Learning and its applications were obtained from the core collection of Web of Science database. WOS is an online platform belonging to the company Clarivate Analytics that provides access to the world's leading citation databases, with multidisciplinary information from over 18,000 high impact journals, over 180,000 conference proceedings, and over 80,000 books from around the world.

The Machine learning and applications (MLAA) data set has been obtained by retrieving the articles, conference proceedings and book chapters published from 1986 to 2017. The concepts (“Machine learning”) AND (Application*) have been searched in the fields Title, Abstract, Author Keywords and Keywords plus, detecting 14805 records that were downloaded in their full record format.

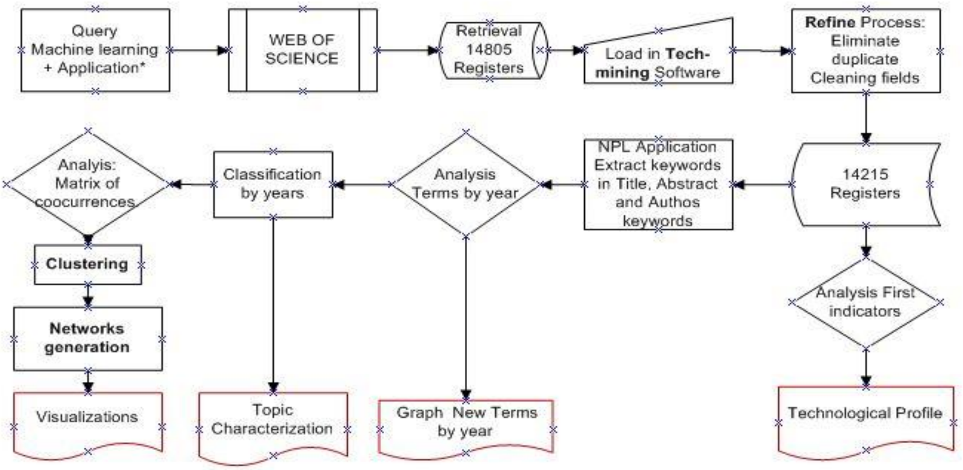

Query Machine learning + Application*
WEB OF SCIENCE
Retrieval 14805 Registers
Load in Tech-mining Software
Refine Process: Eliminate duplicate Cleaning fields
14215 Registers
NPL Application Extract keywords in Title, Abstract and Authos keywords
Analysis Terms by year
Classification by years
Analyis: Matrix of coocurrences
Clustering
Networks generation
Visualizations
Topic Characterization
Graph New Terms by year
Analysis First indicators
Technological Profile

Figure 1. Method of analysis . Source: own elaboration(2018).

After the refining and cleaning processes, the records to be analyzed were reduced to 14215. To determine the life cycle of the machine learning discipline and define the temporal classification, the terms and phrases of the Title, Abstract and Author Keyword fields have been extracted using NPL. These Terms were refined and cleaned for further analysis.

A Term data set is created with the terms defined by the author themself, as they are usually ahead of the thesaurus classifications of the journals themselves. If the data extracted from the abstract and title fields is also added, it produces a good data set of terms that define the discipline.

Next, the analysis and detection of its first year of appearance is carried out, highlighting the evolution of the discipline. Once the temporality had been defined, three databases were created, one for each previously defined period.

For each sub-data set, autocorrelation maps of the Web of Science categories field were created to visualize the network structure and the strength of the connections between the nodes. The generated networks were visualized with the support of VosViewer software.

3. Analysis

3.1 Technological Profile

The countries with the highest number of publications in the field are the USA (4270), China (2002), UK (1124), Germany (914), India (820), Canada (593)...and so on. However, to analyze the technological profile focus should not be placed on the number of publications but rather on the most relevant ones in the field. To analyze the impact we have to study the number of citations of the articles.

In the case of Machine Learning and Application, a total of fifteen publications account for 49.54% of all citations. A single article in the sample receives 10390 citations, representing 12% of all citations. This is the article LIBSVM: A library for support vector machines written by Chang, C., & Lin, C. in Taiwan in 2011. Of the remaining fourteen, twelve are led and/or co-authored by the United States but we have to wait until 2007 to see highly cited publications led by other countries such as Spain, China and Taiwan.

The terms defined by the authors in the fifteen most cited articles of the analyzed discipline are as follows: NEURAL NETWORKS, TEXT CATEGORIZATION, SUPPORT VECTOR MACHINES, LANDSCAPE, MODELS, SPECIES DISTRIBUTION, GENE-EXPRESSION DATA, NEURAL-NETWORKS, COMPONENT ANALYSIS, K-MEANS ALGORITHM, PATTERN-RECOGNITION, HIDDEN MARKOV-MODELS

All highly cited articles belong to the WOS Computer Sciences Artificial Intelligence or Computer Sciences Information Systems Category.

3.2. Topic characterization

Text mining allows us to apply text classification to solve the categorization problems of a discipline. The most representative terms are extracted from the keywords defined by the author themself, to which the words and phrases identified in the title and abstract fields are added. The natural language processing application of Vantage point data mining software is used for this purpose. Once cleaned using fuzzy filters, 22181 terms are available. This number is reduced to 2612 by discarding the terms whose frequency of appearance is less than 3 in order to apply a macro that determines the first year that the term appeared.

As shown in Figure 2, most of the terms are used for the first time between 1997 and 2017, with a peak generation of 203 new terms in 2009, after which time the generation of new terms stagnated and began to decline in 2014. However, the number of records that include these terms has grown since 2000 and in 2016 the maximum of the series is published, at 2356 records.

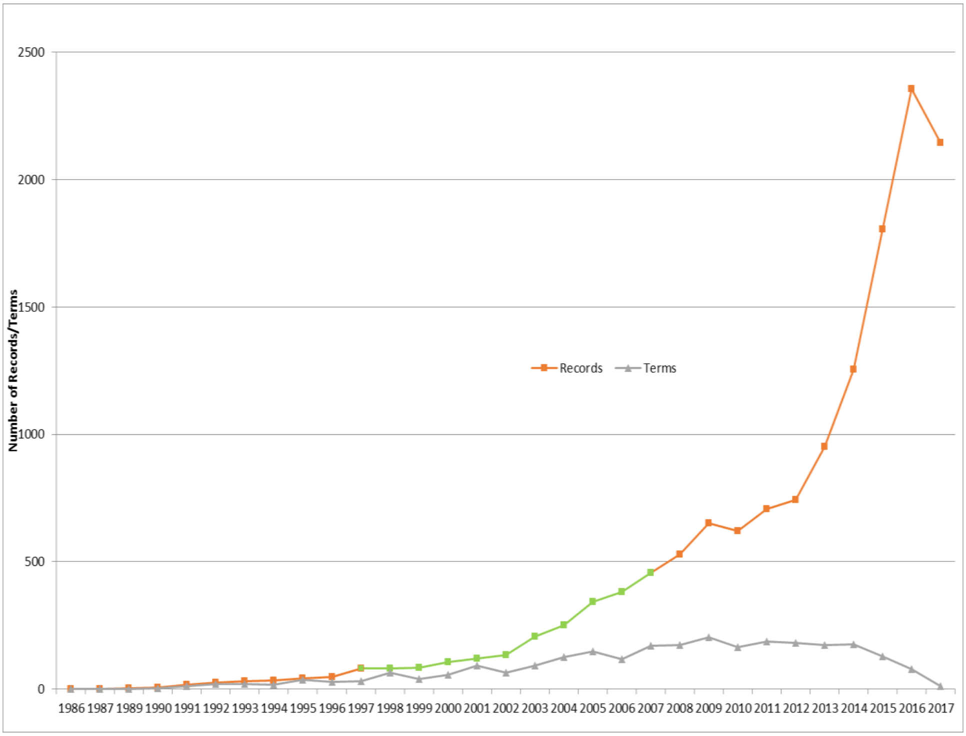


Figure 2. Number of new Author Keywords any year versus the number of records of that year. Source: own elaboration(2018).

If we carry out an analysis of the new terms that arise every year we can say that in 1990 the four terms that appear for the first time are: Machine learning (later repeated in Author keyword field 4236 times), Expert system (38), knowledge acquisition (14) and Knowledge base (14).

In 1997, 32 new terms appeared for the first time in the author's words, such as Data mining (515), Regression (72), Evolutionary algorithm (45), Rule extraction (14), Graph theory (14), fuzzy clustering (12), times series prediction (11), Fuzzy system (10) or Neural net (9). Later in 2007, 169 new terms were generated in the author's words such as: wireless sensor network (50), affective computing (37), Machine Supervised Learning (34) artificial neural network (ANN) (20), machine learning application (19). Finally, in 2017, only 11 new terms were generated, including Landslides (5), Precision medicine (5), and Age prediction (3). Based on the development of the terms, the evolution of the Machine Learning field is divided into three stages: Emergence stage from 1986 to 1996, development stage from 1997 to 2006 and maturation stage from 2007 to 2017. As the technology matures into its own field, the number of new terms each year is shrinked.

3.3 Networks and visualizations

It is considered appropriate to approach the scientific field through the use of the categories assigned by the Web of Science (WOS) to publications (papers, proceedings or book chapters), Leydersdoff (2013). All books and journals included in the Web of Science Core Collection, the leading provider of scientific and technological publications, which includes references to leading scientific publications in any discipline of knowledge since 1945, are assigned at least one of the 242 subject categories predefined by Clarivate Analytics. This makes it possible to determine the scientific classification of the document.

On the other hand, technology maps have the potential to become fundamental tools in science policy planning and therefore in the innovative development of a country. However, their interpretation is difficult because they are complex structures whose representation is difficult to interpret. In figure 3, on the right side, we can see the tentacles of the category called Computer science Artificial Intelligence, the scientific category father of machine learning for the period 1986-1996. In the upper left corner we can see a network composed of nodes, WOS categories, which are connected by means of edges. The strength of the line represents the number of records in the line, the stronger the representation the more common records between the nodes. In the period 1986-1996, there were 205 publications categorized into 66 items and 132 edges. This is a relatively low number and, as can be seen, the main collaborations are carried out between the same science; Computer Sciences Artificial, C.S. information, C. S. interdisciplinary, CS Cybernetics, although there are tenuous connections with Information Sciences, Automation Control systems and Electrical Engineering. This is an EMERGENCE STAGE, where science is developing and focusing on itself.

In the subsequent period 1997-2006, the central part of Figure 3, is seen as the network grows and doubles the number of nodes, WOS Categories , reaching 133. The records from this period date back to 1788, so that more connections and edges are generated (563). Computer Sciences Artificial Intelligence maintains its central position in the network, however its relationships are extended to Medical, Bioinformatics, Biotechnology, Imaging Science photographic, Business Finance, Neuroscience, Biochemical Research methods,… We can define it as a DEVELOPMENT STAGE. Keywords at this time include terms such as: supervised learning; Bayesian decision theory; parametric, semi-parametric, and nonparametric methods; multivariate analysis; hidden Markov models; reinforcement learning; kernel machines; graphical models; Bayesian estimation; and statistical testing.

Finally, in the last ten years, 2007-2017, the network has become too extensive. This is a MATURATION STAGE where most of the 12222 records are generated and the nodes almost double again, 216 nodes and 1295 edges. The area expands to almost all WOS categories (216/242). The graph on the right shows how its position has changed, Computer Sciences Artificial Intelligence has lost its central position in the network, no longer domineering the game but remaining as a base. Areas of major applicability such as Industrial Engineering, Electrical Engineering, Robotics, Material Sciences, Nanosciences, Biomedical Engineering, Optics, Instrumentation... show their clear progress in the life cycle of science. The following figures can be see better in https://doi.org/10.6084/m9.figshare.6302039.

4. Conclusion and future work

The application of text mining techniques combined with visualizations allows us to understand and interpret the evolution of a scientific discipline. Machine learning was born in the heart of Computer Sciences as a subdiscipline of Artificial Intelligence and has few links with other areas. From 1997 to 2006 it began to grow and branch out, connecting other areas of CS. From 2007 to 2017 we can see how the CS category branches out, goes beyond its own scope and expands into areas of applied techniques. In the future, it will become cross-cutting knowledge, as using examples from past experiences has been the basis for problem solving. The change is driven by the generation of large amounts of data on past experience and the high capacity of processors to process them and therefore the generation of solutions and automatic outputs that move the system forward.

For the generation of new terms the number of publications increased until 2009, when it stagnated and began to decrease, however, this is the period of greatest number of publications due to the strong appearance of Industrial Engineering and related areas. As a future extension of this study, WOS data will be combined with patent databases and the flows generated through the non-patent literature collected in the industrial property registers will be analyzed. The aim is to highlight, from various points of view, the current incorporation of Machine Learning and its collaboration in the Industrial Organization 4.0.

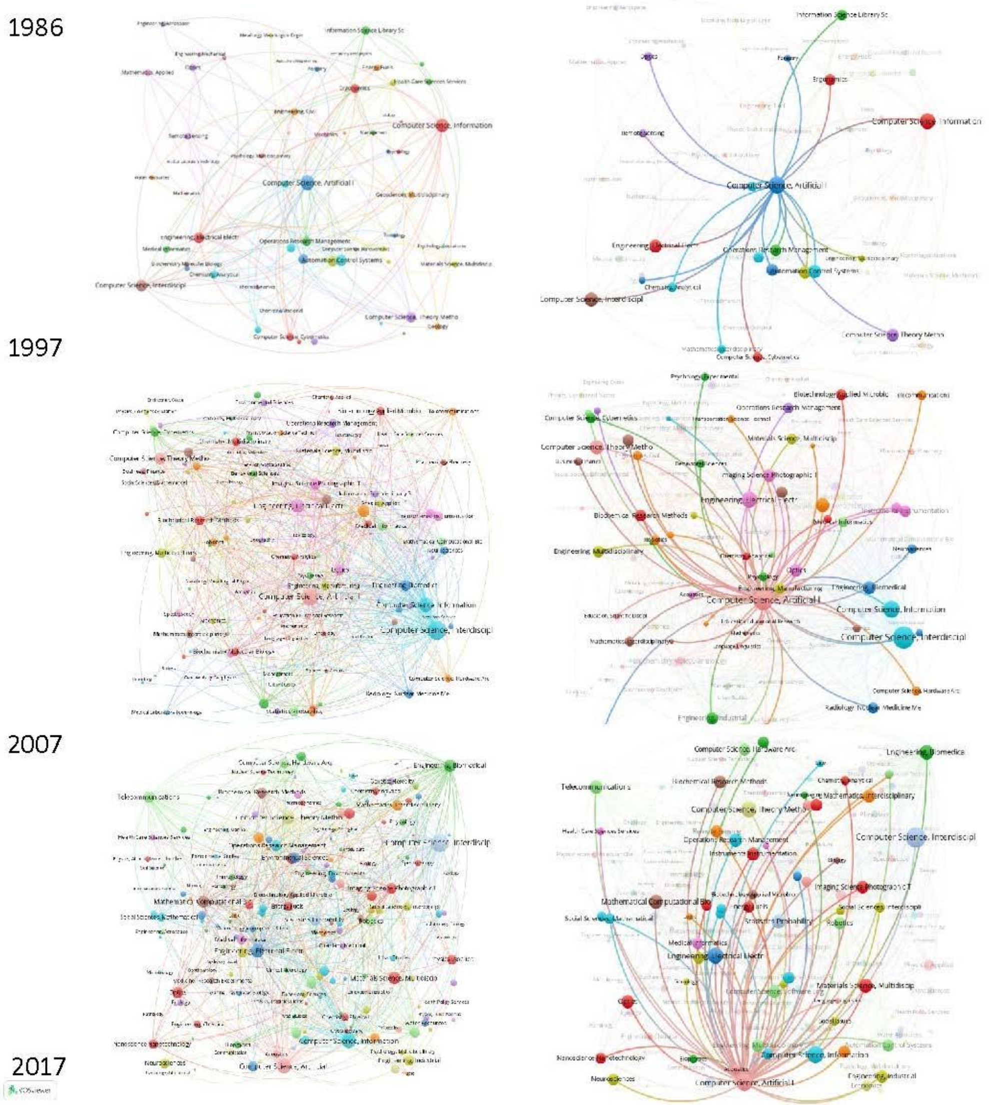

1986
1997
2007
2017
Information Science Library Sc
Computer Science, Information
Computer Science, Artificial I
Operations Research Management
Automation Control Systems
Engineering, Electrical Electr
Computer Science, Interdiscipl
Computer Science, Theory Metho
Computer Science, Cybernetics
Chemistry, Analytical
Ergonomics
Remote Sensing
Optics
Psychology, Experimental
Biotechnology Applied Microbio
Telecommunications
Materials Science, Multidiscip
Imaging Science Photographic T
Biochemical Research Methods
Engineering, Multidisciplinary
Engineering, Biomedical
Computer Science, Hardware Arc
Radiology, Nuclear Medicine Me
Engineering, Industrial
Mathematics, Interdisciplinary
Mathematical Computational Bio
Social Sciences, Mathematical
Statistics Probability
Robotics
Medical Informatics
Nanoscience Nanotechnology
Neurosciences
Health Care Sciences Services
Social Sciences, Interdiscipli
VOSviewer

Figure 3. Machine learning evolution in the Web Science Categories. Source: own elaboration(2018).

References

Alejo-Machado, O. J., Manuel Fernandez-Luna, J., & Huete, J. F. (2015). Bibliometric study of the scientific research on "learning to rank" between 2000 and 2013. Scientometrics, 102(2), 1669-1686.

Alpaydin, E. (2014). Introduction to machine learning. Cambridge: The MIT Press.

Garechana, G., Rio-Belver, R., Cilleruelo, E., & Larruscain J. (2014). Clusterization and mapping of waste recycling science. evolution of research from 2002 to 2012. Journal of the Association for Information Science and Technology, 66, 1431-1446.

Gore, R., Diallo, S., & Padilla, J. (2016). Classifying modeling and simulation as a scientific discipline. Scientometrics, 109(2), 615-628.

Leydesdorff, L., Carley, S., & Rafols, I. (2013). Global maps of science based on the new web-of-science categories. Scientometrics, 94(2), 589-593.

Noyons, E. C. M., & Calero-Medina, C. (2009). Applying bibliometric mapping in a high level science policy context. Scientometrics, 79(2), 261-275.

Peña-López, I. (2015). OECD digital economy outlook 2015. On line.

Porter, A. L., Guo, Y., & Chiavatta, D. (2011). Tech mining: Text mining and visualization tools, as applied to nanoenhanced solar cells. Wiley Interdisciplinary Reviews-Data Mining and Knowledge Discovery, 1(2), 172-181.

Ranaei, S., & Suominen, A. (2017). Using machine learning approaches to identify emergence: Case of vehicle related patent data. 2017 Portland International Conference on Management of Engineering and Technology (Picmet),1-8.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8331

Fishing for Errors in an Ocean Rather than a Pond

Wilson, John [a] and Te'eni, Dov [b]

[a]Ivey Business School, University of Western Ontario, London N6G ON1, Canada, [b]Coller School of Management, Tel Aviv University, Israel.

Abstract

In the internet age, a proliferation of services appear on the web. Errors in using the internet service or app are dynamically introduced as new devices/interfaces/software are produced and are found to be incompatible with an app that is perfectly good for other devices. The number of users who can detect various errors changes dynamically: for instance, there may be new adopters of the software over time. It may also happen that an old user might upgrade and thus run into new incompatibility errors. Allowing new users and errors to enter dynamically poses considerable modeling and estimation difficulties. In the era of Big Data, methods for dynamically updating as new observations arise are important. Traditional models for detecting errors have generally assumed a finite number of errors. We provide a general model that allows for a procedure for finding maximum likelihood estimators of key parameters where the number of errors and the number of users can change.

Keywords: *Errors in software apps; Big Data; reliability; software development and testing.*

1. Introduction

The internet and mobile have changed the way that software is distributed and used. Cloud computing, open source software and continuous connectivity, in particular, allow for the operation and connection of many services, many applications, many devices and many users. The advent of the Internet of Things will magnify the criticality of securing uninterrupted operation and connectivity. As early as the 1990s, there was a recognized need for research in the economics of software development and maintenance (e.g. Banker et al. 1998 and Chan et al.1994). Issues such as the timing of software releases, the development and management of interoperability, and the allocation of resources to testing are critical to management. Once software is released, it is important to have models that track errors as software is being used. Real-time detection of errors encountered by users is often the norm. There is a need, therefore, for research that builds on the new realities of the software industry and that leads to practical tools.

We concentrate on the probability of errors in software, which must be a parameter of any managerial model and can directly affect managerial decisions such as when to stop software testing and how to it. We are interested in the behavior of errors over time because software management is dynamic.

Finding errors is like fishing and open computer systems are analogous to a pond linked to the ocean. Our proposed model describes fishing in an ocean rather than the pond of previous models. In a pond, the rate of catching fish depends on how many are left in it. When the pond is opened to an ocean, waves bring in new fish and will not find all the fish when confronted with a practically infinite stream of fish coming in from the ocean. The potential number of failures due to communication software, printers, operating systems and I/O devices is practically infinite. In the new context of open systems, we distinguish between *errors of content* and *errors of incompatibility*. The former are code contained in the system that is incorrect with respect to its specification (e.g., an incorrect loop or a select construct that does not cover all required cases). The latter are code that is incompatible with conditions external to the specified system, e.g., problems in working parallel to new versions of other packages.

In this paper, our focus is the interaction between the system and its environment: hence the notion of incompatibilities. This shift has already occurred in industry. For example, Mercury Interactive, a software testing company, realized back in 1998 that current software, unlike the past, cannot be contained in a single system (see Forbes, 1998). A similar shift is needed in the operations research modeling of error detection, for instance by extending extant models to include different patterns of error behavior and detection that break the assumptions of instant removal of detected errors and of no new sources of errors (e.g., Gaudoin, 1999; Yang et al., 2016). Another example is the distinct behavior of performance errors that occur after release (Zaman et al., 2012).

A *failure* is an unexpected result of a program execution. Failures are the external phenomena that the user experiences. An *error* (or fault) is incorrect code that, under certain conditions, will produce a failure. Errors are hidden from the user but are perceived as the cause of failures. Therefore the more failures experienced, the more errors assumed to exist in the software. A failure is related to a specific error, called a *detected error*. Several failures may be related to the same error, in which case there would only be one detected error. In reliability growth models, for instance, the number of errors is supposed fixed at the time a prototype is produced and the goal is to systematically eliminate them. (See, e.g., Heydari and Sullivan, K. M. 2017).

Open systems are related one to another. An error may be the result of a combination of conditions in two distinct applications that is incompatible with the code. Errors of incompatibility are practically endless and cannot be determined as a function of the code alone. This is different to traditional models.

We develop a general framework for modelling errors that are continually created. A probabilistic model is formulated that can lead to the development of the likelihood function from which estimates can be derived. This is a complex process since at any period an error may have been introduced at any previous time. In addition, only some users can detect a new error since they are the only ones to have upgraded and thus only they can be exposed to a current error of incompatibility.

2. Model Development

The intuitive discussion is now formulated mathematically. Certain practices may blur some of the theoretical distinctions made above. For instance, a Beta site may fall between testing and production and actual reports of failures may entangle the two types of errors. The mathematical formulation ignores such difficulties and assumes, for simplicity, some additional constraints as discussed below. Moreover, one general functional form is built, which will describe both pre and post-release stages. For clarity of presentation, we use the term users to denote both units of testing and units of use, although the former relates to the pre-release stage and the latter to the post-release stage.

There are three aspects to modelling this problem: (1) The process whereby customers, internet users, arrive to use the app and perhaps cancel their subscriptions at a later time; (2) The modelling of who upgrades their software/equipment and thus may encounter new errors of incompatibility; (3) The error detection process which involves modelling the arrival of new errors into the system and the number of users who can detect them.

The description will be given for a general continuous time process. (Unfortunately, in order to capture the real time complexity of the various processes, a lot of notation and definitions are involved.) Then, the simpler case of discrete time periods will be considered.

2.1. The Arrival and Cancellation Processes

Suppose the $X(t)$ denotes the number of new customers who use the software at time t. For instance, $X(t)$ could represent the number of people who sign up at time t for an online transcription service such as that provided by *Nuance*. Over time, some subscribers will cancel and no longer use the service. Let $C(t)$ denote the time a customer who signed up at time t cancels the service. (A very large value for $C(t)$ means that the customer never cancels.)

2.2. The Upgrade Process

For an individual completely current at time t (i.e. someone who is “new” at time t or who has been using the system before time and upgrades at time t), let $U(t)$ denote the time this customer next upgrades. (A very large value for $U(t)$ means that the customer never upgrades.)

Let $Z(s,t)$ denote the number of people who were current at time s, did not upgrade between times s and t, but did upgrade at time t. The quantity $Z(s,t)$ is a function of $X(u)$, $C(u)$ and $U(u)$ for $u \leq t$.

2.3. The Error and Detection Processes

At time t, let $Y(t)$ denote the number of new errors that are introduced into the system. For instance, a new version of the iPad might be introduced at time t. A subscriber using the software with this device might ultimately encounter an incompatibility error: the software works perfectly well but there is an as yet undiscovered error when the new device is used. Prior to the introduction of the new iPad, this error did not exist.

Let $D(t)$ denote the number of users who can detect *new* errors introduced at time t. ($D(t)$ will depend on the variables $X(t)$, $C(t)$ and $U(t)$) If one assumes that users who adopted prior to time t cannot detect errors introduced at time t, then $D(t)$ is simply equal to $X(t)$, the number of new users introduced at time t. This can be a reasonable assumption if one assumes that most users will not upgrade to new technology until a fair amount of time has elapsed. However, it is not necessary to make this assumption: any upgrade pattern can be accommodated.

For an error introduced at time s, let $P(s,t)$ denote the probability that a user current at time s detects an error during period number $t+1$ given that the user has not detected it prior to this period.

2.4. Analysis for Discrete Time Periods

In this paper, we will concentrate on the special but useful case where tracking is done over discrete periods. This is often the most realistic way to proceed and gives some useful practical and theoretical results. In order to make this clear the notations $X(t)$, $C(t)$, $Y(t)$, $U(t)$) and $D(t)$ will be replaced, respectively, by X_i, C_i, Y_i, U_i and D_i, where $i =$ 0,1,2,3 ... denotes the period ($i = 0$ correponds to release of the app, $i = 1$ corresponds to the end of the first period, etc. The quantity $Z(s,t)$ will be replaced by and $Z_{i,j}$ where i and $j$with $i < j$ are period numbers.

2.5. Example: Discrete Time Periods

Assume that no one cancels a subscription. Suppose that, at the beginning of any time period, a user who is current in the prior period will upgrade with probability 0.1 (i.e. is an "early adopter") , a customer who was last current two periods ago will upgrade with probability .15, those last current three periods ago will upgrade with probability 0.3 and those current more than three four periods ago will definitely upgrade. Then the probability distribution for $\mathrm{U_t}$, the time at which a customer current at time t will upgrade is given by:

$$U_t = \begin{cases} t+1 \text{ with probability } 0.1 \\ t+2 \text{ with probability } (0.9)(0.15) = 0.135 \\ t+3 \text{ with probability } (0.9)(0.85)(0.3) = 0.2295 \\ t+4 \text{ with probability } (0.9)(0.85)(0.7) = 0.5355 \end{cases}$$

Let $B(n,p)$ denote the value of a Binomial random variable with parameters n and p. The number who can detect errors at time 0 is D_0, the initial number of subscribers. At the end of period 1, the number of people who can detect new errors at time 1 equals the number of new subscribers $X(1)$ plus the number who have upgraded from form time - $B(D_0 0.1)$, i.e

$D_1 = X_1 + B(D_0, 0.1)$.

Using a similar argument, the number of subscribers who can detect errors at any time i can be found. For instance, the values of $D(2)$ and $D(3)$ are as follows:

$D(2) = X_2 + B(D_1, 0.1) + B(D_0, (0.15)(1 - 0.1))$

$D_3 = X_3 + B(D_2, 0.1) + B\big(D_1, (0.15)(1 - 0.1)\big) + B(D_0, (0.3)(1 - 0.15)(1 - 0.1))$.

Values for $Z_{i,j}$ can also be found. $Z_{i,i+1}$ is the number of customers who were current at time i and upgrade at time $i + 1$ and thus equals $B(D(D_i, 0.1)$. $Z_{i,i+2}$ is the number of people current at time i who do not upgrade at time $i + 1$ but do upgrade at time $i + 2$ and thus equals $B(D_i, (0.15)(1 - 0.1))$. Similarly, $Z_{i,i+3}$ = $B(D_i, (0.3)(1 - 0.15)(1 - 0.1))$ and $Z_{i,i+4}$ = $B(D_i, (1 - 0.3)(1 - 0.15)(1 - 0.1))$.

3. Discrete Time Periods and Constant Error Detection Probability

In this section, the problem will be simplified. We assume that $P(s,t) \equiv p$ and that $Y(t) \equiv \mu$. Ultimately, The goal is to estimate the quantities p, μ and $\upsilon \equiv X_0$ or, equivalently, D_0. (In the case of a subscription service X_0 is known but in other cases-for instance "free" software, the number of initial users may not be known.)

Consider a particular user who has the potential to discover a particular error. Then p denotes the probability that this user discovers this error during any given period. All users and all errors are assumed to be independent. (In a more general setting non-indepence may be allowed. For instance, Y_i, the number of errors introduced at time i, could have a distribution where the number if errors introduced in a given period depends on those introduced in a prior period. However, here we will focus on the simpler case of indepence which is difficult in its own right.) The p, however, may have interpretations that depend on the context: one p may be used for errors of content while a different p might be used for errors of incompatibility. Let $p(k,i)$ denote the probability that an error introduced at time k is detected for the first time during period i. This quantity can be shown to satisfy the following (proof omitted):

$p(k,i) = (1-p)^{N(k,i)}[1-(1-p)^{M(k,i)}]$,

where $N(k,i) \equiv (i-k-1)D_k + \sum_{j=k+1}^{i-2}(i-j-1)(X_j + Z_{k-1,j})$ and $M(k,i) \equiv D_k + \sum_{j=k+1}^{i-1}(X_j + Z_{k-1,j})$.

This quantity is key to writing down the likelihood. Suppose one has collected the data $x_1, \ldots, x_i$—the numbers of errors observed during each of the first i periods. Then the goal is to find the values of p, μ and υ that maximize the probability of observing this data stream. For given values of the parameters, it is necessary to construct an expression for $L(x_1, \ldots, x_i)$, the probability of observing $x_1, \ldots, x_i$. Note that

$$L(x_1, \ldots, x_i) = L(x_1) \prod_{i=1}^{n} L(\, x_i | x_{i-1}, \ldots, x_1)$$

where $L(x_1)$ is the probability of serving x_1 errors during the first period and $L(x_i | x_{i-1}, \ldots, x_1)$ is the conditional probability of observing x_i errors during period i given that $x_{i-1}, \ldots, x_1$were discovered during the previous periods. (These probabilities, of course, depend on the values of the parameters p, μ and υ.) From the expression above for $p(k,i)$ and noting that the number of errors observed in a given period is binomial with number of trials equal to the number of people who can detect an error, the above likelihood may be calculated. (It is somewhat complex since, during a given period one has to keep track of when errors were introduced and which consumers can see them.) In the worst case scenario, a grid search can be performed over the possible values for the quantities p, μ and

υ in order to find maximum likelihood estimates. For given values of the quantities p, μ and υ, the arrival of new data entails only a minor calculation to update the likelihood values. Thus in a big data context, the size of the data set does not hinder computational efficiency.

From the expression for $p(k,i)$ the expected number of errors detected during period i equals

$$\upsilon p(0,i)+\sum_{k=1}^{i}\mu p(k,i).$$

The variance of the number of errors detected in period i is given by

$$\upsilon p(0,i)(1-p(0,i)+\sum_{k=1}^{i}\mu p(k,i)(1-p(k,i).$$

Note that for given values of υ, μ and p, the above expressions are straightforward to evaluate.

From a management viewpoint, the above expressions can become effective tools. For many processes, control charts have become an important managerial tool for tracking quality, including software maintenance (Haworth, 1996). Control charts can be constructed tracking software errors: at the end of each period, compute the maximum likelihood estimate for the parameters; then compute the mean curve for the number of errors and the upper and lower control limits using the above expressions. Most applications of control charts are relatively straightforward and result in a constant central line. For software error tracking, however, the situation is more complex. For instance, the central line of a control chart based on the above expressions are not constant but its interpretation is similar to those of industrial applications. Points outside the upper and control limits indicate to the manager that the process is "out of control." This would happen, for instance, if any of the assumptions of the software error model were suddenly violated. The above expressions for mean and variance are therefore useful not only for predicting the flow of software errors but can also be used to warn a manager that the underlying marketplace is changing in an unexpected manner.

4. Conclusions

Traditionally, errors were defined within the system's boundaries. Goel (1985) suggests that "software faults can be attributed to an ignorance of the user requirements, ignorance of the rules of the computing environment, and to poor communication of software requirements between user and the programmer …" (ibid. p. 1411). This perspective focuses on mistaken code and is manifested in the quests for estimating error frequency according to various characteristics of the code. Now, with software being used by many users on the internet and mobile, errors rather than being drawn from a finite pool are constantly being introduced. In this paper, we shift the focus to the interaction between the system and its environment which leads to the notion of incompatibilities. The importance of incompatibility errors will grow with the growing impact of cloud computing and Big Data (Wang and Wu, 2016) as well as the Internet of Things (Prehofer, 2015). We formulate a robust and general model. In a Big Data setting, there is more freedom in allowing for more complex models since estimation of certain quantities (such as cancellation patterns and customer flow) is now much easier due to the sheer size of the data set. Error detection, even in a Big Data, context requires careful modelling since by design, even in large data sets, there is (and should be) a paucity of observations. We show how to calculate key quantities needed to construct the likelihood equation from which maximum likelihood estimators may be derived. A follow-up paper, rather than considering a grid search for finding these estimators will provide an algorithmic procedure that removes the need for a grid search. The important special case of discrete time periods and a constant rate of error introduction has been condidered in detail.

References

Banker, R. D., Davis, G. B., & Slaughter, S. A. (1998). Software development practices, software complexity, and software maintenance performance: A field study. *Management science*, *44*(4), 433-450.

Chan, T., Chung, S., & Ho, T. (1994). Timing of software replacement. *ICIS 1994 Proceedings*, 22.

Forbes. Shake those bugs out by A. Linsmayer. *Forbes*, May 18, 1998; 198-199.

Goel AL. Software reliability models. Assumptions, limitations and applicability. *IEEE Transactions on Software Engineering*, 1985. SE-11;(12); 1411-1423.

Gaudoin, O. (1999). Software reliability models with two debugging rates. *International Journal of Reliability, Quality and Safety Engineering*, *6*(01), 31-42.

Harworth D. A. Regression control charts to manage software maintenance. *Software Maintenance: Research and Practice*, 1996;8; 35-48.

Heydari, M., & Sullivan, K. M. (2017). An Integrated Approach to Redundancy Allocation and Test Planning for Reliability Growth. *Computers & Operations Research*.

Prehofer, C., & Chiarabini, L. (2015, July). From internet of things mashups to model-based development. In *Computer Software and Applications Conference (COMPSAC), 2015 IEEE 39th Annual* (Vol. 3, pp. 499-504). IEEE.

Wang, J., & Wu, Z. (2016). Study of the nonlinear imperfect software debugging model. *Reliability Engineering & System Safety, 153*, 180-192.

Zaman, S., Adams, B., & Hassan, A. E. (2012, June). A qualitative study on performance bugs. In *Mining Software Repositories (MSR), 2012 9th IEEE Working Conference on*(pp. 199-208). IEEE.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8333

Validation of innovation indicators from companies' websites

Héroux-Vaillancourt, Mikaël and Beaudry, Catherine
Department of Mathematics and Industrial Engineering, Polytechnique Montréal, Canada

Abstract

In this exploratory study, we use a web mining technique to source data in order to create innovation indicators of Canadian nanotechnology and advanced materials firms. 79 websites were extracted and analysed based on keywords related to the concepts of R&D and intellectual property. To understand what our web mining indicators actually measure, we compare them with those from a classic questionnaire-based survey. Formative indices from the surveys variables were built to better represent all the possibilities resulting from the web mining indicators. A MTMM matrix lead us to conclude that the formative indices are a good representation of the web mining indicators. As a consequence, the data extracted via our web mining technique can be used as proxies for the relative importance of R&D and the importance of IP, which would have previously only been measured using conventional methods such as government administrative data or questionnaire-based surveys.

Keywords: *Multi-Traits Multi-Method, construct validity, Web-mining, innovation measurement, nanotechnology and advanced materials*

1. Introduction

The majority of companies working in highly technological areas have an up-to-date website to inform potential customers, potential business partners and investors about their activities. Although the online information is made available by the companies themselves, suggesting the possibility of a strong desirability bias, this source of information can be suitable for the study of technological innovation (Domenech et al., 2015; Gök, Waterworth, & Shapira, 2015). The information obtained is as rich as it is diversified, including products, services, business models, R&D activities, etc. Would it be possible to extract this information and convert it into useful data to research? Moreover, is the information available on the various business websites is reliable and it is sufficient to give a good picture of some characteristics of companies? In other words, can the content of a commercial website be used to identify different innovation characteristics of a business?

When visiting a company's website, recurring themes that emerge from groupings of synonymous words may be noticed. These themes may actually describe factors appearing to be particularly important to the business. This study aims to validate whether the importance of factors emerging from a website is a good representation of the real importance a company actually gives to these factors.

In this study, we analysed and compared 2 sets of measures of innovation of nanotechnology and advanced materials in Canada stemming from two different data gathering techniques: Web scraping/mining and questionnaire-based survey. Comparisons between results from both methods were obtained via correlations. To ensure a convergent and discriminant validation of our results, we performed a Multi-Traits Multi-Method (MTMM) technique.

2. Methodology

2.1 Data acquisition

We started by conducting a classic questionnaire-based survey of which the core is based on the Oslo Manual (OECD & Eurostat, 2005) and explored the following themes: innovation, commercialisation, collaboration and intellectual property. We contacted 2971 Canadian high technological firms from which, 89 subjects were eligible and accepted to participate to our study. In order to build the two factors described above, R&D and intellectual property, we identified all the relevant questions from the questionnaire-based survey and transformed the answers to these questions into different types of variables. In the end, we generated a total of 9 variables pertinent to R&D, and 2 variables measuring intellectual property.

Then, we treated the websites of these 89 companies with the process described in Figure 1, from which we successfully extracted the website content of 79 companies (88%). The keywords related to R&D were selected from the literature (Gök et al., 2015), while the keywords related to intellectual property were identified from our own investigations of the literature. The most relevant keywords of any paper are generally listed on its first page, particularly under the abstract and served as a basis for the list of keywords used for the construction of our factors.

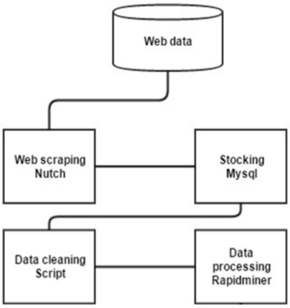


Figure 1. Web mining process

Clustering using keyword frequency analysis with a text mining software enabled us to count the number of occurrences of each keyword for each factor. We transformed these clusters of occurrences into 2 continuous variables. Because the 79 companies are different in structure and size, and therefore present different amounts of information in their websites, we standardized each variable by dividing all occurrences by the total number of words appearing in their website and multiplied the resulting value by 1,000. For each continuous variable, we calculated the Kurtosis and Skewness measures to determine whether they were following a normal distribution. None of our variables followed a normal distribution and were thus transformed by applying a natural logarithm (LN) or an inverse function (INV).

2.2 Construction of formative indices and validation construct

Given the vast field of words used to construct the web variables, treating each questionnaire-based variable individually may not be appropriate. To illustrate the large lexical field of possible words related to the factors studied, it is conceptually sound to build one single measure, one formative index with all the questions related to R&D and IP. Moreover, Principal Component Analysis (PCA) was performed on all the items related to R&D and to IP and did not produce any significant K-M-O and Cronbach's Alpha measures. This situation suggests the use of a formative index comprising several sub-elements explaining our R&D and IP factors.

Partial Least Square (PLS) regressions were estimated to determine whether it is possible to create valid formative indices for these two factors, R&D and IP. In order to use PLS regressions, the methodology requires only the use of complete data sets (Nelson, Taylor, & MacGregor, 1996). Non-response is usually treated either by weight adjustment, i.e. delete incomplete data entry and weigh remaining respondents to compensate for the deletion, or by imputation, i.e. adding artificial values based on average by classes and editing methods (Särndal, Swensson, & Wretman, 1992) to replace the missing values (Haziza & Beaumont, 2007). Since our sample size for IP is already low for one of the items (39 for the number of patents), we could not afford to treat the missing data with a weight adjustment. Thus, we replaced the missing data with their imputation class based on control variables. We sorted by sector, then by the number of employees and then by revenue. Depending on the situation, we used the mean of the class or the most conservative nearest-neighbour, a method commonly used in the literature (Haziza & Beaumont, 2007; Little, 1986; Thomsen, 1973).

Since not all the items shared the same scale, we transformed each variable into a Z-score. PLS regressions were then estimated using the WarpPLS 5.0 software with the following settings: MODEL B BASIC Warp3 Stable 3 and MODEL B BASIC Linear Stable 3. The two different settings produced the same conclusions. The details of the construct comparing the Web mining technique and the questionnaire-based survey are shown in Table 1.

All weights are significant (*p-value* < 0.01), indicator weight-loading signs are all positive, variance inflation factors (VIF) are all very low (<1.5) and the Effect sizes (ES) are all greater than 0.02. All the criteria are met to indicate that the indices generated are valid (Cenfetelli & Bassellier, 2009; Cohen, 1988; Diamantopoulos, 1999; Diamantopoulos & Siguaw, 2006; Diamantopoulos & Winklhofer, 2001; Petter, Straub, & Rai, 2007). For each factor, the sum of each weighted variable generated both indicator RD_INDEX and IP_INDEX.

Table 1. The validation construct

Concepts	Web Mining	Questionnaire	
R&D	LN_WEB _RD (Continuous, normal)	RD_INDEX (Continue, normal)	Z_NUMBER_RD (Continue, normal) Z_INT_INTERN_INFO_RD (Continue, normal) Z_INT_EXT_INFO_RD (Continue, normal) Z_INT_CONT_RD (Continue, normal) Z_INT_PROV_RD (Continue, normal) Z_TIME_RD (Continue, normal) Z_PROP_RD (Continue, normal)
Intellectual property	INV_WEB _IP (Continuous, normal)	IP_INDEX (Continue, normal)	Z_SUM_IP (Continuous, normal) Z_NB_PATENT (Continuous, normal)

3. MTMM Analysis results

First introduced by Campbell & Fiske (1959), the Multi-Trait Multi-Method (MTMM) allows for the convergent and discriminant validation of a construct where a set of t traits (interchangeable with factors in our case) are measured with m different methods.

This MTMM matrix includes the two data mining indicators along with RD_INDEX and IP_INDEX (see Table 2). The reliability diagonal will be neglected in our analysis since the measures are made with single items from the web method and with formative indices from our questionnaire. The monotrait-heteromethod diagonal shows high and significant correlations for R&D ($r = 0.419$; *p-value* < 0.01) and for IP ($r = 0.52$; *p-value* < 0.01), which hints at strong convergent validity. The heterotrait-monomethod value is low and not significant for the Web mining method ($r = 0.182$; *p-value* > 0.05) but the questionnaire-based survey method value is high and significant ($r = 0.32$; *p-value* < 0.01). However, the monotrait-heteromethod value is much higher than the heterotrait-monomethod values (< 0.419 > 0.182 and 0.52 > 0.32 for R&D and IP respectively), which indicates good nomological validity and that there are no mono-method biases. The first heterotrait-heteromethod value is low and not significant ($r = -0.17$; *p-value* > 0.05) while the other is moderate and significant ($r = 0.294$; *p-value* < 0.05). However, and more importantly, the correlations are lower than the corresponding values found in the validity diagonal, which shows good discriminant validity. All the conditions are satisfied under the original guidelines proposed by Campbell and Fiske (1959), and therefore, no risk of potential biases is induced within the methods, the traits or a combination of both. The results based on this methodology suggests that our web mining indicators reflect the importance given to innovation factors such as R&D and Intellectual property.

Table 2. MTMM matrix for RD_INDEX and IP_INDEX

	Traits	**Method 1 (Web)**		**Method 2 (Questionnaire)**	
		RD	IP	RD (RD_INDEX)	RD (IP_INDEX)
Method 1 (Web)	RD	N/A[a]			
	IP	-0.182	N/A[a]		
Method 2 (Questionnaire)	RD (RD_INDEX)	0.419**	0.294*	N/A[a]	
	IP (IP_INDEX)	-0.17	0.52**	0.32**	N/A[a]

Note: All traits from Method 1 are measured by single items, there are no reliability statistic that can be calculated. All traits from Method 2 are measured by a formative index and thus, reliability statistic is irrelevent.

p < .05.

p < .01.

4. Limitations and future research

Obviously, more data would allow our research to be more robust. Another limitation of our methodology is the fact that we did not take into account the context around our keywords, possibly leading to multiple false positives. The addition of machine learning techniques, such as Recurrent Neural Network or Natural Language Processing or Bag-of-Words model, is a promising avenue to improve the level of precision by adding to the method the necessary context around keywords. Moreover, we started with theoretical factors for the conceptual framework, then identified the keywords related to these factors, and finally mined the website for these specific keywords. An interesting alternative would be to do this the other way around, i.e., to start with the website content and to identify the factors that can be naturally found via unsupervised machine learning algorithms. The term frequency inverse document frequency technique (TF-IDF) could be used to provide insight into the importance of keywords relative to the rest of a document.

In a nutshell, our methodology seems it can be used as a valid approach to provide data for future innovation and technology management studies for the relative importance given to a factor such as R&D and IP, and to test the validity of the measures thus created. In most questionnaire-based surveys, that information is gathered using 1 to 7 Likert scale questions. If the goal of a study is to determine the degree of importance of core factors such as R&D or IP for a firm, the use of Web mining indicators is reasonable. However, if the goal is to gather more specific information, such as the precise actions undertaken by a firm, these web mining indicators may lack the necessary context to behave as expected. The importance given by a company to certain types of activities represents which activities that are supported and encouraged by the culture of the company (Herzog, 2011). Therefore, it is possible that our methodology suggest a novel way to measure quantitavely innovation culture.

Of course, company websites are willingly structured in a cooperative and agreeable manner toward whomever is seeking information concerning products, services, activities, and so on. The self-reporting bias induced by this methodology is inevitable. However, it is important to note that questionnaire-based surveys and most national official public directory are all subject to self-reporting biases as well. Fortunately, the bias induced by the web mining technique is as much a quality as a flaw, in that it provides insight on how the company wants to be perceived. Indeed, companies write on their websites about what they care about, what is important for them and who they are as a company. This qualitative information represents the essence of the company. Future research is needed to determine whether this qualitative information could be used as a proxy to understand a company’s culture for instance. Furthermore, future research will be performed to assess how these indicators can be used in actual regressions to understand innovation patterns. It will be especially interesting to assess whether these web indicators tend to be substitutes or complements to the traditional measures use in innovation management studies. This will be performed in the coming months with a sample of 1700 companies.

References

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Cenfetelli, R. T., & Bassellier, G. (2009). Interpretation of Formative Measurement in Information Systems Research. *MIS Quarterly*, *33*(4), 689–707.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, N.J.: L. Erlbaum Associates.

Domenech, J., Rizov, M., and Vecchi, M. (2015). *The Impact of Companies' Websites on Competitiveness and Productivity Performance* (Conference Paper: First International Conference on Advanced Research Methods and Analytics).

Diamantopoulos, A. (1999). Viewpoint – Export performance measurement: reflective versus formative indicators. *International Marketing Review*, *16*(6), 444–457. https://doi.org/10.1108/02651339910300422

Diamantopoulos, A., & Siguaw, J. A. (2006). Formative Versus Reflective Indicators in Organizational Measure Development: A Comparison and Empirical Illustration. *British Journal of Management*, *17*(4), 263–282. https://doi.org/10.1111/j.1467-8551.2006.00500.x

Diamantopoulos, A., & Winklhofer, H. M. (2001). Index Construction with Formative Indicators: An Alternative to Scale Development. *Journal of Marketing Research*, *38*(2), 269–277. https://doi.org/10.1509/jmkr.38.2.269.18845

Gök, A., Waterworth, A., & Shapira, P. (2015). Use of web mining in studying innovation. *Scientometrics*, *102*(1), 653–671. https://doi.org/10.1007/s11192-014-1434-0

Haziza, D., & Beaumont, J.-F. (2007). On the construction of imputation classes in surveys. *International Statistical Review*, *75*(1), 25–43.

Herzog, P. (2011). Innovation culture. In *Open and Closed Innovation* (pp. 59–82). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-8349-6165-5_3

Little, R. J. A. (1986). Survey Nonresponse Adjustments for Estimates of Means. *International Statistical Review / Revue Internationale de Statistique*, *54*(2), 139–157. https://doi.org/10.2307/1403140

Nelson, P. R. C., Taylor, P. A., & MacGregor, J. F. (1996). Missing data methods in PCA and PLS: Score calculations with incomplete observations. *Chemometrics and Intelligent Laboratory Systems*, *35*(1), 45–65. https://doi.org/10.1016/S0169-7439(96)00007-X

OECD, & Eurostat. (2005). *Oslo Manual*. Paris: Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd-ilibrary.org/content/book/9789264013100-en

Petter, S., Straub, D., & Rai, A. (2007). Specifying Formative Constructs in Information Systems Research. *MIS Quarterly*, *31*(4), 623–656.

Särndal, C. E., Swensson, B., & Wretman, J. (1992). Model assisted survey sampling Springer. *New York*.

Thomsen, I. (1973). A note on the efficiency of weighting subclass means to reduce the effects of nonresponse when analyzing survey data. *Statistisk Tidskrift*, *4*, 278–283.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8334

A combination of multi-period training data and ensemble methods to improve churn classification of housing loan customers

Seppälä, Tomi; Thuy, Le
Department of Information and Service Management, Aalto University, School of Business, Finland

Abstract

Customer retention has been the focus of customer relationship management in the financial sector during the past decade. The first and important step in customer retention is to classify the customers into possible churners, those likely to switch to another service provider, and non-churners. The second step is to take action to retain the most probable churners.

The main challenge in churn classification is the rarity of churn events. In order to overcome this, two aspects are found to improve the churn classification model: the training data and the algorithm. The recently proposed multi-period training data approach is found to outperform the single period training data thanks to the more effective use of longitudinal data. Regarding the churn classification algorithms, the most advanced and widely employed is the ensemble method, which combines multiple models to produce a more powerful one. Two popularly used ensemble techniques, random forest and gradient boosting, are found to outperform logistic regression and decision tree in classifying churners from non-churners.

The study uses data of housing loan customers from a Nordic bank. The key finding is that models combining the multi-period training data approach with ensemble methods performs the best.

Keywords: *churn prediction, ensemble methods, random forest, gradient boosting, multiple period training data, housing loan churn*

1. Introduction

Customer retention has been the focus of customer relationship management research in the financial sector during the past decade (Zoric, 2016). Retaining existing customers is argued to be more economical over the long run for companies than acquiring new ones (Gur Ali & Ariturk, 2014). Van den Poel & Lariviere (2004), in their attempt to translate the benefits of retaining customers over a period of 25 years into monetary terms, concludes that an additional percentage point in customer retention rate contributes to an increase in revenue of approximately 7% (Van den Poel & Lariviere, 2004).The first step in customer retention is to classify the customers into binary groups of possible churners, indicating to customers that are likely to switch to another service provider, and non-churners, referring to those that are probably staying with the current provider. The second step in customer retention is to take action to retain the most probable churners to either minimize costs or maximize benefits. As a result, churn classification is an important first step in customer retention.

However, the main challenge in churn classification is the extreme rarity of churn events (Gur Ali & Ariturk, 2014). For example, the churn rate in the banking industry is usually less than 1%. In order to overcome this rarity issue, a great deal of research has been found to improve the two main aspects of a churn classification model: the training data and the algorithm (Ballings & Van den Poel, 2012). Regarding the training data, the recently proposed multi-period training data approach is found to outperform the single period training data thanks to the more effective use of longitudinal data of churn behavior (Gur Ali & Ariturk, 2014). Regarding the churn classification algorithms, the most advanced and widely employed is the ensemble method, which combines multiple models to produce a more powerful one (Yaya, et al., 2009). Two popularly used ensemble techniques are random forest and gradient boosting (Breiman, 2001), both of which are found to outperform logistic regression and decision tree methods in classifying churners and non-churners.

2. Research questions

To the best of the authors' knowledge, the proposed multi-period training data has not been applied with the ensemble methods in a churn classification model. As a result, in this study we examine whether the multi-period training data approach, when employed together with ensemble methods in a churn classification model, produces better churn prediction than with logistic regression and decision tree approach.

The research problem is detailed into the following research questions:

1. In models that employ logistic regression and decision trees, does the multi-period training data approach improve churn classification performance compared to the single period training data approach?

2. In models that employ single period training data, do random forest and gradient boosting improve churn classification performance compared to logistic regression and decision trees?

3. Do models that employ both the multi-period training data approach and ensemble methods perform better in churn classification than those in the first question?

4. What are the best churn predictors in the housing loan context?

3. Methods and Data

In order to answer the research questions, four methods are employed in this study as churn classification algorithms: Logistic Regression, Decision Tree, Random Forest, and Gradient Boosting. All of the four methods are employed with both multi-period training data and single period training data to create competing models. Specifically, logistic regression and decision trees are used to run the baseline models with single period training data. The other models are then compared with the baseline models in order to answer the research question.

This study uses empirical data of housing loan customers from a Nordic bank. The data was collected and analysed for time period between August 1, 2015 and March 31, 2016, and was divided into observation and performance periods. The predictors employed in this study include information from all the four main groups as recommended in the literature: demography, customer behavior, characteristics of customer relationship and macro-environmental factors. The following variable selection methods are employed in the SAS Enterprise Miner before running the logistic regression models: a decision tree with the CHAID method, step wise selection, and the variable selection node using the R square procedure. The churn models are evaluated based on three criteria: misclassification rate, Receiver Operating Characteristics (ROC) index and top decile lift.

4. Results and discussion

This study validates that both multi-period training data and ensemble methods actually improve the churn classification performance compared to their counterparts in the housing loan context. More importantly, when employed together, the models with the combination of the proposed multi-period training data approach and ensemble methods such as random forest and gradient boosting have the best performance among all the created models based on the misclassification rate, ROC index and top decile lift. Type II error refers to misclassifying churners as non-churners and it is more severe than misclassifying non-churners as churners (Type I error) since potential churners will be highly likely to churn without receiving any retention action. Using multi-period training data, the best models are produced with the random forest with a reduction of more than 10% in type II error rate compared with the worst performing models that employ single period training data and logistic regression without variable selection. Such improvement in the misclassified churn events can considerably prevent the bank from a considerable loss of those customers without taking any retention action. The improvement is mainly thanks to the more effective use of churn events that are usually scarce in real life data. Specifically, in contrast to the single period training data approach that captures churns only at a specific period of time and discards the churn events that have happened prior to that period, the multi-period training data approach allows the employment of historical churn events, providing the models with more churn events and mitigating the rarity issue in churn prediction. Therefore the imbalance between the classes is not as severe a problem when using the multiperiod training data approach. Consequently, the authors highly recommend other studies in churn classification to employ the multi-period training data approach together with ensemble methods to achieve the best possible classification models.

Regarding the last research question, this study shows that the most important churn predictors belong to the demographic group, in which the number of family members has the most significant effect on churning. It makes sense that a change in the number of family members, will considerably impact the decision related to a housing loan.

References

Breiman, L.(2001). Random Forests. *Machine Learning*, 45(1), 5-32.

Ballings, M. & Van den Poel, D. (2012). Customer Event History for Churn Prediction - How Long Is Long Enough?. *Expert Systems with Applications*, 39 (18), 13517-13522.

Gur Ali, Ö. & Ariturk, U. (2014). Dynamic Churn Prediction Framework with More Effective Use of Rare Event Data: The Case of Private Banking. *Expert Systems with Applications,* 41(17). 7889-7903.

Van den Poel, D. & Lariviere, B. (2004). Customer Attrition Analysis for Financial Services Using Proportional Hazard Models. *European Journal of Operational Research*, 157 (1), 196-217.

Yaya, X., Xiu, L., E.W.T., N. & Weiyun, Y. (2009). Customer Churn Prediction Using Improved Balanced Random Forests. *Expert Systems with Applications,* 36(3),5445–5449.

Zoric, A. B. (2016). Predicting Customer Churn in Banking Industry Using Neural Networks. *Interdisciplinary Description of Complex Systems,* 14(2). 116-124.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8336

Technical Sentiment Analysis: Measuring Advantages and Drawbacks of New Products Using Social Media

Chiarello, Filippo[a] ; Bonaccorsi, Andrea[a] ; Fantoni, Gualtiero[a];Ossola, Giacomo[a]; Cimino, Andrea[b] and Dell'Orletta, Felice[b]

[a]Department of Energy, Systems, Territory, and Construction Engineering, University of Pisa, Italy. [b]Institute for Computational Linguistics of the Italian National Research Council (ILC- CNR)

Abstract

In recent years, social media have become ubiquitous and important for social networking and content sharing. Moreover, the content generated by these websites remains largely untapped. Some researchers proved that social media have been a valuable source to predict the future outcomes of some events such as box-office movie revenues or political elections. Social media are also used by companies to measure the sentiment of customers about their brand and products.

This work proposes a new social media based model to measure how users perceive new products from a technical point of view. This model relies on the analysis of advantages and drawbacks of products, which are both important aspects evaluated by consumers during the buying decision process. This model is based on a lexicon developed in a related work (Chiarello et. al, 2017) to analyse patents and detect advantages and drawbacks connected to a certain technology.

The results show that when a product has a certain technological complexity and fuels a more technical debate, advantages and drawbacks analysis is more efficient than sentiment analysis in producing technical-functional judgements.

Keywords: *Social media; Twitter; Sentiment analysis; Product Success*

1. Introduction

Nowadays, social media have become an inseparable part of modern life, providing a vast record of mankind's everyday thoughts, feelings and actions. For this reason, there has been an increasing interest in research of exploiting social media as information source of knowledge although extracting a valuable signal is not a trivial task since social media data is noisy and must be filtered before proceeding with the analysis. In this domain, sentiment analysis, which aims to determine the sentiment content of a text unit, is considered one of the best data mining method. It relies on different approaches (Collomb et al. 2013) and it has been used to answer research questions in a variety of fields comprised the measure of customers perception of new products (Mirtalaie et al. 2018).

In this work, we try to understand if sentiment analysis is really the best available method to analyse consumer's perception of products, expecialy when we want to measure the perception of the technical content of the product. Thus we compare State of the art sentiment analysis techniques with a lexicon of advantages and drawbacks related to products. This tool relies on a lexicon developed by Chiarello (2017) to extract advantages and drawbacks of inventions from patents.

Our work started with the selection of an event able to polarise Twitter users' attention and products to analyse. In particular, we chose a premiere tradeshow for the video game industry, and two video game consoles disclosed during the event. We collected about 7 milions tweets about products published before, during and after the tradeshow. Since social media data is noisy (for example it may contains spam and advertising), before proceeding with the analyses, we filtered our dataset. In particular, after removing too short and non-English tweets, we manually classified a randomly extracted subset of posts to train a classifier which provide us the cleansed dataset. Then we conducted a sentiment analysis of the tweets using state of the art machine learning techniques. We classified each tweet as positive, negative or neutral. At this point we applied our lexicon identifying advantages tweets and drawbacks tweets. Finally we compared the outputs of the two analyses for the two product-related clusters of tweets.

We found consistent differences between the extractions. The results shows that when a product has a certain technological complexity and fuels a more technical debate, advantages and drawbacks analysis is more able than sentiment in producing technical-functional judgements. For this reason we think that the proposed methodology peforms better then standard sentiment analysis techniques when a product has a certain technological complexity and fuels a more technical social media discourse.

2. State of the art

We provide an overview of the studies about social media forecasting (Table 1, 2). Researchers especially focused on economics (stock market, marketing, sales) and politics (elections outcomes). In economics, predicting fluctuations in the stock market has been the most studied by far. Early work focused largely on predicting whether aggregate stock measures such as the Dow Jones Industrial Average (DJIA) would rise or fall on the next day, but forecasting can also involve making more detailed predictions, e.g., forecasting market returns or making predictions for individual stocks. The simplest task for stock market prediction is deciding whether the following day will see a rise or fall in stock prices. Comparison between studies is complicated by the fact that stock market volatility, and thereby the difficulty of prediction, may vary over time periods. High accuracy (87,6%) on this task was reported by Bollen (2012). However, slight deviations away from their methodology have seen much less success indicating that the method itself may be unreliable (Xu, 2014). A very good result is achieved by Cakra (2015) who use linear regression to build a prediction model based on the output of sentiment analysis and previous stock price dataset.

Social media has also been used to study the ability of online projects to successfully crowdfund their projects through websites like Kickstarter. Li (2016) predicts whether a project will eventually succeed by making use of features relevant to the project itself (e.g., the fundraising goal), as well as social activity features (e.g., number of tweets related to the project), and social graph measures (e.g., average number of followers for project promoters). Using all of these features for only the first 5% of the project duration achieved an AUC of 0.90, reflecting very high classification performance.

Many studies analysed the predictive power of social media to improve or replace traditional and expensive polling methods. The simplest technique is measuring tweet volume (tweet mentioning a political party = votes). Chung (2010) and Tumasjan (2010) employed this method obtaining mixed results. Razzaq (2014), Skoric (2012) and Prasetyo (2015) improved this method taking into account the mood of the posts, considering if a candidate or a party is mentioned in a positive or negative manner.

Table 1: Summary of studies in economics. Data source: T = Twitter, F = Facebook, K = Kickstarter, O = blogs, other. Task: MDA = Mean Directional Accuracy, MAPE = Mean Absolute Percentage Error.

Article	Topic	Data source	Data size	Observation time	Success rate
Xu (2014)	Stock market	T	100K tweets	42 days	$MDA = 58.9\%$
Crone (2014)	Exchange rates	T, F, O	N/A	N/A	$MDA = 60.26\%$
Kordonis (2016)	Stock market	T	N/A	N/A	$MDA = 87\%$
Cakra (2015)	Stock market	T	N/A	2 weeks	$R^2 = 0.9983$
Bollen (2015)	Stock market	T	9.8M	10 months	$MDA = 87.6\%$
Brow (2012)	Stock market	T	13K	9 days	$r = 0.62$
Rao (2012)	Stock market	T	4M	14 months	$R^2 = 0.95$ (DJIA); $R^2 = 0.68$ (NASDAQ).
Kim (2014)	Hit songs	T	31.6M	68 days	$F1 = 0.841$
Korolov (2015)	Donations	T	15M	10 days	$R^2 = 0.9286$
Le (2015)	Sports book	T	1.2M	30 days	$AUC = 80\%$ $Return = 8\%$
Tuarob (2013)	Smartphone sales	T	800M	19 months	$r = 0.8837$
Asur (2010)	Movie revenues	T	2.8M	3 months	$R^2\ adj = 0.94$
Ahn (2014)	Car sales	T, F, O	26K posts	N/A	$RMSE = 0.170$ (Sedan A); $RMSE = 0.232$ (Sedan B).
Chen (2015)	Advertising	T	5.9K users	N/A	66% gain (click rate); 87% gain (follow rate).
Li (2016)	Crowdfunding success rate	T, F, K	106K tweets	6 months	$AUC = 0.90$

Researchers employ different tools and methods for social media mining, varying from easy to somewhat more complex. The most employed tool is sentiment analysis (with its various approaches: knowledge-based techniques, statistical methods, and hybrid approach) which usually achieves good results. Other researchers use more complex tools such as neural networks or a combination of techniques. At end of the analysis of the state of the art we are able to identify some best practices: (i) implementing suitable techniques to deal with noisy data, (ii) evaluating statistical biases in social media data, (iii) collecting data from heterogeneous sources, (iv) incorporating domain-specific knowledge to improve statistical model.

Table 2: Summary of studies in politics. Data source: T = Twitter. Task: Acc. = Accuracy, MDA = Mean Directional Accuracy, MAPE = Mean Absolute Percentage Error.

Article	Topic	Data source	Data size	Observation time	Success rate
Chung (2010)	Renewal of US senate	T	235K tweets	7 days	Acc. 41% - 47%
Tumasjan (2010)	German federal election	T	104K tweets	36 days	MAE 1.65%
Razzaq (2014)	Pakistani election	T	613K tweets	N/A	Acc. 50%
Skoric (2012)	Political election	T	7M tweets	36 days	MAE 6.1%
Prasetyo (2015)	Indonesian political election	T	7M tweets	83 days	MAE 0.62% (State level)

3.Methodology

3.1 Selection of a triggering event and products

We chose the Electronic Entertainment Expo as event able to polarise users' attention. Commonly referred to as E3, it is a premier trade event for the video game industry, presented by the Entertainment Software Association (ESA). We chose two new video game consoles, disclosed at E3 2017, as products of which predicting the success or failure. The first is Xbox One X, a new high-end version of Xbox One with upgraded hardware and the other product is New Nintendo 2DS XL, a streamlined version of the handheld console New Nintendo 3DS XL.

3.2 Data collection

Twitter provides two possible ways to gather tweets: the Streaming Application Programming Interface (API) and the Search API. The first one allows user to obtain real-time access to tweets from an input query. The user first requests a connection to a stream of tweets from the server. Then, the server opens a streaming connections and tweets are streamed in as they occur, to the user. However, there are a few limitations of the Streaming API. First, language is not specifiable, resulting in a stream that contains tweets of all languages, including a few non-Latin-based alphabets, that complicates further analysis. Instead, Twitter Search API is a Representational State Transfer API which allows users to request specific queries of recent tweets. It allows filtering based on language, region, geolocation, and time. Unfortunately, using the Search API is expensive and there is a rate limit associated with the query. Because of these issues, we decided to go with the Twitter Streaming API instead. For each product, we detected related hashtags and keywords an constructed a query to download relevant tweets.

We chose to collect tweets not only after the tradeshow, but also before. For these reason, we initially identified some products keywords with their provisional names and we updated them at a later stage. Tweets have been downloaded from CNR (Consiglio

Nazionale delle Ricerche, Istituto di Informatica e Telematica, Area di Pisa) since 11th June 2017 h. 10:00 to 31th July 2017 h. 15:00.

3.3 Data filtering

The initial dataset resulted to be very noisy, containing tweets written in different languages, advertising and posts related to different products or subjects. We chose to keep into account only English tweets because sentiment and advantages/drawbacks lexicon is in this language. The data set is filtered removing tweets with less than five words and non-English posts with a language classifier. We obtained 7.165.216 of filtered tweets.

At this point we created a golden set of relevant tweet to train a Supported Vector Machine classifier able to recognize relevant and unrelevant tweets. We defined characteristics that make a tweet: (i) relevant (posted by users or containing words or opinions related to our products of interests and their functionalities), (ii) irrelevant (tweets containing advertisings, links to e-commerce websites or messages related to other products or subjects). A researcher manually classified a subset made up of randomly extracted tweets. In particular, we exctract a subset composed of 6.500 finding 105 positive results and 6.395 negative. SVM model was then trained using this dataset, and computed a probability for each tweet to be relevant or irrelevant. A threshold of 0.7 has been chosen to label a tweet as relevant. The final dataset of filtered tweets, made up of 66.796 posts. We clustered tweets using product-related keywords. Clustering posts allowed us to further filter the final dataset which contained a small number of irrelevant tweets (Table 3).

Table 3. Clusters of tweets

	N° of tweets	% of tweets
Xbox One X	64.885	97,14 %
New N2DS	1.706	2,55 %
Irrelevant tweets	198	0,30 %

Table 4. Sentiment analysis classification

	Positive	Negative	Neutral
Xbox One X	35,99%	4,65%	59,37%
New N2DS	52,99%	1,58%	45,43%
Overall	36,42%	4,57%	59,01%

3.4 Sentiment analysis

Table 4 presents the results of the sentiment analysis. We classified each tweet according to its sentiment into positive, negative, or neutral. We used an established methodology developed by Cimino (2016). We pre-processed the tweets by removing mentions (@ character), URLs, product hashtags, emoticons and single characters. As a result, for each tweet we obtained a probability of belonging to a mood class. After a manual analysis, we used a class prediction probability threshold of 0.6 to filter out low confidence prediction, i.e. tweets that cannot be classified as positive or negative with a high confidence are classified as neutral instead.

3.5 Advantages and drawbacks analysis

To extract technical advantages and drawbacks from tweets we used the lexicon developed in Chiarello (2017) that contains 657 Advantages words and 297 Drawbacks clues. These words are searched on our dataset finding different percentages of tweets with words from the lexicon in the two product-related clusters of tweets. Table 5 reports the results.

Table 5: Percentages of tweets containing or not words from our lexicon.

	Tweets with adv	Tweets with drw	Tweets with adv & drw	Tweets with no adv or drw	Tweets with adv or drw
Xbox One X	8,84%	3,74%	0,37%	87,05%	12,95%
New N2DS XL	6,62%	0,94%	0,00%	92,44%	7,56%

4. Results: Comparison Between Sentiment Analysis and Technical Advantages and Disadvantage Extraction

We adapted the advantages & drawbacks analysis to give as output a classification ef each tweet. We classified data coming from the latter analysis in this way: (i) **positive** (tweets containing only advantages words), (ii) **negative** (tweets containing only drawbacks), (iii) **neutral** (tweets with no words of our lexicon or controversial tweets). As we can see in figure 2, sentiment analysis is more able to polarise tweets. In fact, with this analysis we found lower levels of neutral tweets, respectively 59.37 % for Xbox One X and 45.43% for the New Nintendo 2DS XL.

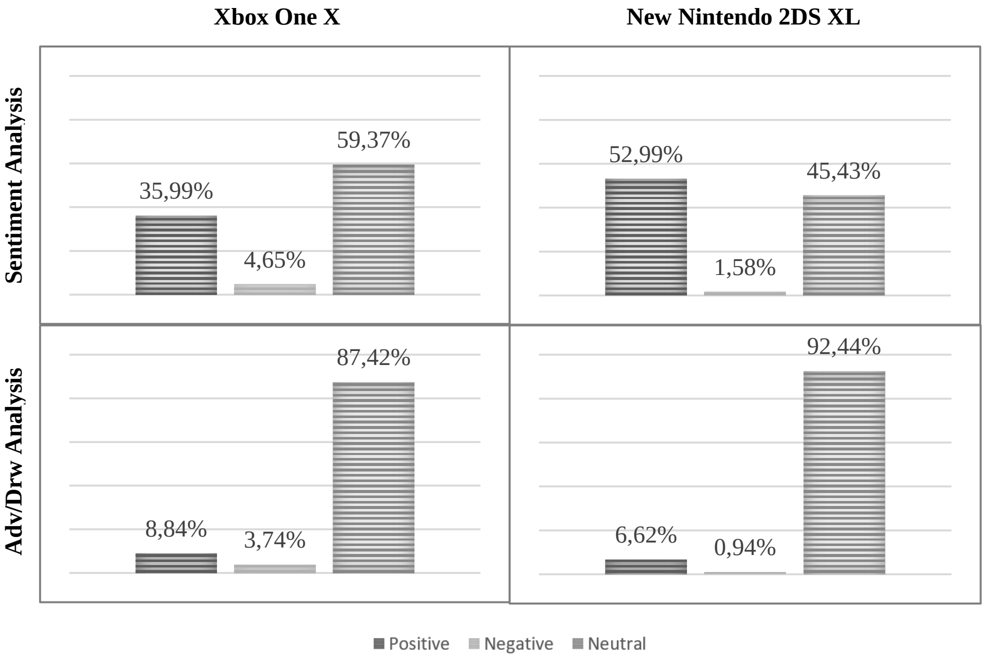


Figure 1: Comparison between Sentiment analysis and Advantages/Drawbacks analysis

This was an expected result since this kind of analysis is designed to deal with colloquial language while our lexicon is technical, being derived from patents analysis. What surprised us is the different polarisation of the products that we see comparing the two analyses. In fact, while with sentiment analysis Nintendo achieves lower percentages of neutral tweets, with advantages and drawbacks analysis is the opposite, since Xbox tweets are more polarised. We also noted that we found more tweets with words of our lexicon in the Xbox subset than in the Nintendo one (Table 5). We did the hypothesis that the differences between the percentages of tweets with words found for each product, and the differences of polarisation between the two analyses depend on the different marketing focus, target customer, and technological complexity of the two new video game consoles. Xbox One X targets hard-core gamers who really wants a premium experience1. With its marketing campaign, Microsoft pushed the technical supremacy of its new machine over the competitors' products, fuelling a debate about its technical features amongst the potential users.

As a result, the campaign produced a more technical social discourse that allowed us achieving better results. Instead, the new Nintendo handheld console has been developed

[1] http://www.businessinsider.com/why-xbox-one-x-costs-500-2017-6?IR=T (last access: 17/11/2017)

targeting children and families providing a model that falls somewhere in the middle of the line of 3DS consoles2.

We initially checked our hypothesis using Google Trend to compare users' search interest about technical review of the two products during the data collection period (Figure 2). Then, we analysed the number of technical articles related to the new products published by the 25 most popular video games and technology websites in the U.S, according to the ranking of SimilarWeb, a digital marketing intelligence company which publishes insights about websites. We entered queries reported in Table 6 into Google search engine to retrieve technical article within the web domains previously identified: we obtained 1.117 articles about Xbox and only 52 about Nintendo, proving that technical debate concerning Xbox is greater. This is and evidence of the fact that when a product has a certain technological complexity and fuels a more technical debate, advantages and drawbacks analysis is more able than sentiment in producing technical-functional judgements. The greater number of neutral tweets found with advantages and drawbacks analysis can also be explained with the Means-end chain model (Reynolds, 1995). Consumers express themselves basing on personal consequences linked with product use or basing on personal values satisfied by the product itself. For these reasons, tweets contain a more colloquial language which sentiment analysis is more able to interpret than the latter tool.

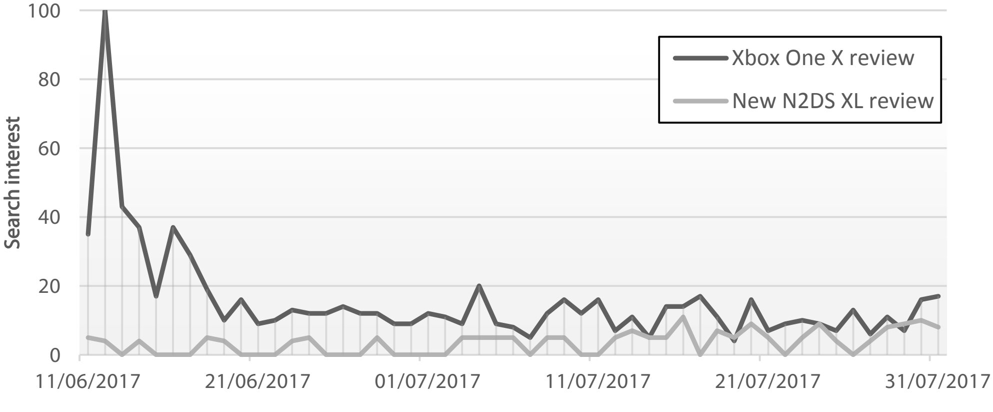


Figure 2: Google Trends comparison of search-terms "Xbox One X review" and "New Nintendo 2DS XL review" during the data collection period, since 11th June 2017 to 31st July 2017. Values on the vertical axis depict search interest compared to the highest point in the graph during the observation time. A value of 100 is the peak popularity for the term. On average, users searched for Xbox reviews with an approximately five times higher frequency.

[2] http://www.nintendolife.com/news/2017/05/reggie_explains_the_reasoning_behind_the_new_2ds_xl (last access: 17/11/2017)

Table 6: Queries entered into Google search engine to search for technical articles within selected web domains. We selected keywords related to technical features of the products. The example report queries used for one of the analysed website: ign.com

Xbox One X	`allintitle: (4k OR hdr OR hardware OR graphics OR review OR resolution OR fps OR fast OR comparison OR frame OR enhanced OR performance OR cpu OR gpu OR ram) AND ("xbox one x") site: ign.com`
New Nintendo 2DS XL	`allintitle: (graphics OR review OR screen OR comparison OR enhanced OR performance OR cpu OR gpu OR ram OR battery OR weight) AND "new nintendo 2ds xl" site: ign.com`

5. Conclusion

Methods and techniques for social media mining with sentiment analysis is one of the most appreciated tool amongst researchers, having a very good reputation in the informatic fields. Also, big companies make use of it because it can be a rich source of information to adjust marketing strategies, improve campaign success, advertising message, and customer service. Nevetheless sentiment analysis is designed to extract feelings related sentiment polarity from tweets of user and not other kinds of polarity, like polarity related to technical advantages and drawbacks of products the users are experiencimg.

In this paper we shown how using a technical lexicon to analyse technical polarity of tweets is a a more effective approach in giving technical-functional judgements about a product we respect to state of the art sentiment analysis techniques. It is particulartly true when a product has a certain technological complexity.

References

Ahn H., and Spangler W. S. (2014) "Sales prediction with social media analysis".SRII Global Conference. IEEE, 2014.

Asur S., and Huberman B. A. "Predicting the Future With Social Media". In IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology. IEEE, 2010.

Bollen J., Mao H., and Zeng X. , 2015, "Twitter mood predicts the stock market". Ref: http://arxiv.org/abs/1010.3003

Brown, Eric D., "Will Twitter Make You a Better Investor? A Look at Sentiment, User Reputation and Their Effect on the Stock Market" (2012). SAIS 2012 Proceedings. 7.

Cakra Y. E., and Trisedya B. D. "Stock price prediction using linear regression based on sentiment analysis". In 2015, International Conference on Advanced Computer Science and Information Systems (ICACSIS). IEEE, 2015.

Chen J., Haber E., Kang R., Hsieh G., and Mahmud J. "Making use of derived personality: The case of social media ad targeting". In Proceedings of the International AAAI Conference on Web and Social Media (ICWSM), 2015.

Chiarello F., Fantoni G., Bonaccorsi A. (2017) Product description in terms of advantages and drawbacks: Exploiting patent information in novel ways. ICED 2017

Chung J., and Mustafaraj E. "Can collective sentiment expressed on twitter predict political elections?". In *Twenty-Fifth AAAI Conference on Artificial Intelligence*. AAAI, 2010.

Cimino A., Dell'Orletta F. (2016) "Tandem LSTM-SVM Approach for Sentiment Analysis". In Proceedings of EVALITA '16, Evaluation of NLP and Speech Tools for Italian, 7 December, Napoli, Italy.

Collomb A. ,Costea C.,Brunie L. (2013). A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation.

Crone S. F., and Koeppel C. "Predicting exchange rates with sentiment indicators". In 2014, IEEE conference on computational intelligence for financial engineering & economics (CIFEr). IEEE, 2014.

Kim Y., Suh B., and Lee K. "#nowplaying the future billboard: mining music listening behaviors of twitter users for hit song prediction". In Proceedings of the first international workshop on social media retrieval and analysis, pages 51-56. ACM, 2014.

Kordonis J., Symeonidis S., and Arampatzis A. "Stock price forecasting via sentiment analysis on Twitter". In Proceedings of the 20th Pan-Hellenic Conference on Informatics, article no. 36. ACM, 2016.

Korolov R., Peabody J., Lavoie A., Das S., Magdon-Ismail M., and Wallace W. "Actions are louder than words in social media". In IEEE/ACM International Conference on Advances in Social Network Analysis and Mining. IEEE, 2015.

Le L., Ferrara E., and Flammini A. "On predictability of rare events leveraging social media: a machine learning perspective". In Proceedings of the 2015 ACM on Conference on Online Social Networks. ACM, 2015.

Li Y., Rakesh V., and Reddy C. K. "Project success prediction in crowdfunding environments". In Proceedings of the Ninth ACM International Conference on Web Search and Data Mining, pages 247–256. ACM, 2016.

Mirtalaie M.A., Hussain O.K., Chang E., Hussain F.K. (2018) Sentiment Analysis of Specific Product's Features Using Product Tree for Application in New Product Development. Lecture Notes on Data Engineering and Communications Technologies

Prasetyo N. D., and Hauff C. "Twitter-based election prediction in the developing world". In *Proceedings of the 26th ACM Conference on Hypertext & Social Media,* pages 149-158. ACM, 2015.

Rao T., and Srivastava S. "Analyzing stock market movements using twitter sentiment analysis. In Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012), pages 119–123. IEEE Computer Society, 2012.

Razzaq M. A., Qamar A. M., and Bilal H. S. M. "Prediction and Analysis of Pakistan Election 2013 based on Sentiment Analysis". In *IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2014)*. IEEE, 2014.

Reynolds T. J., Gengler C. E. e Howard D. J (1995). . A Means-End Analysis of Brand Persuasion through Advertising, "International Journal of Research in Marketing", Vol. 12, No. 3, October, pp. 257–266.

Sang E. T. K., and Bos J.. Predicting the 2011 Dutch senate election results with Twitter. In Proceedings of the Workshop on Semantic Analysis in Social Media, pages 53-60. ACM, 2012.

Skoric M, and Poor N. “Tweets and Votes: A Study of the 2011 Singapore General Election”. IEEE, 2012.

Tuarob S., and Tucker C. S. “Fad or here to stay: predicting product market adoption and longevity using large scale, social media data”. In ASME 2013 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, Volume 2B: 33rd Computers and Information in Engineering Conference Portland, Oregon, USA, August 4–7, 2013.

Tumasjan A., Sprenger T. O., Sandner P. G., and Welpe I. M. “Predicting elections with twitter: what 140 characters reveal about political sentiment”. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*. AAAI, 2010.

Xu F., and Keselj V. “Collective sentiment mining of microblogs in 24-hour stock price movement prediction”. In IEEE 16th conference on Business Informatics. IEEE, 2014.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8337

Mining for Signals of Future Consumer Expenditure on Twitter and Google Trends

Pekar, Viktor
Finance Department, Business School, University of Birmingham, United Kingdom.

Abstract

Consumer expenditure constitutes the largest component of Gross Domestic Product in developed countries, and forecasts of consumer spending are therefore an important tool that governments and central bank use in their policy-making. In this paper we examine methods to forecast consumer spending from user-generated content, such as search engine queries and social media data, which hold the promise to produce forecasts much more efficiently than traditional surveys. Specifically, the aim of the paper is to study the relative utility of evidence about purchase intentions found in Google Trends versus those found in Twitter posts, for the problem of forecasting consumer expenditure. Our main findings are that, firstly, the Google Trends indicators and indicators extracted from Twitter are both beneficial for the forecasts: adding them as exogenous variables into regression model produces improvements on the pure AR baseline, consistently across all the forecast horizons. Secondly, we find that the Google Trends variables seem to be more useful predictors than the semantic variables extracted from Twitter posts, the differences in performance are significant, but not very large.

Keywords: *Google Trends and Search Engine data; Social media and public opinion mining; Internet econometrics; Machine learning econometrics; Consumer behavior, eWOM and social media marketing.*

1. Introduction

Consumer expenditure constitutes the largest component of Gross Domestic Product in developed countries: in the US, it accounts for about 70% of GDP, in the UK 66%, in Germany 60% (Pistaferry, 2015). Significant changes to consumer spending are key to predict the depth of a recession or the speed of recovery, and central banks use consumer spending forecasts as an important tool for monetary policy-making.

Government institutions and market research agencies compile their consumer spending indices on a regular basis. Among best-known examples are the University of Michigan Consumer Sentiment Index for the US, or the Household Final Consumption Expenditure by the UK Office for National Statistics. Currently, such indices are measured by market research surveys, but these have significant drawbacks: they are expensive to organize, they have sampling problems, the amount of effort required to collect and compile the data often entails that the indices are out of date by the time they are published.

This paper examines the hypothesis that user-generated content, such as search engine queries or social media posts, offers a better alternative to traditional surveys when it comes to estimating consumer expenditure. Effective methods to extract signals about future consumer spending from this data may help to produce forecasts more efficiently, based on much larger data samples, and in near-real time.

Previous work studied models of consumer spending trained on search engine data, based on the intuition that web searches for product names indicate intended purchases (Vosen and Schmidt, 2011; Scott and Varian, 2015; Wu and Brynjolfsson, 2015). Another direction of research has been to estimate economic confidence and purchase intentions of consumers from social media using automatic sentiment analysis (O'Connor et al., 2010; Daas and Puts, 2014; Najafi and Miller, 2016).

In this paper we study the relative utility of evidence about purchase intentions found in search engine queries versus those found in social media, for the problem of forecasting consumer expenditure.

2. Google Trends

The Google Trends (GT) site provides data on the volume of all Google queries based on geographic locations and time, collected since 2004. The frequencies of queries is not the absolute number of actual queries, but a normalized index, such that for any given retrieval criteria, the index is always between 0 and 100, 100 being the count of the most common query in the retrieved data.

GT contains data not only on individual queries, but also on categories of queries. In our study we use the data on search volumes of the 18 subcategories of the top-level "Shopping" category in GT. Examples of the subcategories are "Apparel", "Consumer Electronics", "Luxury Goods", "Ticket Sales". The search volume on each category is used as an exogenous variable in the Support Vector Regression model.

The time period we analyse spans 43 months (from 1st January 2014 to 31.07.2017). Because GT returns only weekly search volumes in one request for periods longer than six months, we are able to retrieve only weekly search volumes for the entire time period. To obtain daily search volumes, we first retrieve daily data in separate queries for each 6-month period. Then, within each such subset, we fit a linear regression on the monthly data and use it to obtain daily volumes for the entire 43 months dataset.

3. Purchase intentions on Twitter

Our method aims to predict a consumer spending index from the mentions of purchase intentions in Twitter posts. The method consists of the following steps. First, tweets mentioning a purchase intention are identified. Second, noun phrases referring to the objects of the intended purchases are extracted and represented as semantic vectors using the word2vec method. Finally, a regression model of the consumer spending index is trained that uses semantic vectors as explanatory variables. These steps are detailed below.

3.1. Detecting purchase intentions

To obtain tweets mentioning purchase intentions, we issue a set of queries to the tweet collection, which are meant to capture common ways to express an intention to buy something. They are created from combinations of (1) first-person pronouns ("I" and "we"), (2) verbs denoting intentions ("will", "'ll", "be going to", "be looking to", "want to", "wanna", "gonna"), and (3) verbs denoting purchase ("buy", "shop for", "get oneself"), thus obtaining queries such as "I will buy" or "we are going to buy".

The text of each tweet is then processed with a part-of-speech tagger. PoS tag patterns are then applied to extract the head noun of the noun phrase following the purchase verb (e.g., "headphones" in "I am looking to buy new headphones"). After that, daily counts of the head nouns are calculated.

3.2. Semantic vectors

To represent the semantics of the nouns, we use the word2vec method (Mikolov et al., 2013) which has proven to produce accurate approximations of word meaning in different NLP tasks. A word2vec model is a neural network that is trained to reconstruct the linguistic context of words. The model is built by taking a sequence of words as input and learning to predict the next word, using a feed-forward topology where a projection layer in the middle is taken to constitute a semantic vector for the word, after connection weights have been learned. The semantic vector is a fixed-length, real-valued pattern of activations reaching the projection layer. For each word, the input text originally has a dimensionality equal to the vocabulary size of the training corpus (typically millions of words), but the semantic modelling provides reduction to the size of the vector (typically several hundreds).

For each date, we map each noun that was observed on that day to a semantic vector, using 100-dimensional word2vec vectors trained on a large corpus of Twitter posts. The semantic vectors of all the nouns for each day are then averaged to obtain a single vector. The components of the vectors will then be used as exogenous variables in regression models.

To allow for some time between the stated purchase intention and the actual purchase, we experiment with the "intention lag", different numbers of days between the day on which intentions were registered and the day for which the value of the consumer spending index is predicted.

4. Experiments

4.1. Data

Indicator of Consumer Expenditure. As the forecast variable in our model, we use the Gallup Consumer Spending Index (CSI)[1]. The index represents the average dollar amount Americans report spending on a daily basis. The eventual index is presented as a 3-day and a 14-day rolling averages of these amounts. In our evaluation, we used the 3-day values of CSI, between January 1, 2014 and July 31, 2017, i.e. 1,310 days in total.

Twitter. For the same time period, we collected Twitter posts that originate from the US and that express intentions to buy, obtaining the total of 288,730 messages. Counts of nouns referring to purchases were extracted and rolling averages for each noun for three-day periods were calculated. To eliminate noisy data, we selected the 1000 most common nouns to construct semantic vectors.

Google Trends. Also for the same period, we obtain frequencies of searches in the 18 subcategories of the top-level Shopping category from the GT site, limiting the data to the US.

Train-validation-test split. The available data was divided into the training, validation and test parts, in proportion 60%-20%-20%.

[1] http://www.gallup.com/poll/112723/gallup-daily-us-consumer-spending.aspx

4.2. Modelling strategies

We experiment with four methods to ensure stationarity of the time series data: differencing, detrending, seasonal adjustment and detrending with seasonal adjustment. Detrending and seasonal adjustment are performed using the STL method (Cleveland et al., 1990). Before evaluating the quality of forecasts on test data, the forecasts are de-differenced, and the trend and the seasonal component estimated on training data are added to the forecasts.

4.3. Support vector regression

The Support Vector Machines learning algorithm (Cortes and Vapnik, 1995) is one of the most popular machine learning methods for supervised learning. In our experiments we use Support Vector Regression (SVR), a version of SVM adapted for regression. During evaluation, we experimentally determine free parameters of SVR (the cost parameter, the gamma parameter and the kernel type) on a validation dataset using the grid search technique. The model with the best parameter configuration is then evaluated on the test set.

4.4. Evaluation method

Once a model was trained on the training set and its parameters optimized on the validation set, it was evaluated on the test set using dynamic forecasting: given the first day *t* of the test set, and the forecast horizon *h*, the model predicted *h* days in the future, for each day from t_2 to t_h the values predicted by the model for previous days were input as endogenous variables. In the following, we report results for *h* = 7, 14, and 28. As evaluation metrics, we use the Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE).

As the baselines, we use SVR models trained with the same algorithms but only on endogenous variables, i.e. lagged values of CSI. Because CSI displayed weekly seasonality, we used seven lagged variables in the baseline model.

5. Results and discussion

5.1. Modelling strategies

An inspection of the correlogram of the CSI time series suggested that it is likely to have weekly seasonality. Furthermore, considering the long time period the data covers, the data may contain a trend. Therefore we first examined the effect of different techniques to “whiten” the time series on the quality of the forecasts. Table 1 details the results (the baseline refers to the raw original data, the best RMSE and MAE scores are in bold).

Table 1. Forecast accuracy for different time series transformation methods.

	h=7		h=14		h=28	
	RMSE	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
Baseline	12.78	9.9	14.42	11.23	14.42	11.38
Differencing	12.11	9.57	21.51	18.0	11.77	9.04
Detrending	10.64	8.18	11.49	8.9	11	8.42
Deseasonalizing	12.62	9.85	14.57	11.36	14.35	11.33
Detrend+Deseason	**10.5**	**7.96**	**11.48**	**8.64**	**10.96**	**8.23**

These results show that applying both detrending and seasonal adjustment consistently resulted in the best forecasting results, for all forecasting horizons. Therefore, in the subsequent experiments, the CSI data was detrended and deseasonalized.

5.2. Effect of Google Trends variables

We next examined the effect of supplying GT data as exogenous variables into the regression model, in addition to the autoregressive variables. The results are shown in Table 2 (improvements on the baseline are in bold).

We find that the GT variables do often perform better than the purely endogenous baseline, for all the three forecast horizons. It appears also that shorter intentions lags (between 0 and 4) produce better quality models. The best-performing model is a lag of one model, which beats the baseline by 3-7% across all the horizons.

Table 2. Forecast accuracy with GT variables.

	h=7		h=14		h=28	
Intention lag	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
0	**10.49**	8.03	**10.9**	**8.19**	**10.57**	**8.0**
1	**10.21**	**7.69**	**10.81**	**8.06**	**10.52**	**7.85**
2	**10.4**	**7.64**	**10.78**	**7.97**	**10.7**	**7.98**
3	10.56	8.0	**10.78**	**8.02**	**10.68**	**8.02**
4	10.84	8.3	**11.15**	**8.34**	11.5	8.63
5	10.57	8.06	11.52	8.65	11.62	8.73
6	10.46	7.99	11.48	8.65	**10.9**	8.24
7	11.19	8.58	**11.28**	**8.39**	11.09	8.33

5.3. Effect of Twitter variables

Table 3 presents results on the effect of semantic variables extracted from Twitter posts. As with GT variables, improvements are found across all the horizons. However, the baseline is consistently outperformed only when the intention lag is 0, and the improvements are more modest, ranging between 1 and 5%.

Comparing the performance of the models with GT variables and with Twitter variables, we observe that the GT model tends to fare better, but the gain on the Twitter model is not more than 0.3 points (3.6%) in either RMSE and MAE. Still, the differences in forecasts between the two types of models at corresponding horizons are statistically significant.

6. Conclusions

In this paper we presented a study comparing indicators of purchase intentions obtained from Google Trends to those obtained from Twitter using NLP analysis of the messages, on the task of forecasting consumer expenditure. Our main findings are that, firstly, both kinds of purchase intention indicators are beneficial for the forecasts: the improvements on the baseline are consistent across all the forecast horizons and in terms of both evaluation metrics. Secondly, the study found the Google Trends variables seem to be more useful predictors than the semantic variables extracted from Twitter posts, although the differences in performance are not very large.

Table 3. Forecast accuracy with Twitter variables.

	h=7		h=14		h=28	
Intention lag	**RMSE**	**MAE**	**RMSE**	**MAE**	**RMSE**	**MAE**
0	**10.34**	**7.76**	**10.92**	**8.21**	**10.88**	**8.15**
1	10.73	7.99	**11.37**	**8.49**	11.09	8.29
2	10.75	8.06	**11.46**	**8.61**	11.14	8.37
3	10.57	8.01	**11.3**	**8.55**	10.97	8.34
4	10.65	8.25	**11.33**	**8.57**	11.04	8.42
5	10.71	8.03	**11.31**	**8.46**	11.51	8.67
6	10.92	8.15	**11.32**	**8.43**	11.52	8.69
7	10.86	8.14	11.54	8.68	11.53	8.7

Future directions for this work may involve a further analysis of models that use Google Trends data, such as an analysis of more fine-grained Google Trends subcategories, automatic selection of the most relevant predictors among them, and their semantic clustering.

References

Cleveland R., Cleveland W., McRae J., & Terpenning I. (1990). STL: A Seasonal-Trend Decomposition Procedure Based on Loess. *Journal of Official Statistics*, Vol.6, No.1, 1990. pp. 3–73.

Cortes C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*. 20 (3): 273–297.

Daas P., & Puts M. (2014). Social media sentiment and consumer confidence. In *Workshop on using Big Data for forecasting and statistics*.

Mikolov T., Chen K., Corrado K., & Dean J. (2013). Efficient estimation of word representations in vector space. In *Proceedings of CoRR*.

Najafi H. and Miller D. (2016). Comparing analysis of social media content with traditional survey methods of predicting opening night box-office revenues for motion pictures. *Journal of Digital and Social Media Marketing*, 3(3):262–278.

O'Connor B., Balasubramanyan R., Routledge R., & Smith N. (2010). From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of ICWSM*.

Pistaferri L. (2015). Household consumption: Research questions, measurement issues, and data collection strategies. *Journal of Economic and Social Measurement*.

Scott S. & Varian H. (2015). Bayesian Variable Selection for Nowcasting Economic Time Series. In *Economic Analysis of the Digital Economy*, pages 119–135. University of Chicago Press.

Vosen S. & Schmidt T. (2011). Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*, 30(6):565–578.

Wu L. & Brynjolfsson E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic Analysis of the Digital Economy*, pages 89–118. University of Chicago Press.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8338

Towards an Automated Semantic Data-driven Decision Making Employing Human Brain

Fensel, Anna
Semantic Technology Institute (STI) Innsbruck, Department of Computer Science, University of Innsbruck, Austria

Abstract

Decision making is time-consuming and costly, as it requires direct intensive involvement of the human brain. The variety of expertise of highly qualified experts is very high, and the available experts are mostly not available on a short notice: they might be physically remotely located, and/or not being able to address all the problems they could address time-wise. Further, people tend to base more of their intellectual labour on rapidly increasing volumes of online data, content and computing resources, and the lack of corresponding scaling in availability of the human brain resources poses a bottleneck in the intellectual labour. We discuss enabling direct interoperability between the Internet and the human brain, developing "Internet of Brains", similar to "Internet of Things", where one can semantically model, interoperate and control real life objects. The Web, "Internet of Things" and "Internet of Brains" will be connected employing the same kind of semantic structures, and work in interoperation. Applying Brain Computer Interfaces (BCIs), psychology and behavioural science, we discuss the feasibility of a possible decion making infrastructure for semantic transfer of human thoughts, thinking processes, communication directly to the Internet.

Keywords: *Semantic Technology; Decision Making; Brain Computer Inerface; Data Value Chain; Artificial Intelligence; Data Management.*

1. Introduction

Decision making is time-consuming and costly, as it requires direct intensive involvement of the human brain. The variety of expertise of highly qualified experts is very high, and the available experts are mostly not available on a short notice: they might be physically remotely located, and/or not being able to address all the problems they could address time-wise. Generally, exchanging and managing the data on the Internet in a dynamic and efficient manner are among key challenges for the information systems requested nowadays by enterprises, institutions and citizens. People tend to base more of their intellectual labor on rapidly increasing volumes of online data, content and computing resources, and the lack of corresponding scaling in availability of the human brain resources poses a bottleneck in the intellectual labor. Finally, communication of the results of the intellectual labor requires further efforts, of putting the outcomes in a commonly processible representation form, such as spoken words or written texts.

To approach the optimal data management of the future, we discuss the possibility of ***enabling of direct interoperability between the Internet and the human brain, developing "Internet of Brains", similar to "Internet of Things"***. On the latter, one can semantically model, interoperate and control real life objects, and the applications of the semantic Internet of Things are numerous, see e.g. the areas of smart homes or transport. The Web, "Internet of Things" and "Internet of Brains" will be connected employing the same kind of semantic structures, and work in interoperation. Applying Brain Computer Interfaces (BCIs), psychology and behavioral science, ***an infrastructure for semantic transfer of human thoughts, thinking processes and communication directly to the Internet*** can be designed. This will facilitate the intellectual labor and its representation in human and machine readable forms, and address the aspects difficult to account so far, f.e. non-verbal communication.

Service-based enablers for discovery of interdependencies across human reasoning and senses and heterogeneous datasets for assisting humans in making decisions and changing their behavior and workflows can be created, as well as making these decisions and workflows more transparent and traceable. The latter can be performed taking into account the currently existing developments and standards in the related fields, particularly, semantic data licensing (Pellegrini et al., 2018), and smart contracts – as these are being exploited already broadly in practice (Underwood, 2016).

The ***high-level results*** of the envisioned solution will include:

- a framework for applying human thoughts and senses in decision making, through the semantic interfaces, and its concrete design and implementation,
- synthesizing concrete semantics from abstract thoughts and emotions, and

- automation of an intellectual labor (in the way the robotics is replacing manual labor), with employment of these capacities.

And the corresponding technical ***objectives*** are as follows:

- Design and development of a semantic infrastructure capturing the domain of human reasoning and senses, as well as decision making and intellectual work processes that are based on them,
- Mapping the output of state of the art BCIs to the semantic infrastructure, producing a corresponding mappings library,
- With the framework for streaming human brain activity online, enabling easier modeling of the data in both design time and the run time of the digital workplace scenario – and eventually the organizations creating their own applications and workflows basing on these models,
- Speed up the velocity of the data flow in an information system that are currently bottlenecked by the slow speed of human decision making abilities, or are even performed with mistakes due to their imperfection (e.g. in scenarios connected with reporting),
- Making the decision processes transparent, traceable, and easier to optimize (e.g. it can be easily established which nodes are causing delays),
- Integrate new techniques facilitating easier data reuse, such as semantic information on how the data and content can be licensed (licenses library and tools can be applied out of our development in DALICC project[1]),
- Visualization of the data, decisions in a form that is actionable to humans in a digital workplace scenario.

2. State of the Art and Progress Beyond it

The proposed solution will aim to advance the state of the art in the following areas: (1) semantic modelling, knowledge representation, (2) data-empowered reasoning and decision making, (3) sensor technology. Further, we overview of the state of the art in these fields and how the aimed results are expected to advance the state of the art.

[1] DAta LIcenses Clearing Center: https://dalicc.net

2.1. Semantic Technology as a Communication Means on the Internet

"The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation." - this statement of Tim Berners-Lee has gained even more relevance since the start of this century (Berners-Lee et al., 2001). The vision of the Semantic Web seminal paper came closer, starting with the appearance of the basic semantic languages such as RDF, RDFS, OWL, semantic web service languages. Early from the appearance of the Semantic Web, the challenge to use semantics to facilitate human communication is being addressed, with the "semantic desktop" initiative and NEPOMUK project (Decker and Frank, 2004) being among the first ones. Now there exist IT multi-stakeholder ecosystems and infrastructures to interoperate across different marketing data and content resources using Linked Data (Bizer et al., 2009) and semantic technologies (Domingue et al., 2011), enhancing interoperability of distributed resources for allowing meaningful searches and efficient information dissemination for humans alike as machines. Research and developments on combining human and computing resources are abundant – see e.g. developments such as "social machines" (Hendler and Mulvehill, 2016), and use cases of that kind in the infrastructures such as Wikipedia (Smart et al., 2014), however, none of them yet comprise processing of the direct input from the human brain.

In this development, skipping of the step involving natural language processing technology and communicating the outcomes of human thinking to machines in a semantic form will become possible. This will be used in the first place by the people with limited abilities, but also by people who are able to represent their thinking in an intermediate representation format (spoken words, written texts, etc.) – for the reasons of efficiency and scalability.

In research and commercial developments, methods and tools to extract semantics from intermediary communication means have been developed, for example, extracting emotions and sentiments from the Web (Baldoni et al., 2012), as well as from the natural language texts (Mathieu, 2005). Such methods work already work in practice with a relatively large success, but inherently presume the availability of intermediary knowledge representation sources. On the contrary, here, extraction of emotions and sentiments from the human brain would take place directly. Some of the modeling and representation of the sentiments and emotions from the state-of-the-art research can be taken into account when modelling the framework, also including dedicated efforts to build the relevant ontologies (López et al., 2008; Borth et al., 2013).

2.2. Reasoning and Decision Making in the "Data Tsumami" Conditions

As it is known, in the human reasoning and decision making, there are normally "soft" and "hard" factors involved. For example, if someone is hiring an employee to work with, usually both sides are important: whether a potential employee has an adequate qualification and experience (“hard” factors), and whether he or she would fit well in the team (“soft” factors). Frameworks and models are currently starting to appear in the literature in application to various tasks and domains e.g. forecasting (Bańbura & Rünstler, 2011), as well as the approaches towards explaining human decision making (Rosenfeld and Kraus, 2018). Such works approach the possibilities to formalize the decision making and reasoning process semantically.

In the world overflown by a “data tsunami”, *humans are standing at the edge of their decision making and behavior change capacities*, and the need to overcome these is unavoidable. The reasons here are as follows:

- Drastic increase in the amounts of the data and information that can be potentially relevant for making right decisions, de facto, the current “hard” reasoning we perform now is mostly always “incorrect and incomplete” – and methods and tools to address such reasoning (Fensel et al., 2008) have been developed, particularly, in EU LarKC project[2],
- Limitations and restrictions of human mind in taking decisions (such as due its limited immediate storage capacity, irrationality caused by bias (Boutang and De Lara, 2015)),
- Increase in the dynamicity: often, the situations change on the fly, and the used data, workflow models may need to be replaced – as well as the behavior changed, this again, poses a challenge in choice of goals, methods, and implementation of tasks to a human brain,
- Effective decisions and human behavior changes are essential parts of success, in particular economic success; even more dramatic: in some areas such as climate change or energy efficiency, the change of human behavior may mean the difference between “to be” and “not to be” for the human kind.

Given the ability to process and analyze large amounts of data, the *machines already arguably outperform humans when it comes to the intellectual labor, where only “hard” factors are involved*. However, many decisions carried out solely on the “hard” facts remain unviable in the real world, as they may go in contrary with the human senses, emotions, feelings, intuition, and eventually safety of the humans and acceptance with them. Applying on emotions (fear, curiosity, enjoyment and many others), human brains are able to rather successfully filter out the “right” contexts and defining the new ones (Kahneman, 2011), i.e. possessing the “soft factor” capacity which machines do not possess.

[2] Large Knowledge Collider: http://larkc.sti2.at (archived web site)

Online communication, on the other hand, is not trivial, as it still hides most, or a large part of the semantics, e.g. transferred over non-verbal communication in face-to-face communication. Leaving alone the fact that a human, in order to communicate, needs to create a representation of the thought or an emotion, e.g. spoken words, images, text, which is of course not 100% identical to the original thought or an emotion. Here, we will pursue elimination of the typical intermediate representation layer for a human brain activity, and will create a precise machine and human readable semantic layer for it instead, and map the signals coming out of the bio-sensing equipment directly into this layer. In communication infrastructures, heterogeneous communities of stakeholders need to be addressed, and semantics is a very suitable instrument for this, as the essence of ontologies inseparably reflect the communities using them (Mika, 2005; Zhdanova, 2008).

An additional challenge here is that humans also have a tendency to conceal the outcomes of their thinking, or even communicate the facts that do not correspond to them, if they feel like they would be getting an advantage, in particular, a match to the desired limited resource on the market (Roth et al., 2015), or a better perception by the society. Again, semantics has a potential to resolve this challenge, and test/simulate the realities which would take place under the conditions of humans expressing their actual thoughts and feelings.

2.3. Hardware and Sensors Availability

A better understanding of the human brain stands high on the priority of the European Commission, e.g. The Human Brain Project[3] is ongoing as a H2020 FET Flagship Project which strives to accelerate the fields of neuroscience, computing and brain-related medicine, since 2013, with the duration of 10 years. On the side of the technical development, also the BCIs have been investigated, and the forecasts and scenarios have been roadmapped, confirming the expected broad spread and varying spectrum of the application scenarios where BCIs will be used (Brunner et al., 2015).

Now it is the right time to base the project on this technology, due to the following technical reasons:

1) Now the BCI technology is becoming mature and available. Companies selling advanced consumer-oriented products in the sub-1000 dollar range include Emotiv (http://www.emotiv.com) and Neurosky (http://www.neurosky.com). There are also very inexpensive open source biosensing solutions, such as OpenBCI (http://openbci.com). Generally, currently tools for make technical connections to the brain cost as little as starting from 30 dollars.

[3] Human Brain Project: https://www.humanbrainproject.eu/en/

2) The user acceptance level of the technology is also becoming sufficient, to have an expectation that a system developed on a base of a BCI will be used and usable.

Now BCI are already actively used beyond the typical for them medical scenarios, and e.g. are employed in games – however, also there a systematic approach of the related data management is missing (Gurkok et al., 2017). Thus, a systematic approach for integrating human thinking and reasoning activity on the Internet needs to be designed and developed practically.

3. Conclusions

We have described initial principles and prerequisites of the direct integration of the data stemming from the human brain into decision making processes of the future. The main measurable success criterion of this work can be characterized as ***a transition step from Big Data to Smart Data***. It typically involves enablement of more efficient adding value participation, i.e. increasing efficiency and/or provisioning and take up of new types of interactivity which bring benefits to the involved stakeholders (for example, faster decision making, less effort to transform the brain activity into the intermediary communication formats such as spoken words or written text). Further, the societal externalities of Big Data use (Cuquet and Fensel, 2018) will be accounted for, even when humans are unaware of them. The approach is to be realized in the chosen application domains going beyond the current state of the art of ontology-based service interfacing, integration and bio- and crowd- sensing. The results are to be evaluated within real scenarios, with real life data and services, as well as with real end users. The evaluation outcomes are to confirm the technical feasibility of ontology-based intervention networked services enablement, as well as its added value from the originated new usage scenarios, and its acceptance by the end users.

Acknowledgements

This work has been partially funded by project DALICC, supported by the Austrian Research Promotion Agency (FFG) within the program “Future ICT”.

References

Baldoni, M., Baroglio, C., Patti, V., & Rena, P. (2012). From tags to emotions: Ontology-driven sentiment analysis in the social semantic web. *Intelligenza Artificiale*, *6*(1), 41-54.

Bańbura, M., & Rünstler, G. (2011). A look into the factor model black box: publication lags and the role of hard and soft data in forecasting GDP. *International Journal of Forecasting*, *27*(2), 333-346.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). The semantic web. *Scientific American*, *284*(5), 34-43.

Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data-the story so far. *Semantic Services, Interoperability and Web Applications: Emerging Concepts*, 205-227.

Borth, D., Chen, T., Ji, R., & Chang, S. F. (2013, October). Sentibank: large-scale ontology and classifiers for detecting sentiment and emotions in visual content. In *Proceedings of the 21st ACM international conference on Multimedia*, 459-460. ACM.

Boutang, J., & De Lara, M. (2015). *The Biased Mind: How Evolution Shaped Our Psychology Including Anecdotes and Tips for Making Sound Decisions*. Springer.

Brunner, C., Birbaumer, N., Blankertz, B., Guger, C., Kübler, A., Mattia, D., del R. Millán, J., Miralles, F., Nijholt, A., Opisso, E., Ramsey, N., Salomon, P., & Müller-Putz, G.R. (2015) BNCI Horizon 2020: towards a roadmap for the BCI community. *BCI Journal*. URL: http://bnci-horizon-2020.eu/roadmap

Cuquet, M., & Fensel, A. (2018). The societal impact of big data: A research roadmap for Europe. *Technology in Society*, Elsevier.

Decker, S., & Frank, M. (2004). The social semantic desktop. *Digital Enterprise Research Institute, DERI Technical Report May*, *2*, 7.

Domingue, J., Fensel, D., & Hendler, J. A. (Eds.). (2011). *Handbook of semantic web technologies* (Vol. 1). Springer Science & Business Media.

Fensel, D., Van Harmelen, F., Andersson, B., Brennan, P., Cunningham, H., Della Valle, E., Fischer, F., Huang, Z., Kiryakov, A., Kyung-il Lee, T., Schooler, L., Tresp, V., Wesner, S., Witbrock, M., & Zhong, N. (2008). Towards LarKC: a platform for web-scale reasoning. In *IEEE International Conference on Semantic Computing*, 524-529, IEEE.

Gurkok, H., Nijholt, A., & Poel, M. (2017). Brain-Computer Interface Games: Towards a Framework. *Handbook of Digital Games and Entertainment Technologies*, 133-150.

Hendler, J., & Mulvehill, A. M. (2016). *Social machines: the coming collision of artificial intelligence, social networking, and humanity*. Apress.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

López, J. M., Gil, R., García, R., Cearreta, I., & Garay, N. (2008, September). Towards an ontology for describing emotions. In *World Summit on Knowledge Society*, 96-104, Springer, Berlin, Heidelberg.

Mathieu, Y. Y. (2005, October). Annotation of emotions and feelings in texts. In *International Conference on Affective Computing and Intelligent Interaction*, 350-357, Springer, Berlin, Heidelberg.

Mika, P. (2005). Ontologies are us: a unified model of social networks and semantics. In: *Proceedings of the 4th International Semantic Web Conference*, LNCS 3729, 522–536, Springer.

Pellegrini, T., Schönhofer, A., Kirrane, S., Steyskal, S., Fensel, A., Panasiuk, O., Mireles-Chavez, V., Thurner, T., Dörfler, M., & Polleres, A. (2018). A Genealogy and

Classification of Rights Expression Languages – Preliminary Results. In: *Trend and Communities of Legal Informatics - Proceedings of the 21st International Legal Informatics Symposion*, IRIS 2018, 243-250, Salzburg, Austria.

Rosenfeld, A., Kraus, S. (2018). *Predicting Human Decision-Making:* From Prediction to Action. Morgan and Claypool.

Roth, A. E. (2015). *Who Gets What—and Why: The New Economics of Matchmaking and Market Design*. Houghton Mifflin Harcourt.

Smart, P., Simperl, E., & Shadbolt, N. (2014). A taxonomic framework for social machines. In *Social Collective Intelligence*, 51-85, Springer, Cham.

Underwood, S. (2016). Blockchain beyond bitcoin. *Communications of the ACM*, *59*(11), 15-17.

Zhdanova, A.V. (2008). Community-driven Ontology Construction in Social Networking Portals. *International Journal on Web Intelligence and Agent Systems*, 6(1), 93-121, IOS Press.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8341

Limits and virtues of a web survey on political participation and voting intentions. Reflections on a mixed-method search path

Faggiano, Maria Paola
Department of Communication and Social Research (CoRiS), Sapienza University, Italy, Rome

Abstract

The Internet offers new opportunities for the empirical research, especially if we consider that nowadays most citizens are made up of web surfers: on the one hand, we are seeing the transfer of some traditional methodologies on Internet, on the other hand we are witnessing the development of new innovative data collection and analysis tools. The study was conducted through a classical survey tool (the questionnaire), using it as part of a web survey. Secondly, we chose Facebook as an instrument which is particularly suitable for the investigated topic (political participation and voting intentions), because the election campaign for the 2018 Italian general election took place, for all parties and candidate leaders, mainly on this Social Network. Two surveys were carried out, the first one in September 2017 and the second one in February 2018, reaching about 850 and 1,400 cases, with similar percentages over the whole block of variables and with stable connections among them. The aim is to highlight the advantages and disadvantages of a Web survey on the topic of political participation, showing particular attention to strategic choices and decisions that impact positively on the data quality, according to a mixed-method approach.

Keywords: *web survey; political participation; voting intentions; Social Network; mixed method approach.*

1. Old and new sociological survey tools in the Digital Age: A focus on the Web survey

The rapid spread of digital technologies in each dimension of everyday life has inevitably produced changes in practices, styles, relationships and social interactions. The interest of human sciences researchers in the effects produced by the digitalization process is growing; at the same time the Internet offers new opportunities for empirical research (Cipriani, Cipolla, Losacco, a c. di, 2013; Boccia Artieri, a c. di, 2015): on the one hand, we are witnessing the transfer on the Web (and the adaptation process) of some traditional methodologies (primarily the questionnaire), on the other hand we can observe the development of new innovative data collection and analysis tools. The Internet represents a great chance on the political participation side and for many other topics of sociological interest, considering that nowadays the majority of citizens are made up of Internet users, who are so easily reachable for the purpose of compiling a questionnaire (in Italy today 67% of Italians surf the Web – Report of Demopolis, 2017). However, there is still a generational gap, in fact, only young people use the Internet in almost all cases; even if the "digital divide" is gradually being reduced, through the entry into the mass consumption of mobile devices such as smartphones and tablets, also for the elderly population. Instead, it is more and more complex, for the purpose of conducting surveys, to draw up reliable population lists from which to extract random samples, or to acquire updated and complete telephone registers in order to make telephone or face to face interviews. Moreover, we cannot overlook a detail which applies both to online and offline surveys: in many cases excessive circulation of investigations, opinion polls, market research, etc. have made the voter-citizen elusive and uncooperative, always less inclined to be interviewed. Sometimes voters are very suspicious or unavailable to any initiative that requires on his part to answer to a set of questions (sometimes the respondents show that they do not know how to distinguish the request for collaboration to sociological surveys from that one coming from direct marketing activities). There are in fact many online questionnaires: 1. prepared in the launch phase of a new product/service; 2. designed to record the customers satisfaction about a service or to collect the students approval in a school/university; 3. addressed to citizens to know points of view and needs with respect to a collective service; 4. prepared by political parties to mobilize and get to know voters, etc. Focusing specifically on the Internet, although a self-selected sample poses known problems in terms of its statistical representativeness in comparison with its population, the principle of maximum freedom granted to the respondent seems to produce positive effects in terms of fidelity and quality of the collected data and also of the success of a research initiative: a subject is free to choose to participate or not in an investigation, "released" from the presence or voice of the interviewer at the time of compilation, free to choose the moment compiling into a larger time unit. However, despite a careful selection of the platform and the specific virtual space that hosts a web survey, although there is a continuous monitoring of the phase of data collection by a research team, the participatory outcome of a study condicted on the web is in hands of a rather low number of respondents, considering that the average yield of these surveys is normally less than 1% of the reached contacts. A limited number of completed questionnaires is recorded in many studies in which the online questionnaire was used (Faggiano, 2007; Peruzzi, 2017; Vaccari *et al.*, 2013). Therefore, a sample - more or less consistent and close to the characteristics of the population - will be consist those people who will choose to adhere to the research activity during the period of the data collection. Their participation is based on their available time, their sensitivity and interest, their own skills about the treated topics. Let's dwell on a peculiarity of the online questionnaire, connected with the typical interactivity of virtual environments. It is interesting to note that a respondent can actively participate in the research experience, not limited to the mere filling out his specific questionnaire. Let's imagine an online questionnaire published on a Facebook page: in this case, the respondents can interact via the web with the posts containing the link of the questionnaire, for example by adding a positive or negative reaction towards the initiative of research and/or the sponsoring institution. They can comment, start a debate with the research team and/or other respondents, share or discredit other people’s arguments; write privately to the institution that initiated the investigation, contribute to the circulation of the research initiative through online sharing modes.

2. Doing research to evaluate the survey approach: a voters-based study

In making an evaluation on the chosen survey strategy, argued and supported by data, the study we present was conducted through a classic surveying tool such as the questionnaire, using it as part of a Web survey. More precisely, a post containing the link to the questionnaire was published on the Facebook institutional page of the Department of Communication and Social Research of the Sapienza of Rome. The post was previously sponsored in order to reach social profiles as heterogeneous as possible. These profiles were based on social extraction charateristics, professional activity, level of education and hobbies and interests. The post was also shared on other online platforms and channels. The choice of Facebook is obviously not accidental: it is the most used platform in Italy (and not only), with 33% of the registered population, followed by WhatsApp and Messenger (services also widely used by the research team for the purpose of disseminating the questionnaire, in addition to Telegram, Instagram, LinkedIn, Twitter, mailing lists). Moreover, Facebook is particularly suitable for the treated topic (political participation and voting intentions), because the election campaign for the 2018 Italian general election took place, for all parties and candidate leaders, mainly on this Social Network. Compared to the theme of research, Italy has a large tradition of empirical studies, generally carried out with standardized methodologies of questioning of voters (think of the activity of the Italian

Society of Electoral Studies or the Italian National Election studies). On the eve of the 2018 political elections, the dimensions in analysis were: values, sense of legality, idea of social justice, trust in the institutions, social resentment, social problems perceived as urgent, political orientation and electoral behaviour over time, traditional forms of political and social participation, forms of online political participation and hybrid styles. Two fairly close surveys were carried out, one in September 2017 and the other one in February 2018, reaching about 850 cases in the first case and about 1,400 units in the second one, with similar percentages over the whole block of variables and with stable connections among them. The aim is to highlight the advantages and disadvantages of a web survey on the participation topic, showing particular attention to those strategic choices and decisions that have positive effects on the data quality, in a mixed method approach. Although the reached samples cannot be considered representative from the statistical point of view (these are self-selected samples), we want to underline the particular attention used towards the data quality (specially for the phase of design of the data collection tool, the pretest and the data analysis of the first survey, which has been essential to calibrate the questionnaire with a view of its second use and of the improvement of the whole data collection strategy). We have the numbers to analytically describe and deepen opinions, attitudes, values and social practices related to the most significant Italian electoral targets (*left* area, *right* area, Five Star Movement, Non-voting area, area of indecision), in addition to the typical socio-demographic and economic variables. Furthermore, the offline use of the questionnaire (approximately one fifth of the interviews for each survey) has been an effective solution (Kott, Chang, 2010) for reaching a marginal voter who still exists in Italian society (often elderly and with a low level of education): the non-user of the Web (today it is estimated that a quarter of the Italian population does not use the Internet at all).

3. Measures for improving data quality

In a synthetic way, it is useful to spend a few words about all the measures that the research team has developed in order to obtain reliable data and to widen to the maximum the respondents' catchment, not only from the strictly quantitative point of view (the numerical consistency of the sample), but also with reference to its heterogeneity regarding the strategic variables related to the political participation theme (as a problem of statistical coverage). In fact, the Web data collection mode is definable as "device agnostic", that is the questionnaire that we used in the survey is suitable to be compiled either from PC, tablet or mobile device (for example, the scales of attitude originally prepared with scores from 0 to 10, were brought back to a 0-5 range, in order to optimally display each possible mode of response, even from smartphones). Through a meticulous pre-test on a heterogeneous sample (by age and level of education), conducted both online and offline (with a questionnaire in paper format), we made improvements to the wording and the formulation of the precoded answers. Furthermore, we worked on the order and the number of questions, as well as on the closure of some questions, etc. We have already specify the reasons below the choice of the most used social platform, Facebook, and of the overall strategy of sharing the questionnaire on other channels (Social Networks and mailing lists) in the selected temporal unit (one month for the first and for the second data collection; in the latter case it was the period of the election campaign in a technical sense). With reference to the aspect of sharing, by following the writing of a long list of themes and key words, we tried to identify some Facebook groups connected with social activities and diversified interests, having the electoral targets as a point of reference. It was not easy to penetrate into these groups, especially for increasingly restrictive rules on privacy and spamming. So, an element that we had held in high regard for the success of the survey research, did not help matters much. The surveys conducted in specific groups to which you belong and within which you declare the research activity in an explicit way are very different (case studies, surveys on circumscribed themes, etc.). Many undergraduates in social sciences take advantage of their belonging to youth groups (university, music, etc.) to collect data on lifestyles of their peers; these experiences' success depends both on their role (symmetrical), and on the composition of the target audience (homogeneity compared to a hobby or to age, etc.). Getting the attention of large and heterogeneous populations is more complicated, as it is inevitable that some targets are more receptive than others. The sponsorship of the post, spread throughout the entire research timeframe with a medium economic investment, has led to contain the initial distortion of the sample, that was introduced by a sharing mode of the post based on an university circle and related to the local context of the Lazio Region. This attempt to curb the problem has made possible to reach subjects scattered throughout the whole national territory, differentiated by title of study, age, gender, hobbies and interests. Compared to the second survey, which totaled 1,400 respondents, the statistics on involvement are as follows: more than 30,000 views, about 2,000 interactions with the post, about 100 reactions (especially likes) and hundreds of comments (as expressions of distrust or annoyance towards the research initiative, real invective in the complaining of the Italian political system and its political representatives, expressions of support towards particular parties/declarations, interactions and sometimes animated discussions among commentators of different political orientations, etc.). As mentioned, the precise evaluation of the first survey results allowed to refine the questionnaire of the second survey, with the goal of being able to compare the data of the two research rounds. For example, we refined the text of questions and answers in the direction of clarification and simplification; sometimes we changed the order of items and replies in case of unreliable data; we eliminated items and response modes in the case of redundancies, of excessively unbalanced data, of distortions linked to social desirability, and also in the direction of thinning and agility of the instrument. All the open/semi-open questions of the first survey were closed (an example for all regards the question about the motivations related to voting intentions). The accompanying post of the link (through the explicitation of the research theme, of the institutional subjects involved, of the time required for compiling, etc.) has been prepared in a very accurate way. The post has been "placed at the top" within the Department page for the entire duration of the research, the same time span in which the team has constantly monitored the input data and motivated individuals to participate. The module we prepared for the completion of the questionnaire has graphic characteristics that make it aesthetically pleasing. The number of questions is not excessive. There are numerous indications for the correct filling of the questions. Moreover, the technology in use intervenes in order to reset the different erroneous and partial forms of compilation, that are not containable in the case of the paper questionnaires completion: this makes the online completion, for many ways, simpler and more fruitful (few errors and few missing data) than the offline one.

4. The advantages and disadvantages of online data collection

In considering the advantages of a Web survey (Lombi, 2015) we can count on the substantial containment of the research costs (Groves, 1989) on several levels: the sampling tools, the minor human resources employed, the saving of printing questionnaires/doing phone calls, the avoidance of a wide territory mobility, etc. The Internet also has great potential to spread the data collection instrument in a wide and variegated territorial context. The subjects involved in the survey are automatically listed as records of a data matrix. In other words, we have an immediate availability of data for subsequent analyses. We have also already referred to the greater accuracy of the data collected (absence of errors of insertion, absence of errors of compilation, containment of the missing values), as well as their immediate availability in matrix and the possibility of monitoring the results *in itinere*. Obviously, there are also disadvantages: the first one is the statistical non-representativeness of the samples, due to the fact that not all the population uses the Internet and Facebook, while, the second one regards the mechanisms of autoselection and the effect "ball of snow" triggered by the sharing system. The borders of the universe of Facebook subscribers are not defined and, moreover, are constantly evolving; in addition, there is no coincidence between the Social network world and the respective universe of voters present in a given territory, without taking account of the system of basic features of Facebook users, that are not known in a precise way. Finally, we cannot avoid multiple completions made by a unique subject, although we have the possibility to identify and correct them. In fact, in relation to the study on voting intentions in the electoral campaign phase, there are a series of distortions: those due to the applied techniques (standardized questionnaire published on the net) and some other distortions related to the topic (let's think to distrust and resentment towards politics in this historical moment, or to the distrust of some voters, or again to the need for privacy on the personal intentions of voting, etc.).

In particular, in thinking of a negative combination of "technical effect" and "theme effect", there are electoral targets difficult to contact, sometimes impossible to reach. These are the elderly, the non-internet users, the *right* and extremists voters (who show particular distrust, specific cultural and valuable characteristics and sometimes are afraid to express themselves on ideas and practices that they consider to be wholly personal), people with low degree of study, foreigners. On the other hand, we can observe young people, subjects with a high level of education, *left* voters, people who are interested in politics and rather well-informed (it emerges a highly motivated "respondent-type", that is sensitive to research initiatives and interested in politics). Obviously, some distortions in the sample composition, if contained, can easily be corrected through an appropriate weighting.

5. Concluding notes: Online data collection is not enough

It can be concluded that, without solving every criticality emerged but making the utmost effort in the direction of the quality of the achievable empirical basis, the only assumable perspective is the mixed-method approach (Dillman, Smith, Christian, 2009; Amaturo, Punziano, 2016). In our study we highlight the need to combine an offline data collection instrument with one online, taking for granted the same basic characteristics for both (same questionnaire, same mode of administration), in order to address the problem of the coverage of absent or under-represented targets. The email addresses provided for free by a part of the people that we reached online, allowed to proceed with interviews in depth about the anomalous aspects that emerged during the data analysis. The most collaborative subjects have sometimes offered their support to help us in reaching other subjects poorly represented in the sample, with the same characteristics or not, in order to involve them in the investigation. For some aspects, the statistical non-representativeness of the sample gives way to a strong representativeness regarding the substantial plane (Di Franco, 2010). In considering the aim of securing a numerical consistency for each of the electoral targets - known to be diffused in Italy in this historical moment -, we have committed ourselves in this survey (whose publication of the extended version is being set up) to obtain the maximum heterogeneity of socio-demographic and economic variables, despite the distortions generated by the online publication of the questionnaire. All this has been done to be able to accurately investigate values, opinions, perceptions, behaviours in their synergy, with the purpose of identifying electors' styles and profiles who are prevalent and recognizable in our society. The statistical non-representativeness of the sample certainly does not interfere with the theoretical deepening of the theme of political and social participation, neither with the testing and refining of the instruments of data collection, with the activity of conceptualization and operation, with the study of social trends prevailing through the identification of interconnections among variables at multivariate level. On the other hand, a description, a typing and an accurate interpretation of data is a valuable and fundamental basis also for the preparation of explanatory models and for the identification of predictive factors. All this is based on a deep knowledge of the social and political context of reference, also resulting from the ability to interconnect different and complementary data (for example, analysis of the electoral campaign; analysis of attitudes, voters' intentions and perceptions, aggregate analysis of the outcome of an electoral session, etc.). Mauceri (2003) says about it: "If (...) probabilistic sampling is irreplaceable in research situations in which it is intended to estimate precisely what is the numerical consistency of the diffusion of certain traits within a given population, as when we made opinion polls, the assessment of the relations among variables can require to privilege a comparative logic rather than generalizing, aimed, for example, to compare groups of subjects with opposite action orientations (...) and to establish which are contextual, relational and individual elements that make their courses of action so different".

References

Amaturo, E., Punziano, G., (2016). *I "Mixed-Methods" nella ricerca sociale*, Roma: Carocci.

Boccia Artieri, G., (a cura di), (2015). *Gli effetti sociali del web. Forme della comunicazione metodologie della ricerca sociale*, Milano: Franco Angeli.

Cipriani, R., Cipolla, C., Losacco, G., (a cura di), (2013). *La ricerca qualitativa fra tecniche tradizionali ed e-methods*, Milano: Franco Angeli.

Di Franco, G., (2010). *Il campionamento nelle scienze umane. Teoria e pratica*, Milano: Franco Angeli.

Dillman, D.A., Smyth, J.D., Christian, L.M., (2009). *Internet, mail and mixed-mode surveys. The tailored design method*, San Francisco: Jossey-Bass.

Faggiano, M.P., (2007), *La formazione sociologica nell'università della riforma. La domanda, i percorsi, l'offerta presso il CdL in Sociologia della Sapienza di Roma,* in Fasanella, A. (a cura di), (2007), Milano: Franco Angeli.

Groves, R.M., (1989), *Survey errors and survey costs*, New York: Wiley and Sons.

Kott, P.S., Chang, T., (2010). Using calibration weighting to adjust for nonignorable unit nonresponse, *Journal of the Americn Statistical Association*, 105 (491): 1265-1275, DOI: 10.1198/jasa.2010.tm09016.

Lombi, L., (2015). *Le wen survey*, Milano: Franco Angeli. *Gli effetti sociali del web. Forme della comunicazione e metodologie della ricerca online*, Milano: Franco Angeli.

Mauceri, S., (2003). *Per la qualità del dato nella ricerca sociale. Strategie di progettazione e conduzione dell'intervista con questionario*, Milano: Franco Angeli.

Peruzzi, G., (a cura di), (2017), *Le reti del Terzo Settore. Terzo Rapporto,* Roma: Forum Nazionale del Terzo Settore.

Vaccari *et al.*, (2013), A survey of Twitter users during the 2013 Italian general election, *Rivista Italiana di Scienza politica*, Anno XLIII, n. 3: 381-410, DOI: 10.1426/75245.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8343

Italian general election 2018: digital campaign strategies. Three case studies: Movimento 5 Stelle, PD and Lega

Calò, Ernesto Dario; Faggiano, Maria Paola; Gallo, Raffaella; Mongiardo, Melissa
Department of Communication and Social Research (CoRiS), Sapienza University, Rome, Italy.

Abstract

The advent of the Network Society has brought substantial transformations also in the politics, which, like other areas of society, is affected by important changes. The network, which regulates social relations, has become the place of political discussion and that is where the most substantial part of the electoral campaign for the 2018 general election took place. The object of our research is the observation of the political propaganda of the Movimento 5 Stelle, the Partito Democratico and the Lega (the three most voted parties in the Italian elections) through the institutional accounts of the political parties on Facebook. Once collected a research sample of 1,397 posts officially published online on the three monitored accounts, the aim of our analysis is to investigate the communication strategies of the parties in a phase of hybrid democracy crossed by a deep crisis of political representation. From our analysis it emerges how the three political forces, that refer to different electorates, organize their electoral propaganda, each according to their own strategy.

Keywords: *Political election 2018; Networked politics; Digital campaign; Movimento 5 Stelle; Partito Democratico.*

1. Introduction

On the eve of the vote for the 2018 general election, Italy, like the other mature Western democracies, is undergoing a deep crisis of political representation (Manin, 2016). This crisis is linked to several factors, including: the end of ideologies that has progressively transformed the traditional forms of political participation (Fukuyama, 1992); the digitalization of the social sphere that led to the birth of the Network Society (Castells, 2009) and of the networked individualism (Raine, Wellman, 2012); the distrust of political parties and institutions that operate in a political arena regulated by marketing and communication and which sees in social networks the real places of political participation and discussion.

It is a new society characterized by a hybrid environment. It results problematic for traditional political forces that are dealing with a metamorphosis of representative democracy, of its languages and its communication (Chadwick, 2013).

In addition to this, even from a strictly political point of view, the scenario appears to be complex: the affirmation of the *Movimento 5 Stelle*, as a post-ideological party, has contributed to erode the consensus of the traditional parties and has broken the traditional bipolarity regulated by the *right*/*left* alternation. The splitting of the *Partito Democratico* has led to a fragmentation of the left-wing area, unable to build a unitary political proposal. Abstentionism and non-voting represent a consistent block of voters who in a consolidated democracy are to be considered as a real political force that claims its right to not choose (Manin, 2016). Italy, in line with international and European political trends, is object of a return to nationalist sentiments (Holtz-Bacha, 2016) and, in the European context, represents one of the most interesting case studies of populism. *Populism*, as a consequence of the conflict between people and the elite (Diamanti, Lazar, 2018). Logically, all this events affect on the tones and the languages of political parties, in a phase definited “of permanent postmodern electoral campaign", characterized by the use of the network and by a fluid electorate that must be struck by the tones of a captivating communication (Norris, 2000).

According to the Law[1], the 2018 election campaign is the first in which political parties receive no public funding from the State; therefore, it is an electoral campaign that needs "zero cost" instruments and that sees in the social networks the most democratic communication tool equally available to all political parties.

[1] Decree Law December 28, 2013 n. 149.

The electoral result has given back to Italy an apparent condition of ungovernability: a tripolar scenario characterized by very different political forces and by their irreconcilable nature.

Starting from this assumption, our attention is focused on the analysis of the online campaign of the *Movimento 5 Stelle*, the *Lega* and the *Partito Democratico*, observed and analyzed through the institutional profile of the political parties on Facebook, with the aim of identifying a clear map of the communication strategies of each political parties and of observing the flow of communication addressed to the voters without the mediation of third parties.

This research interest originates from the hypothesis that the new mediated political scene, as an object of substantial distortions, regulated by an immediate communication, represents the privileged seat of negative communication strategies and attacks of the opponents, who aim to obtain the consensus by pointing at the elaboration of strongly emotional and not very rational messagges.

2. Research methodology

The object of our analysis is the online electoral campaign officially managed by those political parties that have obtained the highest number of votes: the *Movimento 5 Stelle*, the *Lega* and the *Partito Democratico*. We have chosen to focus our attention on Facebook, because it is the most widespread social network and it is the one that allows a more direct interaction between politicians and users and also because it is a real social marketing tool, as an indispensable tool to communicate the politics and its propaganda. Our monitoring consisted in the collection of all posts and contents published by the political parties. It took place during two stages: the first and last weeks of the electoral campaign (from 5 to 11 February and from 26 February to 4 March). This preparatory part of our research returned a total *corpus* of analysis of 1,397 posts collected, as said, from the official Facebook pages of the main three political parties.

After having a review of the existing literature on the subject, we have constructed a data matrix to analyze each of the posts in a detailed manner, according to a series of variables considered relevant for the analysis of the communication strategy of the three political parties.

Every data collected were treated from a quantitative point of view to describe the frequency and the intensity of the post publication activities of each considered party and then from a qualitative point of view, in order to investigate the kinds of diffused material, their contents and their related function, witg the im of returning a clear trend of the communication strategies of each political force.

3. The *corpus* of analysis and the intensity of the publication activities

The sample analyzed consists of 1,397 posts published by the *Lega*, the *Movimento 5 Stelle* (M5S) and the *Partito Democratico* (PD): 680 posts during February 5-11 and 717 posts during February 26-March 4).

It is evident (Table 1.) that in the two weeks considered the volume of published posts is almost stable and describes the same intensity of the publication activities. The substantially stable percentages (with a slight increase in activities during the second week) describe that the political parties reseverd the same attention at the opening and closing of the electoral campaign: 48.68% *vs*. 51.32%.

Table 1. Number of posts published by political party per week

Week	%	In absolute terms
February 5-Febaury 11	48.68	680
February 26-March 4	51.32	717
Total	**100**	**1,397**

Source: Our elaboration.

In the two weeks considered, the trend of posts publication remains substantially stable, without significant variations. It cannot be said the same about the communicative intensity of the three political parties. In looking at the posts publication activity of the *Movimento 5 Stelle*, the *Lega* and the *Partito Democratico* (Table 2.), it is evident that the *Lega* represents 73.16% of the total volume of posts of the general sample. The *Lega* published average of about 500 posts a week and about 70 posts a day, marking a substantial difference comparing to the other two political parties. The *Lega* is distinguished by a posts publication activity "almost obsessive", that aims a constant contact with voters throughout the day; it is not the same for the *Movimento 5 Stelle* and for the *Partito Democratico*, that have had much lower percentages. In particular, the *Partito Democratico* represents 7.02% of the entire sample, that corresponds to a poor use of the network as a propaganda tool, slightly oscillating between the first and second week of detection. The *Movimento 5 Stelle*, with a decidedly more contained trend compared to that of the *Lega*, increases the production of electoral propaganda online during the second week, making the most of at the closing phase of the electoral campaign.

Table 2. Number of posts per week (% of the total number of posts)

Political Party	**February 5-11**	**February 26-March 4**	**Total**
Lega	37.44%	35.72%	73.16%
M5S	8.23%	11.60%	19.83%
PD	3.01%	4.01%	7.02%
Total	48.68%	51.32%	**100%**

Source: Our elaboration.

4. Type of posts

To be able to investigate the communication of each party, we have developed an easy tool of analysis aimed at identifying the descriptive categories of the different types of post, in order to better describe the communicative style of each party. We have built a variables based analysis according to the following descriptive methods: Link (posts containing a link that refers to an external page), Photo (photographs and images of electoral propaganda), Video (direct events, electoral spots, excerpts of transmissions television), Status (post of written text only). The most used post type of the total sample of analysis (Table 3.) is the image (45.45%), because, as happens in the offline electoral campaign, is exploited for the immediacy of its communication that can directly catch the attention of the voter. The use of the images, between the first and second week of relevation, is clearly increasing, posts of external links to the social platform are drastically reduced, while a post of only text requires a greater activation effort from the user.

Table 3. Posts publication style

Type of posts	**February 5-11**	**February 26-March 4**	**Total**
Photo	41.32%	49.37%	45.45%
Link	40.15%	23.85%	31.78%
Video	18.24%	25.80%	22.12%
Status	0.29%	0.98%	0.64%
Total	**100.00%**	**100.00%**	**100.00%**

Source: Our elaboration.

The three political parties managed the online electoral campaign according to different communication styles (Table 4.). The *Movimento 5 Stelle*, unlike the *Lega* and the *Partito Democratico* (that prefer the use of images), prefers in its communication the use of video (49.93% posts) to the detriment of the image (25.27%). If the verbal communication

(Status) requires a greater interaction from the user, the *Lega* – which, as we have seen, is the most active political party on the network - does not use this type of post at all, preferring the immediacy of the images, that correspond to over 50% of its sample. Posts of Status, which represent the smallest part of the total sample, with only 9 total posts, have been mainly made by the *Partito Democratico*, which, in considering a lack of capabilities of the use of the network, uses less immediate and captivating languages.

Table 4. Posts publication styles adopted by political parties

Types of post	**Lega**	**M5S**	**PD**
Photo	50.78%	25.27%	46.94%
Link	34.05%	27.44%	20.41%
Video	15.17%	46.93%	24.49%
Status	0	0.36%	8.16%
Total	**100%**	**100%**	**100%**

Source: Our elaboration.

5. Communication strategies

Communication strategy adopted by the political parties during the electoral campaign was investigated by elaborating a variable named "*Post Function*". The variable is articulated on three general macro-categories, describing three main strategies: "*Negative campaign*", which contains adversary's attacks and denigration functions; "*Political proposal*", which illustrates the program points and the actions carried out by the political parties; and "*Engagement"*, that aims to involve the voters proposing to be *militant 2.0*, who act in first person on the digital campaign of the political party.

The variable "*Post Function*" (Table 5.) confirms the hypothesys of the propensity to adopt a *Negative campaign* strategy: in fact, about 48.8%[2] of the published posts aim to persuade the voter by denigrating the political opponent and by using negative feelings. About 11% of posts use clearly *Negative* and *Negative-Comparing* modalities; 9.40% denounce circumstances of political relevance through the use of statistical data; 8.80% generate fear and concern about events reported; while 17.4% recall specific dramatic news events.

[2] Cumulative percentage of Negative, Negative Comparing, Data declaration, Irony/parod/sarcasm, Fear and Current events modes.

Table 5. Post Function

Post Function	**%**
Negative	8.4
Negative Comparing	2.9
Data Declaration	9.4
Irony/Parody/Sarcasm	1.9
Fear	8.8
Current Events	17.4
Past Political Achievements	1
Political Program	12.3
Political Issues	3.1
Identity Membership	1.5
Feeling good	3.6
Media Agenda	5
Territorial Agenda	11.9
Online Mobilization	11.7
Fundraising	1
Total	**100**

Source: Our elaboration.

Considerated the minority of the other two communication strategies, our attention is focused on the analysis of the prevailing negative. The *Lega* is the party most oriented to the Negative campaign. It publishes posts that recall facts of crime or illegality (about 19% of posts) that aim to emotionally shake the voters causing fear and worry (about 11% of posts). Even the *Movimento 5 Stelle* adopts the *Negative strategy*, by directly attacking the opposing political parties in a comparison of the political proposal (about 10% of posts) and using statistical data to highlight what differentiates it from others (about 13% of posts).

The *Partito Democratico*, the outgoing government party, is the only party that does not use negative strategies. It focuses its communication on the policy proposal aimed at explaining and illustrating the actions taken and the program points (about 37% of posts).

6. Conclusive observations

As emerged in the course of the discussion, the online electoral campaign of the three parties has returned a composite and descriptive picture of their communicative peculiarities.

The overall sample of the materials, collected in the two weeks of monitoring, describes a consistent and unequivocal protagonism of the Lega, followed by the *Partito Democratico*, whose online campaign appears truthfully inconsistent, if compared to the other two political forces, and not perfectly in line with languages and tones of immediacy imposed by the online network. If the *Lega* and the *Movimento 5 Stelle* aim to hit the imagination of the electorate with images and videos of immediate use, the *Partito Democratico* produces long text posts that imply voluntary activation by the user. The negative tones, as supposed in the introduction, play a key role. For the *Lega*, news events are the basis for the elaboration of the negative political messages, that are based on verbal and symbolic violence, without any programmatic political proposal. The target of its negative propaganda is the *left*, the European Community and the immigration phenomenon meant as a threat to national security. The negative tones of the *Lega* are full of ideological political references that refer to xenophobia, nationalism and anti-Europeanism. The negative trend of the *Movimento 5 Stelle* is in line with its nature of post-ideological political movement and aims to build its image according to a denouncing and contrasting strategy towards the political adversaries. The *Movimento 5 Stelle* does not express any ideological value and does not take sides on the news events that have strongly influenced the debate in the electoral campaign. If the *Lega* and the *Movimento 5 Stelle* propose an electoral campaign of attack; the *Partito Democratico* plays an electoral campaign in a defense tactics, but, considering the effective electoral results and the poor digital campaign, it did not exert much appeal on voters.

References

Castells, M. (2009). Comunicazione e potere. Milano: Università Bocconi Editore.

Chadwick, A. (2013). The Hybrid Media System: Politics and Power. New York: Oxford University Press.

Diamanti, I., Lazar, M. (2018). Popolocrazia. Bari – Roma: Laterza.

Fukuyama, F. (1992). La fine della storia e l’ultimo uomo. Milano: Rizzoli.

Holtz-Bacha, C. (2016). Europawalkampf 2014. Berlin: Springer Vs.

Manin, B. (2016). Principi del governo rappresentativo. Bologna: Il Mulino.

Norris, P. (2000). A Virtuous Circle. Political Communications in Postindustrial Societies. Cambridge: Cambridge University Press.

Raine, L., Wellman, B. (2012). Networked. Cambridge: Mit Press.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8345

Access and analysis of ISTAC data through the use of R and Shiny

González-Martel, Christian[a] ; Cazorla Artiles, José M.[b]; Pérez-González, Carlos J.[c]
[a]Departamento de Métodos Cuantitativos en Economía y Gestión, Universidad de Las Palmas de Gran Canaria, Spain, [b]Universidad de Las Palmas de Gran Canaria, Spain, [c]Departamento de Matemáticas, Estadística e Investigación Operativa, Universidad de La Laguna, Spain.

Abstract

The increasing availability of open data resources provides opportunities for research and data science. It is necessary to develope tools that take advantage of the full potential of new information resources. In this work we developed the package for R istacr that provides a collection of eurostat functions to be able to consult and discard the data that Eurostat, including functions to retrieve, download and manipulate the data set available through the ISTAC BASE API of the Canary Institute of Statistics (ISTAC). In addition, A Shiny app was designed for a responsive visulization of the data. This develope is part of the growing demand for open data and ecosystems dedicated to reproducible research in computational social science and digital humanities. With this interest, this package has been included within rOpenSpain, a project that aims to promote transparent research methods mainly through the use of free software and open data in Spain.

Keywords: *Economic databases; R; package; Shiny; visualization.*

1. Introduction

The Open Data initiative or data opening is a practice that seeks to ensure that certain data and information belonging to public administrations and organizations are accessible and available to everyone, without technical or legal restrictions.

Ruijer et al. (2017) have studied a context-sensitive open data design that facilitates the transformation of raw data into meaningful information constructed collectively by public administrators and citizens. Thorsby et al. (2017) research on features and content of open data portals in American cities. Their results show that, in general, the portals are in a stage of development and need to improve user help and analysis features as well as inclusion of features to help citizens understand the data, such as more charting and analysis.

The reproducible research defined as the complete analytical workflows, fully replicable and transparent, that span from raw data to final publications can benefit from the availability of algorithmic tools to access and analyse open data collections (Gandrud, 2013; Boettiger et al., 2015). Dimou et al. (2014) presents a use case of publishing research metadata as linked open data and creating interactive visualizations to support users in analyzing data in a research context.

However, the data provided in open access are not in a standardized format and arises the need to adapt the code to specific data sources to accommodate variations in raw data formats, access details so that the end users can avoid repetitive programming tasks and save time allowing simplification, standardization, and automation of analysis workflows facilitating reproducibility, code sharing, and efficient data analytics.

Following this idea, within the ecosystem of R, several packages have been created to work with data from Food and Agricultural Organization (FAO) of the United Nations (FAOSTAT; Kao et al. 2015), World Bank (WDI; Arel-Bundock 2013, wbstats; Jesse Piburn 2018), Open Street Map (osmar; Eugster and Schlesinger 2012) amog others.

The Canary Institute of Statistics (Instituto Canario de Estadística, ISTAC) provides a rich collection of data, including thousands of data sets on Canarian demography, health, employment and tourism and other topics in an open data format.

ISTAC is the central authority of the canary statistical system and the official research office of the Government of the Canary Islands and, among others, among its functions are to provide statistical information and coordinate the public statistical activity of Canary Island autonomous region.

The main access to ISTAC is the web-based graphical user interface (GUI) from where the data can be consulted and downloaded in alternative formats. This access method is fine for the occasional use but is tedious for large selections and when the user must access to data very frequently. The second method uses an Application Programming Interface (API) that can be embedded in a computer code to programmatically extract data from ISTAC. We have developed a R package that integrates the API into the code that allows for the downloaded data to be directly manipulated in R. Based on this package, we have also created a Shiny application that allows a visualization of ISTAC data.

The visualization characteristics is one of the most important features in analyzing information from open data sources. Chen and Jin (2017) have recently proposed a data model and application procedure that can be applied for visualization evaluation and data analysis in human factors and ergonomics. Jones et al. (2016) research innovative data visualization and sharing mechanisms in the study of social science survey data on environmental issues in order to allow the participatory deliberation. Kao et al. (2017) shows how to use a visualization analysis tool for open data with the aim to verify whether there exists sensitive information leakage problem in the target datasets.

This paper provides an overview of the core functionality in the current release version. A comprehensive documentation and source code are available via the package homepage in Github[1]. The package is part of rOpenSpain[2], an initiative whose objective is to create R packages to exploit open data available in Spain for reproducible research.

This paper is structured as follows: firstly, we explain the data extraction procedure implemented in the R library and the workflow to achieve visualization of data. In section 3 we explain the architecture of the visualizations with Shiny. Finally, we present some concluding remarks.

2. The extraction routine in istcar

To install and load the last release version of istacr, the user should type in R the installation from GitHub command from the devtools package.

```
devtools::install_github(“rOpenSpain/istacr”)

library(“istacr”)
```

When the package is loaded the metadata of each dataset available by ISTAC BASE API are also loaded into the `cache` variable. It contains information about the title, topic, subtopic, the url to access to the json data, among other.

For searching about a specific term the `istac_search()` function is provided.

[1] https://github.com/rOpenSpain/istacr

[2] https://ropenspain.es/

```
busqueda.egt <- istac_search("egt", fields = "datos publicadosII")
```

This seeks among all the ISTAC BASE datasets those in which the pattern “egt” appears within the field “datos publicados II”. Other fields can be “titulo” (default), “tema”, “subtemaI”, “subtemaII”, “datos publicados I”, “origen” and ”encuesta”. You can obtain the list of fields with

```
names(cache)
```

The patter can be used with regular expression operators. The output are the rows or row of cache that keep to the pattern. Values in the ID column of the output provide data identifiers for subsequent download commands.

```
busqueda.egt$ID[1]
```

```
## [1] "sec.hos.enc.ser.2528"
```

2.1. Downloading data from ISTAC

We retrieve the data from the dataset with the ID reference using the ISTAC BASE API.

```
df <- istac(busqueda.egt$ID[1])
```

By default the function istacr works with human-readable labels. With the argument `label = FALSE` the function converts the labels into less interpretable codes.

The indicators in the ISTAC open data service are typically available as annual time series grouped by islands, but sometimes at a different granularity or geographic levels.

If the dataset has the “Islas” column it can be filtered by islands using the argument `islas = TRUE`, otherwise this argument is ignored. Valid values for islands are: El Hierro, La Palma, La Gomera, Tenerife, Gran Canaria, Fuerteventura and Lanzarote.

The function allows filtering the dataset by dates using the arguments `startdate`, `enddate`, and `mrv`. The argument `freq` controls the granularity of the data for fetching yearly ("anual"), biannual ("semestral"), quaterly ("trimestral"), monthly("mensual"), bi-weekly("quincenal"), weekly("semanal") values.

2.2. Data visualization

Istacr by itself does not have a dedicated function to plot the data but you can use the potential that R provides to visualize the data retrieved. Figure 1 shows the result of the combination of istacr and ggplot2 package.

```
ggplot (df_p, aes(x = Periodos, y = valor, fill = `Países de
residencia`)) +
      geom_col() +
      facet_wrap(~`Países de residencia`) +
      theme_bw()+
      theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
labs(x = "Years", y = "Total expenditure (euros)", title = "Total
tourist expenditure according to countries of residence")
```

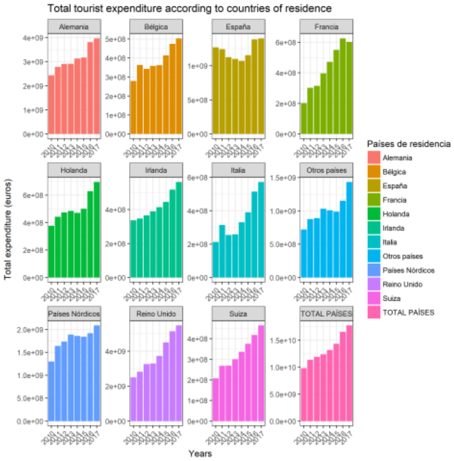


Figure 1. Data obtained through istacr. Visualization with ggplot function from ggplot2 package. Source:ISTAC (2018).

Because that most of ISTAC dataset contains geographical information, map visualization can be represented in a very natural way.

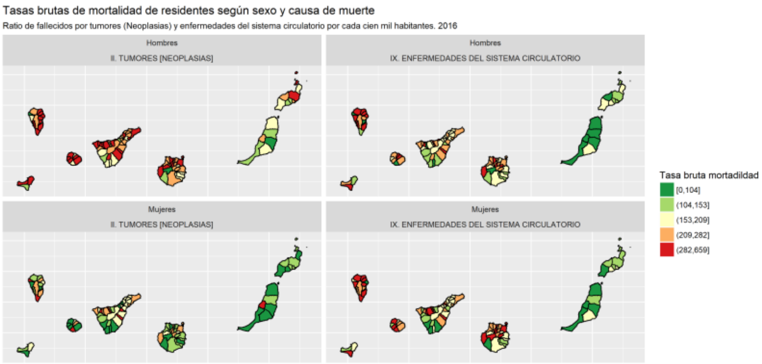

Tasas brutas de mortalidad de residentes según sexo y causa de muerte
Ratio de fallecidos por tumores (Neoplasias) y enfermedades del sistema circulatorio por cada cien mil habitantes. 2016
Hombres
II. TUMORES [NEOPLASIAS]
Hombres
IX. ENFERMEDADES DEL SISTEMA CIRCULATORIO
Mujeres
II. TUMORES [NEOPLASIAS]
Mujeres
IX. ENFERMEDADES DEL SISTEMA CIRCULATORIO
Tasa bruta mortadildad
[0,104]
(104,153]
(153,209]
(209,282]
(282,659]

Figure 2. Geographical visualization with ggplot2 package. Source: Análisis de la Mortalidad/ Series anuales. Municipios de Canarias. 1999-2016. ISTAC.

3. Creating a Shiny Web Application

The last step in this work focuses in generate a web application in R (Shiny). The Shiny is feed up using the ISTAC base API through istacr and its main purposes are the access to ISTAC tourism data and to facilitate the understanding of tourism patterns in the Canary Islands with special attention to exploit the statistics which comes from Tourism Expenditure Survey.

Based on the idea from New Zealand Tourism Dashboard[3], the Canary Islands Tourism Dashboard has been developed[4,5]

The structure of this Shiny shows a navigation menu upper bound. This menu contains several sections. The first one is a brief description of the application. The second one is the tourism expenditure section, this section is also composed subsections related and showed by a dropdown menu. The third sections is referred to the tourist profile socio-demographics characteristics are shown here. The fourth section is about the travel characteristics. Last section is about the authors. In future the purpose it is to improve this options to get a more complex application.

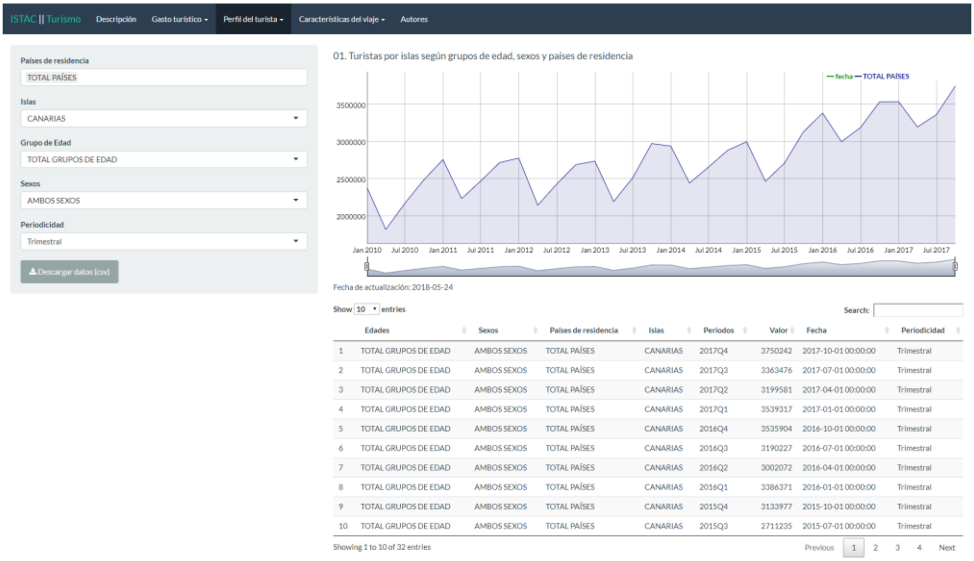
An important reason to use Shiny is the interaction between user and server. In this sense the user can filter data and change the rendered visualization. The Shiny application has an option to download data and export visualization are also available.

[3] https://mbienz.shinyapps.io/tourism_dashboard_prod/

[4] https://jmcartiles.shinyapps.io/canary_islands_tourism_dashboard/

[5] Full code aviable on https://github.com/jmcartiles/canary_islands_tourism_dashboard

A Shiny example of use is shown in Figure 3. The data represented is referred to the number of tourist quarterly to the Canary Islands by age, sex and residence country. In this tab, the user has four filter options in a dropdown menu at left-side. The results are displayed as chart and table. In the data panel a search option to filter data by pattern and a sorting option could be also used.



Países de residencia: TOTAL PAÍSES

Islas: CANARIAS

Grupo de Edad: TOTAL GRUPOS DE EDAD

Sexos: AMBOS SEXOS

Periodicidad: Trimestral

Fecha de actualización: 2018-05-24

Show 10 entries

	Edades	Sexos	Países de residencia	Islas	Periodos	Valor	Fecha	Periodicidad
1	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q4	3750242	2017-10-01 00:00:00	Trimestral
2	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q3	3363476	2017-07-01 00:00:00	Trimestral
3	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q2	3199581	2017-04-01 00:00:00	Trimestral
4	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2017Q1	3539317	2017-01-01 00:00:00	Trimestral
5	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q4	3535904	2016-10-01 00:00:00	Trimestral
6	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q3	3190227	2016-07-01 00:00:00	Trimestral
7	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q2	3002072	2016-04-01 00:00:00	Trimestral
8	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2016Q1	3386371	2016-01-01 00:00:00	Trimestral
9	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2015Q4	3133977	2015-10-01 00:00:00	Trimestral
10	TOTAL GRUPOS DE EDAD	AMBOS SEXOS	TOTAL PAÍSES	CANARIAS	2015Q3	2711235	2015-07-01 00:00:00	Trimestral

Showing 1 to 10 of 32 entries

Figure 3. Shiny application for istacr data visualization. Source:ISTAC (2018).

4. Conclusions

In the last years have emerged a high numer of tools that enable the sharing of public data in open formats across different cloud platforms through API web services. One of the problems that arises is how to collect data of interest from such sources. The istacr library allows the users to query and obtain statistical data series in an efficient and convenient way. The main aspect of this library consists in connecting to API web service of ISTAC to access data and, then, create a dataset into R to work with it. This represents an advantage respect to other procedures that download data in a computer file previously to read from R.

In most of cases, the visualizations are used to demonstrate the provided information in a alternative fashion to the information they present. The visualizations can provide some significant insights of the open data and allow to non-expert users the opportunities discovery in their data analyzes. Therefore, the usefulness of Open Data is revealed to non-expert users. In this use case, it is highlighted how Open Data helps in improving the quality of the data, the diversity of the information and the integration of knowledge.

Considering these visualizations, the potential that offers istacr package could be highly interesting for managing this kind of data.

References

Arel-Bundock, V. (2013) WDI: World Development Indicators (World Bank), URL http://CRAN.R-project.org/package=WDI

Boettiger, C., Chamberlain, S., Hart, E., Ram, K. (2015). Building software, building community: lessons from the rOpenSci project, *Journal of Open Research Software*, 3(1), DOI http://doi.org/10.5334/jors.bu

Chen, X. & Jin, R. (2017) Statistical modeling for visualization evaluation through data fusion, *Applied Ergonomics*, 65, 551-561.

Dimou,A., De Vocht, L., Van Grootel, G., Van Campe, L., Latour, J., Mannens, E., Van de Walle, R. (2014) Visualizing the Information of a Linked Open Data Enabled Research Information System, *Procedia Computer Science*, 33, 245-252

Eugster, M. J. A. and Schlesinger, T. (2010) osmar: OpenStreetMap and R, URL http://osmar.r-forge.r-project.org/RJpreprint.pdf

Gandrud, C. (2013). *Reproducible Research with R and RStudio.* Chapman & Hall/CRC

Jones, A. S., Horsburgh, J. S., Jackson-Smith, D., Ramírez, M., Flint, C. G., Caraballo, J. (2016) A web-based, interactive visualization tool for social environmental survey data, *Environmental Modelling & Software*, 84, 412-426.

Kao,C.-H., Hsieh, C.-H., Chu, Y.-F., Kuang, Y.-T., Yang, C.-K. (2017) Using data visualization technique to detect sensitive information re-identification problem of real open dataset, *Journal of Systems Architecture*, 80, 85-91.

Kao, M.C.J., Gesmann, M., Gheri, F. (2015). FAOSTAT: Download Data from the FAOSTAT Database of the Food and Agricultural Organization (FAO) of the United Nations, URL https://cran.r-project.org/web/packages/FAOSTAT/index.html

Piburn, J. (2018). wbstats: Programmatic Access to the World Bank API, URL https://www.ornl.gov/division/csed/gist

Ruijer, E., Grimmelikhuijsen, S., Meijer, A. (2017) Open data for democracy: Developing a theoretical framework for open data use, *Government Information Quarterly*, 34(1), 45-52.

Thorsby, J., Stowers, G.N.L., Wolslegel, K., Tumbuan, E. (2017) Understanding the content and features of open data portals in American cities, *Government Information Quarterly*, 34(1), 53-61.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8348

Grassroots Market Research on Grass: Predicting Cannabis Brand Performance Using Social Media Scraping

Kregor, Jennifer [a]; Gomez, Bethany [a]; Kelly, J. Steven [b]; Stevenson, Kathleen [b]
[a] Brightfield Group, USA. [b] Department of Marketing, DePaul University, USA

Abstract

Social media listening has become a useful tool to marketers in studying behavior for a wide variety of consumer applications, from political leanings and drug abuse to common product choices. Although most cannabis products are illegal at the U.S. Federal level, it is legal in 30 states for medical use and 8 states and the District of Columbia for recreational use. Despite the legal issues, cannabis is projected to reach over $31 billion in sales world-wide by 2021. The industry is both rapidly evolving and highly fragmented, making it challenging for companies operating in the space to access the insights and the data to help design communications, product development and branding strategies. The research presented here will show that the application of social media listening can be helpful for cannabis brand marketers to gauge size, scope and nuances of these markets and tailored social media mining can accurately predict a brand's future performance. Later research will show that social media scraping will help identify and segment consumers at a fraction the cost of traditional consumer research methods.

Keywords: *Social media listening, brand share, predictive analytics, cannabis industry,*

1. Introduction

Brightfield Group is a market research firm focused on the legal cannabis industry. The company holds a robust ecosystem of data on all aspects of the cannabis industry, including market sizes, brand shares, pricing and analytical reports and customized consumer research on a custom basis. Syndicated data is constructed using a multi-source methodology, including analysis of publicly available sources, expert interviews, data reported by brands and dispensaries and big data scraped from relevant industry sites.

Global marijuana sales are estimated to reach $31.4 billion by 2021 (Zhang, 2017). U.S. Cannabis consumers can acquire it in many formulations, such as edibles, concentrates, tinctures, vapes as well as the standard flower. Product subcategories include infused chocolate, savory snacks, baked goods, drinks, sugar candy, crumble, shatter, vape cartridges, resin and wax. Products have a wide variety of differentiating attributes, based on strain, dosage, cannabinoid profile (levels of THC or CBD) and quality of ingredients. Marketers in the cannabis industry are confronted with decisions for product development, packaging, and branding along with a plethora of environmental issues from federal and state regulations regarding lack of trademark protection (Schuster, 2016), banking, growing, distribution and marketing communications. With more than 1600 brands of infused products on the market in 2017 (Brightfield Group, 2018), savvy marketing strategies are crucial to a brand’s success. With limited access to capital, steep competition and consumer preferences that are constantly in flux, cannabis brands need to access highly cost-effective and agile consumer research methods to drive product development, marketing and advertising strategies.

Making matters more complicated for cannabis companies, strict advertising regulations are in place limiting where and how brands can promote themselves and vary state-by-state. Since many traditional forms of advertising rely on businesses that are licensed at the federal-level (like broadcasters), few businesses have agreed to advertise cannabis-related content (IAB, 2018). This has driven cannabis businesses to focus instead on more grassroots promotions of their brands (Gunelius, 2018). Some of these methods include brand ambassadors, demo days and event sponsorships, but cause the cost of customer acquisition to increase. Social media is increasingly the preferred tool for promoting brands, leaving a tangible digital footprint to be analyzed to gain insights into brand behavior and consumer perceptions (McVey, 2017).

Reporting by the National Survey on Drug Use and Health (NSDUH) showed, through survey self-report, that in 2014, 13% or 35 million Americans over 12 years old had used marijuana in the past year (Azofeifa et al, 2018). But, for business application these reports are limited to demographics of users of cannabis and other drugs. The cannabis products only include blunts, joints and hash and the focus is on drug abuse, not consumer purchasing patterns.

Studying social media usage has given researchers opportunity to explore the relationship of posts to other behavior. McGregor, et al. (2014) employed the monitoring of several general social media platforms (Facebook, Twitter) as well as blog sites. Their goal was to identify themes of conversations by the community of glaucoma patients. In fact, 14 different themes were identified. This kind of research demonstrates that users are openly willing to offer the language they use to discuss their issues and product usage. But there was no clear relationship to specific product usage. A study by Schwartz et al. (2013) demonstrated the correlation between language and personality. A more recent study by Antoniou (2017) demonstrated that social network posts can be related to users' cognitive profiles as measured by the Meyers Briggs, MBTI profile.

Research focusing on product usage, Culotta and Cutler (2016) established that they could monitor Twitter posts to study consumer perceptions of 200 brands along three perceptual attributes. Their social media monitoring showed high correlation with more expensive survey techniques. As for using social media to predict behavior, Lievens and Van Iddekinge (2016) used social media scraping in the staffing area where employer keywords or signals were compared to social media conversations to predict who might be good potential employees.

Other important research also has demonstrated the predictive ability of social media monitoring analytics. St Louis and Zorlu (2012) demonstrated the relationship between Twitter posts and the spread of flu. Sul et al. (2016) found that monitoring emotional sentiment about a company from Twitter conversations demonstrated impact on same-day and longer-term stock prices for those same companies. Yaden et al. (2017) demonstrated correlation between word usage on social media posts and the religion of the media discussant. Social media scraping has been used to study drug abuse such as when Sarker et al. (2016) showed a clear relationship between Twitter posts and drug abuse.

Chen et al. (2015) conducted smoking research in their analysis of users of some blogs about vape devices and Reddit posts to discover experience of users of electronic cigarettes. However, this work was about tobacco use and did not explore the specifics of the product brands. Research in the cannabis field was carried out by Nguyen et al. (2016). They studied Twitter posts as related to marijuana usage. The customer profiling was minimal, correlating to type of phone used, times of day, etc. The research did not focus on or indicate user perceptions of cannabis product types or brands.

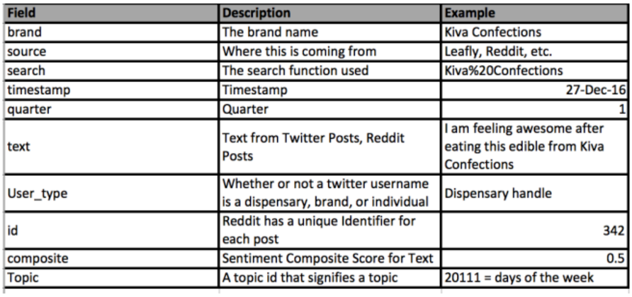
There seems to be no academic research done to relate cannabis users to the brands available in the various marketplaces. What will be presented here is a study of how social media scraping and the results of the analysis therein relates to cannabis product usage by brands and brandshare.

2. Methodology

A total of 3,050,725 words and phrases from 38,014 twitter messages, 2,319 forum messages, and 1,695 professional articles were collected for 86 of the leading cannabis brands spanning a period from January 2016-September 2017. Web crawlers were developed using python, Twitter's API (tweepy), Reddit's API, and Beautiful Soup (Richardson 2007), a search technology system that uses the html structure of websites to more easily extract iterable information. Researchers qualitatively compiled a list of 427 hashtags or search functions that uniquely identified brands (e.g. #kivaconfections for Kiva Confections). The web crawlers used this list to collect messages for each respective hashtag or phrase. Approximately 50,000 posts were scraped that comprise the dataset. Table 1 provides an excerpt of the dataset and structure.

Following the formation of the dataset, the sentiment and topics of these messages were analyzed using python packages which leverage differential language analysis techniques. The sentiment of the Twitter and Forum post language was obtained using VADER (Valence Aware Dictionary and sEntiment Reasoner), a lexicon and rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media. It is fully open-sourced under the MIT License (Hutto & Gilbert, 2014; Rebeiro et al., 2016).

Table 1. Full dataset structure



Field	Description	Example
brand	The brand name	Kiva Confections
source	Where this is coming from	Leafly, Reddit, etc.
search	The search function used	Kiva%20Confections
timestamp	Timestamp	27-Dec-16
quarter	Quarter	1
text	Text from Twitter Posts, Reddit Posts	I am feeling awesome after eating this edible from Kiva Confections
User_type	Whether or not a twitter username is a dispensary, brand, or individual	Dispensary handle
id	Reddit has a unique Identifier for each post	342
composite	Sentiment Composite Score for Text	0.5
Topic	A topic id that signifies a topic	20111 = days of the week

3. Results

Researchers then took the full dataset and aggregated by brand, compiling the total number of twitter posts, professional articles, and forum posts, number of followers, as well as the mean overall sentiment across each of the seven quarters. Social media performance was then aggregated at the monthly basis and compared with monthly brand performance.

Monthly brand shares come from Brightfield Group's proprietary database (Brightfield Group, 2017). Brand shares are calculated based on a combination of sales data provided directly from brands and retailers as well as monthly menu audit scrapes identifying distribution and number of SKUs carried for each brand across each state. Baseline brand share calculation algorithms use distribution of SKUs as a proxy for sales, with algorithms weighted based on sales data provided by dispensaries and brands and validated by qualitative and primary research. An example of the social volume tracking for one brand can be seen in Figure 1.

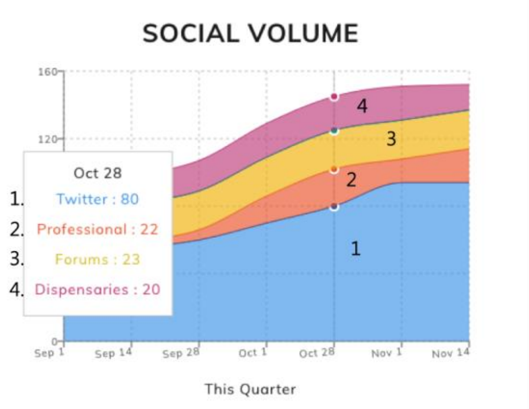


Figure 1. Social volume tracked over time for sample brand

Multiple regression analysis was used to test if the social volume metrics significantly predicted brand SKU's. The results of the regression indicated the six predictors explained 95% of the variance ($R^2 = .95$, $F(120,33) = 4.88$, $p < .000$). Table 2 explains the coefficients and Figure 2 displays the regression analysis.

Table 2. Analysis of coefficients

Variable	B	SEB	t	P-Value
quarter	30.45	55.55	0.55	.04
twitter	4382.39	1112.32	3.94	.00
sentiment	364.88	1331.5	0.27	.08
forums	56.77	19.70	2.88	.00
professional	26.87	11.19	2.40	.02
followers	.0000375	.00000141	2.66	.01

Notes R2 = .9466 (p <.001)

Researchers then used this model to predict 1 quarter in advance after information for the eighth quarter was obtained. The predicted brandshares used by this model accurately explained a significant proportion of the actual brandshares collected in the following quarter, R2 = .95, F(1,152) = 2695.6, p < .000.

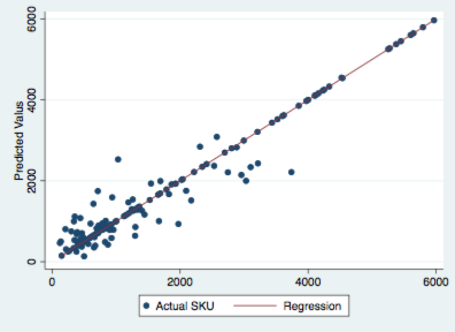


Figure 2. Regression line actual SKU vs. predicted SKU

4. Discussion and Next Steps

The approach presented leverages social listening: tracking both social volume and analysis of linguistic features within Twitter and other online platforms. Social volume metrics predicted outcomes like brandshares and SKUs for 86 cannabis brands. Brands can use social volume metrics to track and predict future brand performance.

The availability and prevalence of social media data now makes it possible to automatically derive characteristics from language use. This research is an example of a highly cost-effective and efficient way to obtain deep insights into an opaque and rapidly changing industry. By mining and analyzing data present on relevant social media channels and key publications, analysts were able to gain insights into the brands, performance and consumers of the legal cannabis industry. This dataset can be used for a variety of applications, including predicting future brand performance, identifying the ROI from social media presence for brands in the space and gaining deeper insights into modern consumer base of this highly sensitive and dynamic industry. An extension of this analysis can be conducted by collecting posts and following trends from individuals that post or follow a particular brand, enabling consumer segmentation into individual personas of each brand (e.g. millennial moms, techie bros) to emerge, which can cut the cost of consumer research down to a fraction of its original cost. Learning more about language patterns and following tendencies may help brands more effectively message to and reach receptive audiences. In a space as grassroots as cannabis advertising, analyses like ours may lead to illuminating insights about a budding industry.

This technique can extend to other industries and consumer research. Companies new to market with low budgets or quickly changing industries can use methods like these to derive automatic cost-effective insights into their consumers. The ability to not only track the messaging around the product, but the aspects of consumers (e.g. collecting posts and liking trends of those messaging or following your brand) allows for thorough capturing of both the active vocal consumers as well as their silent followers.

References

Antoniou, A. (2017). Social network profiling for cultural heritage: combining data from direct and indirect approaches. *Social Network Analysis and Mining*, *7*(1), 39. https://doi.org/10.1007/s13278-017-0458-x

Auerbach, B. (2018, March 15). Looking At The Year Ahead In Cannabis, Technology, Microdosing, Home Brew and Blockchain. Retrieved March 18, 2018, from https://www.forbes.com/sites/bradauerbach/2018/03/15/looking-at-the-year-ahead-in-cannabis-technology-microdosing-home-brew-and-blockchain/

Azofeifa, A., Sherman, L. J., Mattson, M. E., & Pacula, R. L. (2018). Marijuana buyers in the United States, 2010–2014. *Drug & Alcohol Dependence*, *183*, 34-42.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. Journal of machine Learning research, 3(Jan), 993-1022.

Brightfield Group (2017. July). *Understanding Cannabidiol: CBD*. Retrieved from Brightfield Group database.

Chen, A. T., Zhu, S. H., & Conway, M. (2015). What online communities can tell us about electronic cigarettes and hookah use: a study using text mining and visualization techniques. *Journal of medical Internet research*, *17*(9). doi: 10.2196/jmir.4517. http://www.jmir.org/2015/9/e220/

Culotta, A., & Cutler, J. (2016). Mining brand perceptions from Twitter social networks. *Marketing science*, *35*(3), 343-362. https://doi.org/10.1287/mksc.2015.0968

Gunelius, S. (2018, January 24). 5 Proven Marijuana Marketing and Advertising Ideas that Work. Retrieved March 24, 2018, from https://cannabiz.media/5-proven-marijuana-marketing-and-advertising-ideas-that-work/

Interactive Advertising Bureau. (2018). *Marijuana Legalization and Advertising Restrictions in the United States*. Retrieved from https://www.iab.com/marijuana-legalization-advertising-restrictions-united-states/

Lievens, F., & Van Iddekinge, C. H. (2016). Reducing the Noise From Scraping Social Media Content: Some Evidence-Based Recommendations. *Industrial and Organizational Psychology*, *9*(3), 660-666. https://doi.org/10.1017/iop.2016.67

McGregor, F., Somner, J. E., Bourne, R. R., Munn-Giddings, C., Shah, P., & Cross, V. (2014). Social media use by patients with glaucoma: what can we learn?. *Ophthalmic and Physiological Optics*, *34*(1), 46-52. https://doi.org/10.1111/opo.12093

McVey, E. (2017, November 13). Chart: Most effective forms of advertising for cannabis businesses. Retrieved March 24, 2018, from https://mjbizdaily.com/chart-effective-forms-marketingadvertising-marijuana-businesses/

Nguyen, A., Hoang, Q., Nguyen, H., Nguyen, D., & Tran, T. (2017, January). Evaluating marijuana-related tweets on Twitter. In *Computing and Communication Workshop and*

Conference (CCWC), 2017 IEEE 7th Annual (pp. 1-7). IEEE. doi: 10.1109/CCWC.2017.7868364

Richardson, L. (2007). Beautiful soup documentation. Retrieved from https://media.readthedocs.org/pdf/beautiful-soup-4/latest/beautiful-soup-4.pdf

Sarker, A., O’Connor, K., Ginn, R., Scotch, M., Smith, K., Malone, D., & Gonzalez, G. (2016). Social media mining for toxicovigilance: automatic monitoring of prescription medication abuse from Twitter. *Drug safety*, *39*(3), 231-240. https://doi.org/10.1007/s40264-015-0379-4

Schuster, W. M., & Wroldsen, J. (2018). Entrepreneurship and Legal Uncertainty: Unexpected Federal Trademark Registrations for Marijuana Derivatives. *American Business Law Journal*, *55*(1), 117-166. https://doi.org/10.1111/ablj.12118

Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H. (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. *PloS one*, *8*(9), e73791. https://doi.org/10.1371/journal.pone.0073791

St Louis, C., & Zorlu, G. (2012). Can Twitter predict disease outbreaks?. *BMJ: British Medical Journal (Online)*, *344*. doi: 10.1136/bmj.e2353

Sul, H. K., Dennis, A. R., & Yuan, L. I. (2017). Trading on twitter: Using social media sentiment to predict stock returns. *Decision Sciences*, *48*(3), 454-488. https://doi.org/10.1111/deci.12229

Yaden, D. B., Eichstaedt, J. C., Kern, M. L., Smith, L. K., Buffone, A., Stillwell, D. J., ... & Schwartz, H. A. (2017). The Language of Religious Affiliation: Social, Emotional, and Cognitive Differences. *Social Psychological and Personality Science*, 1948550617711228. https://doi.org/10.1177/1948550617711228

Zhang, M. (2017, November 7). The Global Marijuana Market Will Soon Hit $31.4 Billion But Investors Should Be Cautious. Retrieved March 20, 2018, from https://www.forbes.com/sites/monazhang/2017/11/07/global-marijuana-market-31-billion-investors-cautious/#2c79945c7297

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8357

What should a researcher first read? A bi-relational citation networks model for strategical heuristic reading and scientific discovery

Moreno Pascual, Cesar [a]; Martinez de Ibarreta Zorita, Carlos [b]
[a] EAE Business School, Spain, [b] Universidad Pontificia de Comillas, Spain.

Abstract

Scientists usually try to find relevant and updated documents for their research. Also, they face an abundance of information. Most of the methodologies and algorithms look Backwards, so they suffer an inevitable time delay. We propose a recommendation algorithm combining Forward and Backward citation entire networks and Macro, Meso and Micro metrics that concludes in a strategic map and a heuristic reading path. Underlying it, we found an asymmetric bowtie scientific advance model that informs all, solving the abundance problem with a triple reduction and a heuristic reading path.

Keywords: *citation networks, big data, recommendation system, network metrics, science advance model, bibliometrics*

1. Introduction

On the one hand, Scientists look forward to discovering the emergent knowledge or Research Front (Fujita, Kajikawa, Mori, & Sakata, 2014; Huang & Chang, 2014; Price, 1965; Shibata, Kajikawa, Takeda, & Matsushima, 2009; Small, Boyack, & Klavans, 2014). Usually, they include in that Front those documents cited more frequently (Shibata et al., 2009; Upham & Small, 2010). Additionally, the scientists tend to mention the most recent documents to gather the more updated knowledge. This phenomenon is traditionally called "immediacy factor" (Price, 1965). Therefore, relevance can have several conceptualisations that might be contradictory. On the other hand, the number of scientific documents doubles every 1.8 years (Kleinberg, 1999; Wang, Song, & Barabási, 2013). This abundance makes challenging the choice of which papers to read and how to order their reading. Many methodologies, metrics, recommendations systems and information retrieval algorithms were developed to solve the classic problem of relevance and abundance.

This **Paper proposes** an algorithm to guide that reading. It is based on well-proven networks metrics, Micro (Eigenvector Centrality and Betweenness Centrality), Meso (Modularity maximisation (Blondel, Guillame, Lambiotte, & Lefebvre, 2008; Lancichinetti & Fortunato, 2011; Newman, 2006)) and Macro perspectives combination and it applies them to Backward and Forward entire citation networks. Forward citation has been used until now (Belter, 2016; Couteau, 2014) but locally around a document, and to interpret an existing document when a new one cites it, but saying nothing about the citing new one. Namely, Eigenvector Centrality is an energy diffusion vector in a steady state of energy, representing a ranking of documents where the knowledge (the diffused element) of the network is deposited (Rodriguez & Shinavier, 2010). Therefore, these interpretation offers a relevance criteria about new publications without time delay when it is applied to Forward Networks. The former technical novelty, the combination of the metrics to the two networks over a timeline (taking advantage of the acyclical characteristics) and the usage of three levels of analysis permit two new intimately related outputs:

a)A **Strategic Reading map** that classifies the clusters of documents in 4 areas (emerging mainstream, declining mainstream, emerging new stream, and declining new stream) and ranks them. Inside each cluster, we define a document level ranking based on the three Research Fronts defined in the next output informing a document level ranking.

b)A **Science advance model** of the given topic that supports the previous view, defining three Research Fronts (Forward, Intermediary and Backward). The scheme makes possible the comparison of scientific areas.

Some **metrics have been developed to understand "relevance",** aiming to reduce the number of documents considered by the reader. Most bibliometric indicators, understand that the very relevant are the most cited (Garfield, 1972; Moed, 2010) or the more prestigious (Bergstrom, West, & Wiseman, 2008) (defining prestige in different ways) in a given period. For comparison, there are many summaries available (Hu, Rousseau, & Chen, 2011; Salvador-Oliván & Agustín-Lacruz, 2015). Notably, the Inmediacy Index considers the average number of times an article is cited in the year it is published, trying to measure the emergence of a document. It is mainly affected by publication patterns and time delays (Salvador-Oliván & Agustín-Lacruz, 2015). Also, some new indicators based on the citation dynamics are developed to substitute the traditional ones using (Hirsch, 2005; Wang et al., 2013).

There are **other network-based techniques**, such as co-citation (Small, 1973) or bibliographic coupling (Kessler, 1963) or a combination of both (Small et al., 2014), but all of them still need a time elapse. As a consequence, a noticeable time delay appears (Fujita et al., 2014). Additionally, reading is usually done by successively incorporating documents in a recursive search (Vazquez, 2000, 2001). In fact, Scientists include referenced materials that attract their interest that there are not among those initially found, expanding the whole system.

For **example**, if we searched in WOS "*Knowledge diffusion or scientific change*" and "*Knowledge networks*", the system would retrieve 721 and 884 documents, respectively. A Recursive search (Vazquez, 2001) carried out entirely, would involve the revision of 22,986 and 31,925 documents respectively. On the contrary, if this recursive search is not done ultimately, we will most likely be locked in the cluster of the documents in which we started the reading, with little chance of jumping into other groups (Ren *et al.*, 2012).

			Reduction			
Topic	Seed	Expanded network	Clustering	Eigen($\omega_{1,2}^{2forward}$)	Betw($\omega_{1,2}^{2forward}$) (*)	Suggested documents to read
Knowledge diffusion or scientific change	721	22986	11,778	87	20	**107**
Knowledge networks	884	31925	9,184	72	15	**87**

Figure 1 Documents Reduction

This way, when a scientist cites a document, the information is extracted and incorporated into the new one, building a new limit in the research horizon, a new Forward or Destiny Front, that can be quantitatively defined. They also consolidate and reviews the existing literature building an Intermediary Front. Finally, the standard backwards-looking, without controlling the time elapse as some of the mentioned metrics do, makes the Origin or Backward front. All these Fronts can be joint in a whole perspective taking advantage of the citation networks acyclical feature, that is in the basement of the inadequacy of other algorithms. Interestingly, these three Fronts form an asymmetric bowtie connected by the timeline, in analogy with the Internet bow tie (Broder et al., 2000), that conceptualises the scientific advance with a new perspective. The relatively small number of documents in the Forward Front confirms the traditional Price intuition.

2. The model

There are several precedents, coming from search algorithms such Pagerank (Brin & Page, 1998), Hubs and Authorities (Kleinberg, 1999) and other models that combine several network types. Namely, using co-citation and bibliographic coupling (Boyack & Klavans, 2014), direct networks (Caschili, De Montis, Ganciu, Ledda, & Barra, 2014) and local Backward and Forward networks (Belter, 2016; Couteau, 2014). Also, there are some tools like Sci2, Citeseer, Google Scholar or Researchgate that crawls and apply the mentioned or similar bibliometric metrics or algorithms.

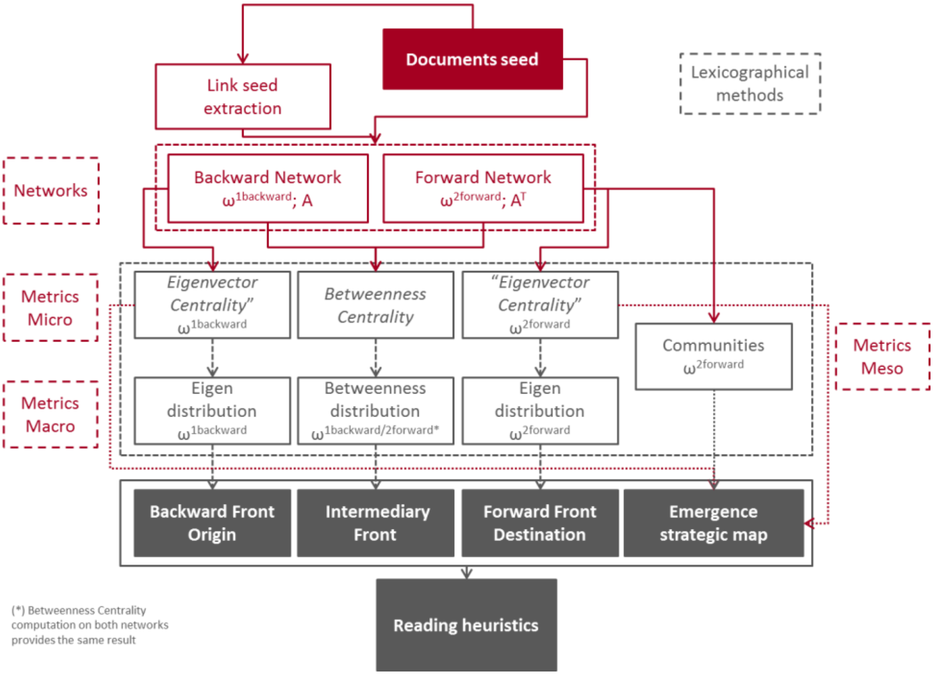

Documents seed
Link seed extraction
Lexicographical methods
Networks
Backward Network $\omega^{1backward}$; A
Forward Network $\omega^{2forward}$; A^T
Metrics Micro
Metrics Macro
Metrics Meso
Eigenvector Centrality" $\omega^{1backward}$
Betweenness Centrality
"Eigenvector Centrality" $\omega^{2forward}$
Communities $\omega^{2forward}$
Eigen distribution $\omega^{1backward}$
Betweenness distribution $\omega^{1backward/2forward*}$
Eigen distribution $\omega^{2forward}$
Backward Front Origin
Intermediary Front
Forward Front Destination
Emergence strategic map
Reading heuristics
(*) Betweenness Centrality computation on both networks provides the same result

Figure 2 Bi-relational algorithm for reading heuristics

The proposed algorithm applies the following steps:

In the beginning, we apply a lexicographic search using, for example, Web of Science (WoS), retrieving the seed of the system.

In a **second step**, we create two citation networks. The Backward network $\omega^{1backward}$ contains directed edges that come from the document that cites the cited document. Conversely, the Forward network $\omega^{2forward}$ uses edges that comes from the cited documents to the one that cites. We understand both networks as a representation of information flow.

In a **third step**, we compute the Eigenvector Centrality to both entire networks and betweenness centrality to any of them, capturing how information circulates through the documents and across its edges in both directions, ranking them. Both, Eigenvector Centrality (Newman, 2012) and Betweenness Centrality (Freeman, 1977; Newman, 2012), are widely used.

Eigenvector Centrality computed in both networks allows us to interpret the relevance of the origin of the information and the destination. The resulting vector is a ranking of documents where the knowledge of the network is deposited, according to the interpretation of the steady state of energy (Rodriguez & Shinavier, 2010). Defining it more clearly, the Eigenvector Centrality measure, applied to the Forward network $\omega^{2forward}$ offers us the documents that, in the researchers' perspective, gather the most relevant information at the current time. The Backward vision has a symmetrical interpretation.

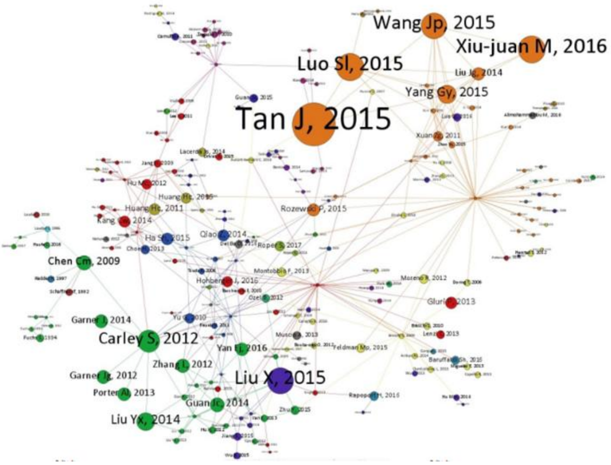


Figure 3 Forward Network example retrieved end 2016. Topic "Knowledge diffusion" or "Scientific change."

However, a complete panorama forces us to analyse what happens between these two states, origin and destination. To do this, we propose the usage of the "Betweenness Centrality" metric. The measure offers a ranking of documents through which the most considerable amount of information passes, mediating the flow of the network. In this case, its calculation also involves the consideration of the entire network structure (Shibata, Kajikawa, & Matsushima, 2007, p. 881). Therefore, all these measures of relevance asses prestige from different and combined perspectives.

The **fourth step** consists of the Community detection. We considered for it **Modularity Maximisation** (Newman, 2006) using the Blondel resolution methodology (Blondel et al., 2008) but applied to the Forward network $\omega^{2forward}$ that detects the clusters formed currently. The clustering step also makes possible to unravel different citation knowledge areas patterns at the same time, as Business and Economics.

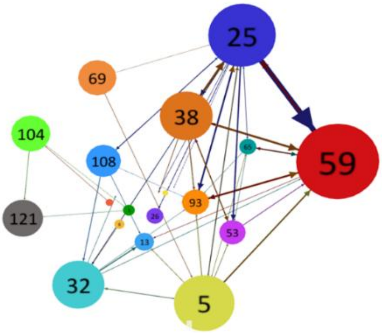


Figure 4 Topic "Knowledge diffusion" or "Scientific change" community detection

The **fifth step** compares the Backward and the Forward Eigenvector relevance in each community to depict the Strategic Map. The vertical axis expresses the emerging factor. For that, we add the Forward Eigenvector Centrality, on the one hand, and in the other, the Backward Eigenvector Centrality score and we measure the distance to the regression line relating the two criteria. The horizontal axis is the Forward Eigenvector Score. Then we can interpret that the more emergency, the higher vertical position, the more current relevance, the more on the right horizontal position, concluding in a four-quadrant interpretation.

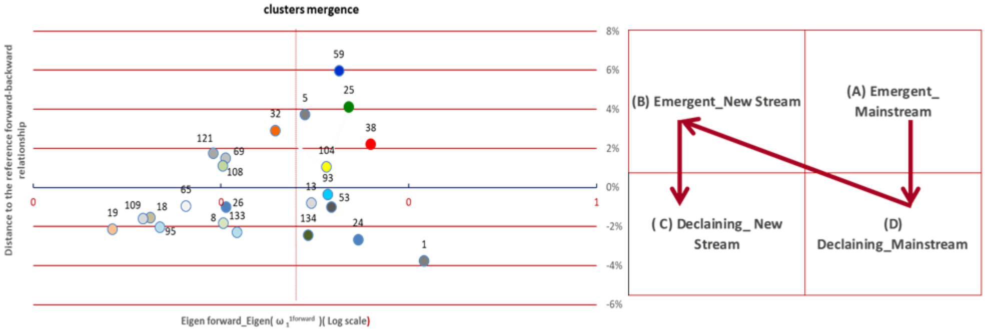


Figure 5 Strategic Map and Reading order example

As a result, in the example depicted in Figure 5, the reading order would be: A) 59,25,5,38,104 B)93,13,53,134,24,1 C) 32,121,69,108. Given that order, inside each

Community, we can logically begin reading the intermediary ranking and then the forward one, concluding with a specific reading suggestion.

Of course, the methodology faces several limitations: “grey literature” (Cooper, Hedges, & Valentine, 2009, pp. 92–93), Ortega hypothesis, the strategic or social citation (Stremersch, Camacho, Vanneste, & Verniers, 2015), obliterated citations (Cole & Cole, 1972; MacRoberts & MacRoberts, 2010) and citation interpretation limits (Amsterdamska & Leydesdorff, 1989; Bellis, 2009), may affect the result. The usage of the whole network as a system, using enough data, might vanish this effects.

3. Reductions and Asymmetric Bowtie

All the mentioned Fronts follows, in all the cases, high skewed distributions that provoke steep decreases in the number of relevant documents. However, the relevant communities detected are only a few attending to the appropriate content and its emergency. Then, a triple reduction its taking place solving the abundance problem.

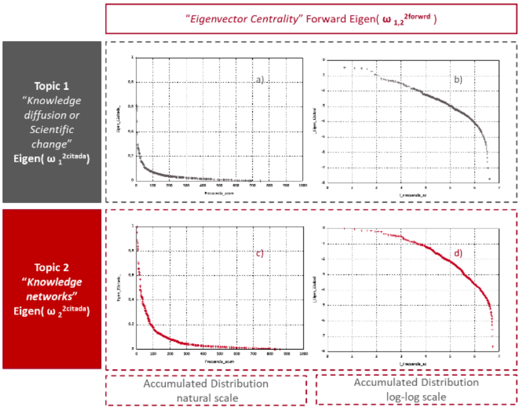


Figure 6 Skewed distribution in several topics. Forward Eigenvector

Finally, a **scientific advance model** is underlying all the reasoning. Interestingly, this model would suggest an asymmetric relation between several fronts, meaning that the Forward is relatively short compared with the Backward front, and the intermediary appears very concentrated, as Price realised many years ago.

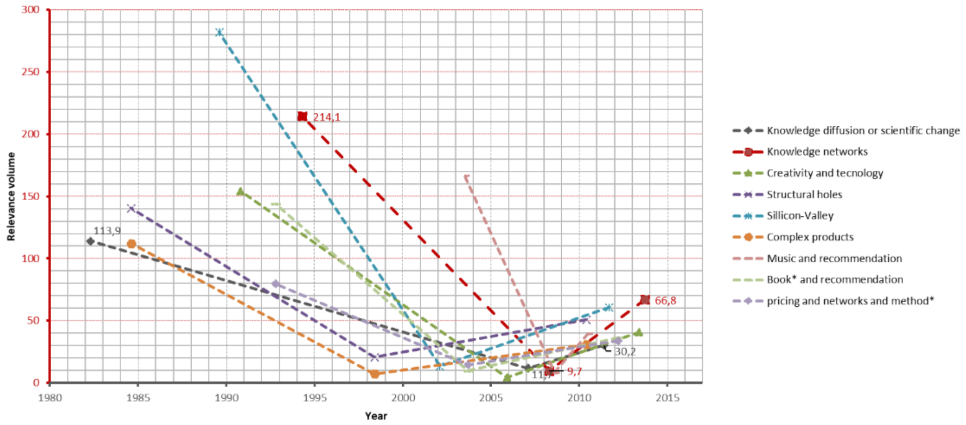


Figure 7 Asymmetric bowtie for several topics

4. Conclusion

A combination of direct Forward and Backward, complete, Networks and Micro, Meso and Macro perspectives not only drives a triple reduction that solves the abundance problem but also gives several "relevance" criteria even for at the very early moments of a document publication. That combination can reconcile relevance and time, understanding the physical interpretation of the Eigenvector Centrality and its application to the forward network as a whole system. The Research fronts conceptualisation, its asymmetric bowtie shape and its relationships, confirm Price intuitions and give light to the reading process. Future research may include weighted edges, an automatic tool and large application for its verification using several databases.

References

Amsterdamska, O., & Leydesdorff, L. (1989). Citations: Indicators of significance? *Scientometrics*, *15*(5–6), 449–471. https://doi.org/10.1007/BF02017065

Bellis, N. De. (2009). *Bibliometrics and citation analysis: From the Science Citation Index to Cybermetrics*. Plymouth United Kingdom: The Scarecrow Press, Inc.

Belter, C. W. (2016). Citation Analysis as a Literature Search Method for Systematic Reviews. *Journal of the Association for Information Science and Technology*, *67*(11), 2766–2777. https://doi.org/10.1002/asi.23605

Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The Eigenfactor metrics. *The Journal of Neuroscience : The Official Journal of the Society for Neuroscience*, *28*(45), 11433–4. https://doi.org/10.1523/JNEUROSCI.0003-08.2008

Blondel, V. D., Guillame, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics*, 10008. https://doi.org/10.1088/1742-5468/2008/10/P10008

Boyack, K. W., & Klavans, R. (2014). Creation of a highly detailed, dynamic, global model and map of science. *Journal of the Association for Information Science and Technology*, *65*(4), 670–685. https://doi.org/10.1002/asi

Brin, S., & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine BT - Computer Networks and ISDN Systems. *Computer Networks and ISDN Systems*, *30*(1–7), 107–117. https://doi.org/10.1016/S0169-7552(98)00110-X

Broder, A., Kumar, R., Maghoul, F., Raghavan, P., Rajagopalan, S., Stata, R., … Wiener, J. (2000). Graph structure in the Web. *Computer Networks*, *33*(1), 309–320. https://doi.org/10.1016/S1389-1286(00)00083-9

Caschili, S., De Montis, A., Ganciu, A., Ledda, A., & Barra, M. (2014). The Strategic Environment Assessment bibliographic network: A quantitative literature review analysis. *Environmental Impact Assessment Review*, *47*, 14–28. https://doi.org/10.1016/j.eiar.2014.03.003

Cobo, M. J. J., López-Herrera, A. G., Herrera-Viedma, E., & Herrera, F. (2012). SciMAT: A new science mapping analysis software tool. *Journal of the American Society for Information Science and Technology*, *63*(8), 1609–1630. https://doi.org/10.1002/asi.22688

Cole, J. R., & Cole, S. (1972). The Ortega Hypothesis: Citation analysis suggests that only a few scientists contribute to scientific progress. *Science (New York, N.Y.)*, *178*(4059), 368–75. https://doi.org/10.1126/science.178.4059.368

Cooper, H., Hedges, L. V., & Valentine, J. C. (2009). *The handbook of Research synthesis and meta-analysis*. (R. S. Foundation, Ed.) (2nd Editio). New York.

Couteau, O. (2014). Forward searching - A complement to keyword- and class-based patentability searches. *World Patent Information*, *37*, 33–38. https://doi.org/10.1016/j.wpi.2014.01.007

Freeman, L. C. (1977). A Set of Measures of Centrality Based on Betweenness. *Sociometry*. https://doi.org/10.2307/3033543

Fujita, K., Kajikawa, Y., Mori, J., & Sakata, I. (2014). Detecting research fronts using different types of weighted citation networks. *Journal of Engineering and Technology Management - JET-M*, *32*, 129–146. https://doi.org/10.1016/j.jengtecman.2013.07.002

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science (New York, N.Y.)*, *178*(178), 471–479. https://doi.org/10.1300/J123v20n02_05

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16569–16572. https://doi.org/10.1073/pnas.0507655102

Hu, X., Rousseau, R., & Chen, J. (2011). Structural indicators in citation networks. *Scientometrics*, *91*(2), 451–460. https://doi.org/10.1007/s11192-011-0587-3

Huang, M. H., & Chang, C. P. (2014). Detecting research fronts in OLED field using bibliographic coupling with sliding window. *Scientometrics*, *98*(3), 1721–1744. https://doi.org/10.1007/s11192-013-1126-1

Indiana University and SciTech Strategies. (2009). Science of Science (Sci2) Tool. Retrieved from https://sci2.cns.iu.edu

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, *14*(1), 10–25. https://doi.org/10.1002/asi.5090140103

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, *46*(May 1997), 668–677. https://doi.org/10.1.1.120.3875

Lancichinetti, A., & Fortunato, S. (2011). Limits of modularity maximization in community detection. *Physical Review E - Statistical, Nonlinear, and Soft Matter Physics*, *84*(6), 1–8. https://doi.org/10.1103/PhysRevE.84.066122

Lawrence, S., & Bollacker, K. (2018). CiteSeerX. Retrieved May 28, 2018, from http://citeseerx.ist.psu.edu/index

MacRoberts, M. H., & MacRoberts, B. R. (2010). Problems of citation analysis: A study of uncited and seldom-cited influences. *Journal of the American Society for Information Science and Technology*, *61*(1), 1–12. https://doi.org/10.1002/asi.21228

Moed, H. F. (2010). Measuring contextual citation impact of scientific journals. *Journal of Informetrics*, *4*(3), 265–277. https://doi.org/10.1016/j.joi.2010.01.002

Newman, M. E. J. (2006). Modularity and community structure in networks. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(23), 8577–8582. https://doi.org/10.1073/pnas.0601602103

Newman, M. E. J. (2012). *Networks: An introduction*. New York: Oxford University Press.

Price, D. S. (1965). Networks of Scientific Papers. *SCIENCE*, *149*. Retrieved from http://apps.webofknowledge.com/full_record.do?product=UA&search_mode=GeneralSearch&qid=17&SID=Z15uohRqct1ddYYr4hv&page=1&doc=4

Rodriguez, M. A., & Shinavier, J. (2010). Exposing multi-relational networks to single-relational network analysis algorithms. *Journal of Informetrics*, *4*(1), 29–41. https://doi.org/10.1016/j.joi.2009.06.004

Salvador-Oliván, J. A., & Agustín-Lacruz, C. (2015). Correlación entre indicadores bibliométricos en revistas de Web of Science y Scopus. *Revista General de Información y Documentación*, *25*(2). https://doi.org/10.5209/rev_RGID.2015.v25.n2.51241

Shibata, N., Kajikawa, Y., & Matsushima, K. (2007). Topological analysis of citation networks to discover the future core articles. *Journal of the American Society for Information Science and Technology*, *58*(6), 872–882. https://doi.org/10.1002/asi.20529

Shibata, N., Kajikawa, Y., Takeda, Y., & Matsushima, K. (2009). Comparative Study on Methods of Detecting Research Fronts Using Different Types of Citation. *Journal of the American Society for Information Science and Technology*, *60*(1971), 571–580. https://doi.org/10.1002/asi

Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for Information Science*, *24*(4), 265–269. https://doi.org/10.1002/asi.4630240406

Small, H., Boyack, K. W., & Klavans, R. (2014). Identifying emerging topics in science and technology. *Research Policy*, *43*(8), 1450–1467. https://doi.org/10.1016/j.respol.2014.02.005

Stremersch, S., Camacho, N., Vanneste, S., & Verniers, I. (2015). Unraveling scientific impact: Citation types in marketing journals. *International Journal of Research in Marketing*, *32*(1), 64–77. https://doi.org/10.1016/j.ijresmar.2014.09.004

Upham, S. P., & Small, H. (2010). Emerging research fronts in science and technology: Patterns of new knowledge development. *Scientometrics*, *83*(1), 15–38. https://doi.org/10.1007/s11192-009-0051-9

Van Eck, N. J., & Waltman, L. (2014). CitNetExplorer: A new software tool for analyzing and visualizing citation networks. *Journal of Informetrics*, *8*(4), 802–823. https://doi.org/10.1016/j.joi.2014.07.006

Vazquez, A. (2001). Statistics of citation networks. *Science*, 12. Retrieved from http://arxiv.org/abs/cond-mat/0105031

Wang, D., Song, C., & Barabási, A.-L. (2013). Quantifying Long-Term Scientific Impact. *Science*, *342*(6154), 127–133. https://doi.org/10.1126/science.1237825

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8361

Historical query data as business intelligence tool on an internationalization context[1]

Carro-Rodríguez, J.M; Lorenzo-Romero, C.; Gómez-Borja, M.A.
Business Administration, University of Castilla-La Mancha, Spain

Abstract

This article reports theory concerning the key strategies of information search behaviour on an international market-orientation context, proposing a common framework to identify search goals using data generated from the keyword planner from google covering the period 2014/17 applied to the clothing retail sector.

A conceptual framework of user goal identification from search data within clothing retail sector is presented and discussed in light of existing empirical studies. For that, a case study of an important fashion retailer has been analyzed: Zara. This firm is situated in the second position within the ranking of most valuable brands of clothes 2015 around the world (Kantar, 2016). A comparative analysis of search patters of this company, between United Kingdom and Spain, has been developed in order to offer the possible internationalization strategies in the online retail sector.

User goals are identifyed and are stable over the period of study, a framework that covers main clothing consumer search goals have been identifyed.

Keywords: *Search behaviour, internationalization strategy, Zara as study case, search patterns by query data framework with KeywordPlanner, comparative analysis with Google Trends.*

[1] This study has been developed within the Research Project funded by Fundación Ramón Areces, entitled “La colaboración abierta en Internet como estrategia de innovación e internacionalización del sector de moda y complementos [open collaboration in the Internet as innovation and internationalization strategies by fashion sector] (2015-2018)”.

1. Introduction

Historical search data is a great source of business intelligence about consumer search in order to describe the number of searches related to a specific query or groups of queries.Therefore search data offers a volume of queries that might convey different types of goals. In order to understand the underling goals from keyword volume, we have to associate user goals with queries, in other words, the why of search behaviour is actually essential to offer or satisfy the user information need.

Our premise is that search data reflects a diverse set of underlining goals and the knowledge of those goals offers the prospect of future improvements and the degree of interest for product purchasing or information for decision-making. we will analyze the web search patters of Zara, as specific study case of the second most important fashion firm in the world. Moreover, we will compare the obtained query data for two different countries, United Kingdom and Spain, in order to offer different international strategies to fashion retailers respect to positioning and web contents.

2. Literature review

Through literature information, we can find a long history of educational behaviour which allow us to understand the consumer behaviour in an online context, based on how people make their searches and the whole search process. Future use of Big Data with the help of Artificial Intelligence is very bright. The use of artificial intelligence will lead to production of machines and computers, which are much more advanced than what we have today. Researchers are continuously working to handle growth of data as well as to convert it into valuable assets. There has been an increasing interest in the application of Aritificial Intelligence tools to IR in the last few years. Concretely, the machine learning paradigm whose aim is the design of system able to automatically acquire knowledge by them. In the light of requirement of intelligent processing of the big data so as to retrieve the information as per the business requirement, authors have proposed a novel architecture.

Lau & Horvitz (1999) analysed a set of server logs and develop a user behaviour model from log data. They were motivated by the idea that informational retrieval could lead to develop of models to diagnose a user’s informational goal, in their model defined a refinement strategy of query sequences’, whereas the refinement class of query represents user’s intent to his prior query. Concluding that there is a difference between query and informational goals. Broder (2002) was motivated by the idea that the intent behind a web search is not often informational. In his work, he recognises that the informational need is associated with some task, and this “need” generates a human-computer interaction where cognitive aspects play a significant role as the need is verbalized mentally and translated into a query posed to a search engine, in his research he has come up with the trichotomy of web search "types" navigational, informational and transactional. Rose & Levinson (2004) developed a framework of user goals with a sample of queries from Altavista logs, they reached the conclusion that the goals naturally fell into a Broder´s (2002) trichonomy at the top level and the full framework was a hierarchy structure, the researches believed that in many cases, user goals could be deduced: By looking at the user behaviour, the query itself, the type of results returned in the search list, results click by user further research or other action by the user. Webb (2009) studied the historical weekly search data from the query "Foreclosure" from google trends in the "US" and correlate with statistics of US home foreclosures, finding that search data from the related domain could predict the foreclosures. Search data has been used to predict financial behaviour. Preis, Moat, & Stanley (2013) quantified trading behaviour using data from Google Trends by analysing changes in query volumes for search terms related to finance. They analysed the performance of 98 terms including stock markets with some terms suggested by the Google set service tool that identifies semantically related keywords. Researchers concluded that conveying large behavioural financial trading data sets with data on search query volumes related to the financial might offer insights into different stages of the decision-making. Despite the increasing popularity of predicting consumer behaviour with historical search data with Google Trends, little research has examined the potential of conveyning query formation and user search goals conditions required for successful in the international market-orientation decision-making contex. Not a lot of research has been carried out or examined with precision how search analysis data using a search goal framework can obtain user behaviour to be used for business intelligence insight.

Currently, Sharma & Srivastava (2017) proposed architecture may help in faster information retrieval with better accuracy and recall.

3. Clasifiying user goals for the clothing sector

This study proposes a search pattern analysis to identify the search goals from the clothing sector. We analysed a specific study case characterized for being an e-commerce website from major-clothing retailer: Zara. Racked Fashion (2016) report includes Zara within the ranking of "retailers which offer intuitive site navigation, a smooth checkout, easy shipping and returns, and awesome products (clothes, shoes, bags, and accessories)". In fact, Inditex has hegemony in the textile sector (being Zara the 50% total sales of group). The Zara's customer invests in clothing average of 266 euros per year (Fintonic, 2016). their search processes. We collected 6 sets of queries per each year (2014/2017) , 3 per country that contained the brand name itself and the brand name plus the main category name (Spain: Zara tienda, Zara hombre; UK: Zara shop, Zara man, Zara woman). The purpose of this examination was to get a wide sample of queries related to the brand name and main product category that will allow us to identify user goals over time, the full download covered 346 groups that contained 4835 suggested queries for Spain and 262 groups and 4,665 queries for the UK. Not related terms to the country in study emerge in the sample, such as “Tienda Zara DF”. They were all removed.

3.1. Identifying user goals from search data

We consider that user goals can be inferred from analysing the user interaction with the search engine, these interactions might be: The query itself, results retrieved by the search engine, type of results clicked on by the user and query reformulation. However, this information is not available or to condense to analyse or at least the last two options.

Our first task was to classify the type of search goals that users might have in mind when interacting with a search engine. Those interactions are the result of information needs from the users during. The queries downloaded where organized in columns,”adgroup” had the key tag word for each group of keywords, Then we manually classified the key tag word from the column adgroup into our framework, in order to identify the underlying goal, first, we identify it by the query itself and second by typing the keyword in the browser and check the list results it they matched the query criteria.We proceed in the following way: A tag query or adgroup like “Zara Locations” has several queries in the group: “Zara store locator”, “Zara store locations”, “Zara men store locator”, and “Zara man store locator”. Due to the relationship among the tag query and the queries within its group, the suggesting goal for all queries was local. The goal was collaborated after submitting the query in the search engine and analysing that the main result link was a store locator and the google maps feature displayed.

3.2. Taxonomy of queries

In our study, we used the following taxonomy of query goals largely based on Rose & Levinson (2004).

We defined **navigational** goal as a demonstrating desire by the user to be taken to the home page. We differentiate queries consisting on the clothing brand named as “Brand” and “Transactional” queries which are formed by the website name or its reformulation. For it to be considered navigational, the query must have a single authoritarian web site that the user has in mind. Not all queries containing the brand name are considered as navigational, user might type an authoritarian web site name to narrow the results within a specific web page. This is because of the relevance relationship between terms is directed, term “A” may strongly suggest term “B” but not vice-versa according to Joshi & Motwani (2006).

Informational category covers goals for answering questions and learning more about the topic.Queries seeking for reviews, sales instructions, customer service and suggestions, we have named them **Advice,** this group can be broadly defined as – I need to know what to do. Product gender could lead as to estimate the interest on female or male products, we split queries related to gender to quantify the volume and interest on those products, categories like woman containing female products and its first link in result is the authoritarian Zara female product link "Zara woman" and the same classification method was applied to add "Man" category. The informational goal **Locale** requests information about stores locations, but we also include contact information and opening hours (e.g. "Zara phone numbers", "Zara opening hours store"). **List** can be define as exploratory search queries, the user is not even considering a purchase, queries are less focused, the main goal is to get a suitable list of results suggested for more in-depth information (e.g. "Zara trouser"). New categories emerged, we categorized under **Deals** when users search for information about products or services that have a specific and temporary status where price is a main constrain, queries such as "clothing sale" or "black Friday Zara deals" are included in this category – I am searching on the web to find only bargains. Another contribution to this taxonomy is the category **Trends** - I am searching on the web for X because I want to know what is currently trendy or new-, queries consisting on topics about "new arrivals", "new season products", "spring summer 2016", "shoes trends", "Zara fashion" will be included into this category.

Another discovery was the category **Target Group** that defines searches that have a define target consumer group and they are not releated to gender (e.g. "Zara kids, "Zara pregnancy clothes"). Many keywords have objetives out of product or future purchase, those are categorized under **Enterprise,** we define this objective as searching information about the company itself and not its prodcuts or services (e.g. "Inditex total turover", "Zara number of shops"). **Clothing attribute** goal are described as searches related to product or service with a high level of search domain by the user, (e.g. "Zara military jeans", "lether jackets").

Resource queries achieve something apart from information. If the resource is, something that we install via App, computer or online the goal is **Download**. When the resource is something that I need to use in the real world, such as logos, turnover statistics, corporate brand information and so forth, we call it **Obtain**. If my goal is to enjoy the resource for leisure and for sharing and social profiles, the goal is **Entertainment and social**; the most common examples are social media profiles, blogs, online catalogues and video. Lastly, the **Interact** goal arise when further interaction will happen after landing the desire page, (such as fashion directories and job vacancies opportunities. we don't include map service) to accomplish user search.

4. Results

We observe that for the set of data under study the searches related to navigational goals in the Spanish market are 7% of the total of searches for the 4 years of study, being 5% for transactional, related to online purchases and 2% as brand awareness.

On the other hand, users in the UK generate a total volume of 13% for the same goal category 8% only for transactional searches and 5% for brand awareness, in both cases the search volume is higher than the Spanish market, we discovered that although the volume is higher for the objective in general only differs by 1% in those that refers to transactional objectives which are related to ecommerce mainly.

Regarding the searches generated for the information goal which receive the most attention for consumers in both countries, the values are very similar, being 37% for Spain and 35% for the UK, the new objectives identified in this research, emerge from this category.

Spainish users have more interest in trends issues by marking a volume of the total of 11% whereas the UK only reaches 5%, in the search goal deals the British market reaches 4% compared to the 3% in Spain.

Users in both countries have the same behavior when performing searches related to the goal named“target” with 2% of the total searches, Attribute and Enterprise goal categories have emerged however they are not consolidated since the volume that show is below 1% for both countries so we will have to take it into account for future research, on the contrary we see that the main objective of users is defined by the search for information since the goal “List” reaches the highest percentage in “informational” goal with a 12% for Spain and a 17% for the United Kingdom, this goal is closely linked to the locale with 4% and 3% respectively. As user may perform a searcg before planning to go to the shop.

Consumers in both markets search for “advise” with a volume of 5% for Spain and 4% for the United Kingdom, Regarding the objectives of resources is very different where UK users do not generate searches related tom this goal , according to data, only Spanish users collect 7%, mainly for “obtain” objectives with 3% and “entertainment” with 4%.

As an overall view Spain reaches 51% of the total volume and the united kingdom 49%, so the data are comparable and can be compared for strategic decision making.

5. Conclusion and implications

In this paper, we have studied the potential of historical search data to identify underlying goals of search. We have collect datasets from a given query using the KeywordPlanner tool and develop a framework based on previous studies. To understand user goals in the fashion sector we established a search goal framework where all queries naturally fell into, analyzing the search patterns for Zara brand in United Kingdom versus Spain.

We observe that queries grouped themselves by common goals from the datasets, we classify query groups into categories for decision making, transactional category was added into the navigational goal, as the proposed of those queries is to reach a site for interact, but we could no conclude whether those transactional queries finally end up into a sale.

From the extracted data we have monitored that user goals identified under this new framework are stable in time as the sample covered 2014 to 2017 therefore we concluide that the current user goals are the most popular for the clothing industry consumer.

As main limitations we have found that datasets were generated by product related queries and brand name, therefore results can be too focused on search goals categories. Moreover, goal classification was carried out by the query itself first and the links generated in the results page, the result list would be organised by the algorithms and might change through time, it can only be taken as a current picture of the current situation.

We suggest for further investigations to extend the framework to analyse user goals not only with historical search data also with internal data from google analytics to categorize inbound traffic queries and search terms into the framework. Another line of interest for investigations should be the data analysis to predict or forecast events in the real world using search data with the search goals framework to compare how the goals correlate with data, and extended to more countries and brands.

References

Broder, A. (2002). A taxonomy of web search. In ACM Sigir forum, Vol. 36, No. 2

Fintonic (2016). El consumo de moda en España [Fashion consumption in Spain]. Available at http://www.elmundo.es/economia/2016/01/12/5694d7feca474159218b45c5.html

Joshi, A. & Motwani, R. (2006). Keyword generation for search engine advertising. Proceedings of Sixth IEEE-ICDM, 123-129. 2. Dongqing Zhu, Ben Carterette, 2010.

Kantar (2016). Las 10 marcas de ropa más valiosas en 2015 [The 10 clothing brands most value in 2015]. Available at http://es.kantar.com/empresas/marcas/2015/mayo-2015-ranking-brandz-de-las-marcas-de-ropa-m%C3%A1s-valiosas-del-mundo-en-2015/

Lau, T., & Horvitz, E. (1999). Patterns of search: analyzing and modeling web query refinement, pp. 119-128. Springer Vienna.

Moat, H. S., Curme, C., Avakian, A., Kenett, D. Y., Stanley, H. E., & Preis, T. (2013). Quantifying Wikipedia usage patterns before stock market moves.Scientific reports, 3.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. Scientific reports, 3.

Racked Fashion (2016). The 38 Essential Online Shops. Available at http://www.racked.com/2015/7/14/8923189/best- online- shopping- stores

Rose, D. E., & Levinson, D. (2004). Understanding user goals in web search. In Proceedings of the 13th international conference on World Wide Web, ACM, 13-19.

Webb, G. K. (2009). Internet search statistics as a source of business intelligence: Searches on foreclosure as an estimate of actual home foreclosures. Issues in Information Systems, 82.

Sharma, L. & Srivastava, V. (2017). Performance Enhancement of Information Retrieval via Artificial Intelligence, International Journal of Scientific Research in Science, Engineering and Technology, 3(1), 187-192.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8364

Big data and official data: a cointegration analysis based on Google Trends and economic indicators

Crosato, Lisa; Mariani, Paolo; Marletta, Andrea and Zavanella, Biancamaria
Department of Economics, Management and Statistics, University of Milano-Bicocca, Italy

Abstract

In this paper the relationship between the Industrial Production Index (IPI), the confidence index for the manufacturing sector and Google searches for several words linked to the economic situation is explored. In particular, time series referred to the period January 2004 - September 2016 on Italian data. An analysis of significant correlations between the selected indicators is achieved to explore the probable comovements of same. Adding one observation at a time since the first forewarning signs of the 2008 crisis, we find that a few Google searches and the IPI cointegrate, particularly during the strong downward trend leading to January 2009, while there is no cointegration between confidence indicators and the IPI. These results suggest that searches in google and the IPI or the confidence indexes are influenced by common circumstances. Finally forecasts of the IPI obtained through VECM models suggest that the evolution of the IPI can be well represented using the real time Gtrends selected variables.

Keywords: *Big Data; Google trends; Confidence indicators, Cointegration*

1. Introduction

The Industrial Production Index (henceworth, IPI) is one of the main monthly indicator attesting the current health of a country's economy. Accordingly, several contributions in the literature proposed to forecast it usually imputing hard data as regressors, from macroeconomic variables to business-specific indicators (Bodo and Signorini, 1987; Bruno and Lupi, 2004; Hassani et al., 2013). Soft data, such as text analysis in media and other sentiment indicators were introduced instead by Ulbricht et al. (2016) to predict the German IPI with more than 17,000 models.

The degree of novelty of the paper consists in the combined use of data coming from hard and soft data. The basic idea is to analyse the comovements of time series related to the general and personal economic situation achieved by different sources. The main goal of the paper is to understand whether web based soft index numbers together with confidence indicators may help in predicting the hard IPI. In their work, Ulbricht et al. (2016) showed that when it comes to the forecast of industrial production models using media data clearly outperform models without media data. Here the aim is to understand whether even models considering Gtrends data performed better than those taking into account only hard data.

The empirical strategy of the paper is to proceed by subsequent selection of variables. Firstly, the selection is obtained by simple visual inspection on the range of variability; secondly it is realized a correlation analysis with the IPI. If the correlations between the IPI and the soft indicators is significant, it is possible to represent this relationship through timeseries modeling. Last selection step of indicators is to exclude the stationary ones in order to proceed to the final cointegration analysis. Finally the presence of more than one cointegration relationship among time serie is tested, to end up with VECM based short term forecasts of the IPI.

The paper is organized as follows. After a brief introduction, data are shown in Section 2, in Section 3 methodology about time series and the choice of selected indicators is presented; finally main results and conclusions are discussed in the last part of the paper.

2. Data

This paper makes use of three data sources, two of which official and a third one non-official. The first is the Industrial Production Index, monthly released by ISTAT (Italian Institute of Statistics) with two months of delay with the reference period. The IPI is a 2010 fixed base Laspeyres index and is the main conjunctural indicator measuring real output for all facilities located in Italy.

The second data source is the Italian confidence index for manufacturing, monthly released by ISTAT with about 15 days of delay with respect to the interviews. In particular, here data refer to opinions on current level of orders, current economic situation, future level of orders and future economic situation.

The third data source we use is Google Trends, a free tool by Google showing the interest of some keyword during time. It allows to monitoring tendencies about a topic detecting the search frequencies on the web. Typing the keyword (or the topic), it is possible to extract frequencies and trends.

The economic literature has been using Google trends since its appearance in 2004 (see Hassani and Silva (2015) for a recent review on forecasting using Big Data). Google trends data are released as monthly frequencies of searches starting from January 2004, therefore this is the initial date for all our time series. Since the interest of this paper regards in understanding whether Google searches can be considered and used as proxies of the IPI, the searched words in Google Trends are related to the economic situation.

The words we have searched for in Google Trends are economic crisis, recovery, GDP, gross domestic product, public debt, spread, recession, unemployment, employment, job. We also construct naive composite Gtrends indicators by summing up frequencies associated to related words so obtaining four more variables: Total cycle = economic crisis + recession + recovery, Total occupation = unemployment + employment + job, Total Debt = public debt + spread plus a mixed-up variable three words = economic crisis + unemployment + public debt.

The official statistics we use in the paper are all expressed as index numbers in base 2010, so in order to have a fair comparison, also the Gtrendsdata have been indexed to 2010. To this end, the single and composite words monthly frequencies were divided by the mean of 2010 respective frequencies.

3. Time series methodology and choice of the selected indicators

The time series from the three data sources here shown differ at least in two aspects. First, the IPI and the confidence indicators (when needed) are published already deseasonalized, while the Gtrends variables must be treated for seasonality. Therefore, the R-interface to X13ARIMA-SEATS method by the United States Census Bureau is applied. Second , they are released with different lags with respect to the date of the information they are referred to. Morevoer the IPI of two months earlier is available at the end of each month, while confidence indicators and Gtrends variables refer to the current month. Accordingly, the data matrix is shaped anticipating all confidence and Gtrends indicators by two months. All the time series thus obtained are represented in figure 1. A quick glance to the series reveals different degrees of variability among the time series, highlighting the structural difference among the indicators.

The flatter series is for sure the IPI, followed by the confidence indicators and the Gtrends variables. Gtrends variables are clearly more volatile and subject to sudden jumps in correspondance of particular events (for instance, see in figure 1 the spikes in economic crisis from spring 2008 onwards and of three words at the end of the Berlusconi Government in summer-fall 2011).

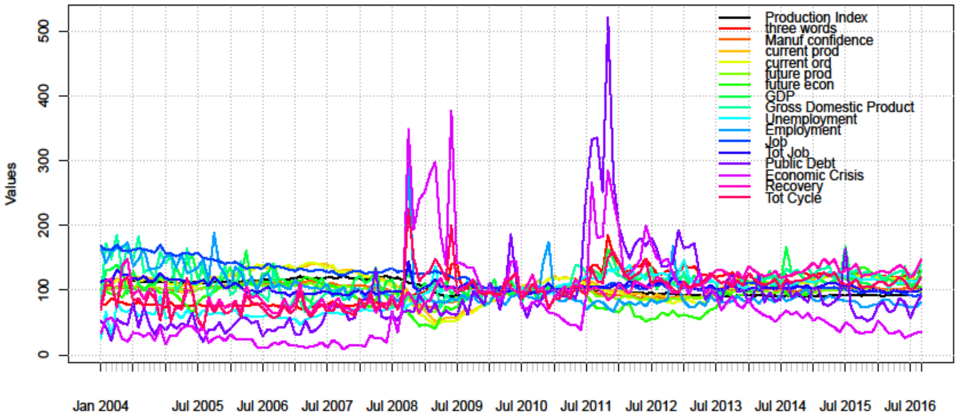


Figure 1. Time series of the selected indicators. Sources: ISTAT official statistics and our own elaborations on google trends data.

If the official sources (IPI and confidence index for manufacturing) and the non-official one (Google Trends) present common movements, it is more simpler to analyse the evolution of the phenomena involved. This could help in terms of prediction with the aim to move up the description of the economic conjuncture in comparison with official data. The cointegration analysis get back the idea according to two or more economic variables, although characterised by a different behaviour in the brief period, the could have some co-movements and tendency in the longrun period.

The final aim of the paper is to explore whether Gtrends variables and confidence indicators may show some predictive power on IPI. For this purpose it is used a multivariate time series model (VAR or VECM if any cointegration relationship appears). In particular, a forward approach is adopted adding one observation at a time from April 2008 onwards to monitor changes in the cointegration relationship during the observed period.

The selection process of the initial variables could be divided into three steps:

- removal of indicators showing too wide a range of variation (spread, recession and total debt);
- restriction to variables which correlate with the IPI and elimination of all variables showing no significant correlation relationship with the IPI and, among the remaining, those presenting a correlation coefficient lower than 0.3 (GDP, Gross domestic product, total job, public debt);
- test for the presence of Unit root in the series, as a preliminary information for the cointegration analysis. using a Phillips Perron test for unit root on the whole set of 100X12 series.

According to the selection process here described, the only variables not discarded are the threewords index (economic crisis + unemployment + public debt), job, economic crisis and all the confidence indicators.

4. Results

After the selection process of indicators correlated to the IPI, the next step consists in the cointegration analysis of the IPI index with one of the remaining variables in turn. Results of the Engle and Granger test for cointegration, reported in figure 2, point to no cointegration neither between the IPI and the confidence indexes, nor between the IPI and job. On the contrary, the IPI and three words do cointegrate and so do IPI and economic crisis, although there are some spikes when the turbolence in the two Gtrends variables is higher. The cointegration analysis between confidence indicators and threewords reveals a similar outcome.

This could be viewed as a first result of the paper contributing to define a selection strategy for Gtrends variables to increase forecasting models, although at present restricted to this particular case. If variables cointegrate when influenced by a common factor or by a combination of common factors, it might tentatively say that a few of the Gtrends variables and the IPI share some pattern drivers.

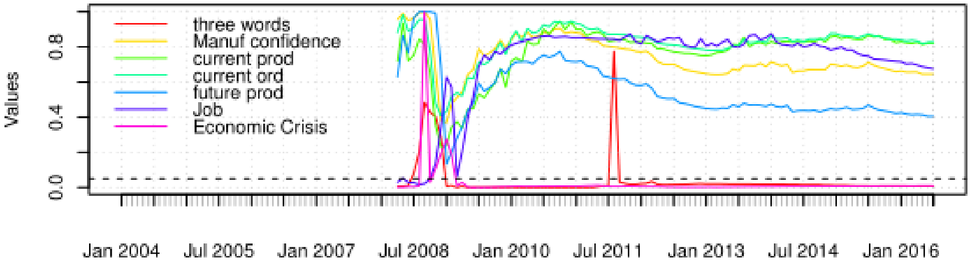


Figure 2. Engle and Granger cointegration test (p-values of the Phillips Perron test on residuals). Both tests are applied adding one observation at a time from April 2008 onwards. Sources: ISTAT official statistics and our own elaborations.

Another important result of this work can be obtained through a simple prediction based on a VECM model estimated on the IPI, the threewords index and one confidence indicator in turn. This is a way to measure the possible contribution to prediction of IPI by one or more confidence indicators and to better exploit pieces of information shared by Gtrends variables and the confidence indicators, although none of the latter cointegrate with the IPI. Again the VECM model was estimated 100 times on the month by month augmented time series, resulting in 100 forecasted values of the IPI, from May 2008 to September 2016 (see figure 3). The preliminary Johansen test confirms one rank of cointegration almost always for the confidence indicator on future orders and to a minor extent for the composite manufacturing confidence index, while the introduction of current orders or current production indicators seems to weaken the cointegration relationship between threewords and the IPI. Therefore, the 100 IPI forecasted values in figure 3 are obtained through VECM models based on IPI, threewords and the manufacturing confidence index or the confidence in future production. As can be seen predictions closely follow the actual values of the production index, in downward as well as in upward changes. The median percentage absolute error is smaller for the confidence in future production (0.9%) with respect to the manufacturing composite confidence (1.1%) mainly due to the protracted fall in the forecast for April 2009, when the IPI had already turned up. Note that these predictions are available two months earlier than the official IPI. For this reason, it is possible to claim that Gtrends data could be useful in prediction models to disclose movements of IPI when they are used in combination with hard data.

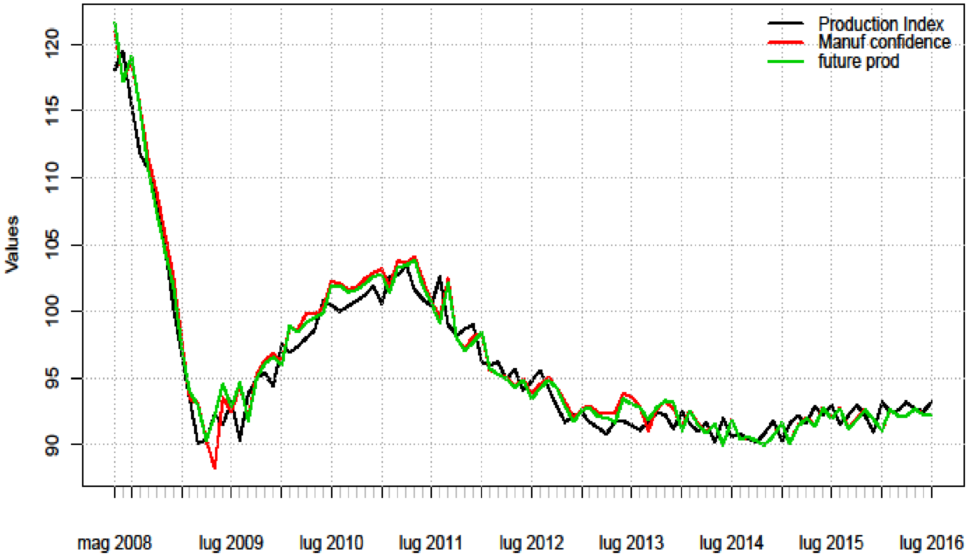


Figure 3. Recursive forecast of the IPI by VECM models using one of the listed variables together with the IPI and the three words Gtrends variable. Sources: ISTAT official statistics and our own elaborations.

References

Bodo, G., & Signorini L. F. (1987). Short-term forecasting of the industrial production index. International Journal of Forecasting 3(2), 245–259.

Bruno, G., & Lupi C. (2004). Forecasting industrial production and the early detection of turning points. Empirical economics 29(3), 647–671.

Hassani, H., Heravi, S., & Zhigljavsky A. (2013). Forecasting UK industrial production with multivariate singular spectrum analysis. Journal of Forecasting 32(5), 395–408.

Hassani, H., & Silva E. S. (2015). Forecasting with big data: A review. Annals of Data Science 2(1), 5–19.

Ulbricht, D., Kholodilin K. A., & Thomas T. (2016). Do media data help to predict german industrial production? Journal of Forecasting.

2nd International Conference on Advanced Research Methods and Analytics (CARMA2018)
Universitat Politècnica de València, València, 2018
DOI: http://dx.doi.org/10.4995/CARMA2018.2018.8576

Using big data in official statistics: Why? When? How? What for?

Mazzi, Gian Luigi
Technical Director, GOPA, Luxembourg

Abstract

This paper analyses the potential usefulness of big data in official statistics starting from four key questions such as Why? When? How? and What for - should we use big data in official statistics? To derive some answers related to empirical cases. This paper presents a big data classification by types, which is then used to identify how big data can answer to specific information needs in key policy areas. Based on the findings of these investigations, some very provisional and subjective answers to the questions raised above are derived.

Keywords: *Big data, nowcasting, indicators, policy areas.*

1. Introduction

Whether or not big data should be used in official statistics and how far we should eventually go in this direction still remain open issues for official statisticians, data scientists and analysts. It is a fact that policy makers claim for more and more information which cannot always be found in official statistics. Big data have the potential to meet policy makers' needs but, due to their nature and characteristics they face official statisticians with new challenges.

The two main obstacles for the regular use of big data in official statistics, are constituted by the evidence that they are often based on non-requested information (not collected within a robust sampling frame) and by their often unstructured nature. There are still non or partially answered questions for their use in producing official statistics such as: Why should we use big data to produce official statistics? When big data could be useful? How big data should be used to produce statistical indicators? What big data should be used for?

In this paper, we are trying to provide some answers to the questions above without pretending to be neither exhaustive nor conclusive, but aiming to provide an additional contribution to the ongoing debate. To achieve this objective, we are showing how big data could be useful in providing relevant information for key economic and socioeconomics policies, also showing advantages and drawbacks with respect to traditional sources of information.

The paper is structured as follows: Section 2 will present a big data typology which will be the basis of our investigation in the rest of the paper; Section 3 will relate the various big data types to the information needs in designing and monitoring some relevant policies; and Section 4: will conclude by providing some tentative answers to the above raised questions.

2. Big data typology

Several ways of classifying and characterizing the big data ecosystem have been proposed in the literature. Probably the most known is the so-called 4V which identify big data according to 4 main characteristics: volume, velocity, variety and veracity. Alternatively the UNECE proposes the following classification in 3 groups of the big data ecosystem: human sourced information including social networks, traditional business systems and internet of things.

With a totally new focus, mainly oriented to big data modeling, Dornik and Hendry (2015) proposed a big data classification according to the size of the data set: tall (not many variables but many observation), fat (many variables and few observations) and huge (many variables and many observations). None of the above classifications/characterizations are really fully satisfactory since they don't emphasize enough the crucial aspects represented by the very different sources originating big data. In this respect, we are proposing a so-called big data typology already proposed in Buono et al. (2017) which distinguishes big data into10 main types presented in the table below:

Table 1: Big data typologies

	Type	Main Utilisation
1	Financial market data	Macroeconomics, financial sector monitoring
2	Electronic payments data	Macroeconomics, inflation, consumers behavior
3	Mobile phone data	Labour market, sustainable development
4	Sensor data and the Internet of Things	Sustainable development, urban and environmental monitoring
5	Satellite image data	Sustainable development, economic growth and land utilisation
6	Scanner prices data	Macroeconomics, inflation, consumers behaviour
7	Online prices data	Macroeconomics, inflation, consumers behaviour
8	Online search data	Macroeconomics, sustainable development, human behavior
9	Textual data	Human sentiments, confidence, uncertainty
10	Social media data	Macroeconomics, sustainable development, human behavior

In the table above the 10 main big data types have been associated to policies and related statistical areas. In this way we have highlighted the potential usefulness of big data in relation to various statistical areas, either as a complement of traditional data sources or as an alternative to provide reliable data and to fill existing gaps. This link between big data types and statistical information needed to design and implement key macroeconomics and socioeconomics policies will be further addressed in the next session.

3. Big data contribution in designing and implementing key policies

Official statisticians are supposed to answer to policy needs providing all necessary information for the implementation, follow up and monitoring of policy actions. They provide a reliable set of statistical indicators based on traditional sources of information, such as census, sampling/surveys, and administrative data, integrated and complemented by grossing up and estimation techniques. Unfortunately not necessarly official statistics meet policy maker needs especially in terms of timeliness, relevance and ability to describe some complex phenomena.

The size of such a gap between policy makers' needs and available data can vary from country to country and also over the time. As described further, extracting information available within the big data ecosystem can help in filling the gap between the policy makers demand and the official statistics supply.

3.1 Macroeconomic growth and stability policies

When looking at macroeconomic policies, the relevance of available official statistics is quite good especially in developed countries. On the other hand, the timeliness and the frequencies at which data can be available do not necessarly meet policy maker expectations. In this respect, big data can contribute to both, increasing the timeliness of macroeconomics aggregates (i.e. GDP and consumption) and to provide higher frequency estimates of the same variables or even of inflation.

3.1.1 Economic growth

In this context financial and electronic payment data have proven to produce good results either in estimating real time economic growth or in deriving higher frequency proxies on GDP or consumption Galbraith and Tkacz (2007) Stock and Watson (2002a), Giannone, Reichlin and Small (2008) and Aprigliano et al (2016). In particular to obtain higher frequency estimates such us at weekly and even daily frequency, it is necessary to build up a quite complex modeling structure combining a data selection or data reduction tool adapted to large scale data set with mixed frequency models such as UMIDAS or MIDAS.

Also, online search data and in particular Google search data pre-synthetized within the Google Trend application provide good quality nowcasting and advanced estimates for macroeconomics variables as shown by Koop and Onorante (2013), who use the Google Trend information to select the most appropriate nowcasting model. Alternatively, several authors such as Baldacci et al (2016) and Buono et al (2018) use Google Trend data as regressors in a generalized regression model.

3.1.2 Inflation

Inflation usually measured by the Harmonized Index of consumer process (HICP) is timely available and produced at monthly frequency, meeting most of the policy makers requests. Nevertheless, starting with the paper of Silver and Heravi (2001), a large literature on the use of scanner data to estimate inflation has demonstrated how this alternative source of information can produce additional information not really present in the HICP data. In particular, they can provide higher frequency estimates of the inflation, more detailed information at product level as well as indications on retailers and consumers behavior. As a last consideration, we have to say that scanner prices data enable us to considerably reduce the burden on retailers and consumers and to reduce the production costs of inflation data. This is the main reason for which several countries such as the Netherlands and Luxembourg are planning to use in an extensive way scanner data for compiling their HICP.

3.1.3 Additional big data related information

As we mentioned in 3.1.2. scanner price data can provide useful insights on the retailers and consumers behavior. Furthermore, electronic payment data can help in better understanding the reaction of consumers to unexpected or exceptional events as shown by Galbraith and Tkacz (2011). By means of text mining and text analytics technics it is also possible to derive measures of the economic uncertainty based on textual information available in newspapers and other media. Examples of such uncertainty measures have been proposed by Baker, Bloom and Davis (2015) and Bacchini et al (2017). The use of textual information has also shown its relevance in calculating a daily business cycle indicator obtained by combining within a complex modeling structure, quarterly GDP data and daily textual information extracted from newspapers (see Thorsrud (2016). Finally, social network data can also provide useful information to study the changes in individual and collective mood or sentiment.

3.2 Labor market policies

They usually require both macro indicators such as employment/unemployment, job vacancies etc. and more and more micro indicators describing individual behavior or changes in the habits related to the employment status especially in developed countries. Official statistics provide a pretty detailed picture of the sector both at the macro and micro level. With some drawbacks related to a certain lack of timeliness and to the frequency at which different kind of information become available. The basis for labour market statistics is represented by the labour force surveys implemented in the large majority of the developed countries as well as in some of the emerging ones. The situation can be much more complex in developing countries or in underdeveloped ones where the lack of information can be really relevant to both at macro and micro level.

3.2.1 Employment and unemployment indicators

Those indicators are generally obtained by labor force surveys in developed countries being usually very reliable even if not necessarily timely available. Big data can help in increasing their timeliness either by using online search data (Google Trend), mobile phone conversation and mobile phone positioning. As an example D'Amuri and Marcucci (2012) and Tuhkuri (2016) investigate the power of big data in nowcasting and forecasting Unemployment data by using Google Trend.

On the other hand Toole et al. (2015), forecasted the employment at regional and European countries level by using the call duration information and changing behavior in social communication related to the employment status. They use an innovative approach based on Bayesian classification models.

In emerging and especially developing and underdeveloped economies, the situation can be radically different, either because labor force surveys are not being yet implemented or because they are not producing fully reliable estimates. In this case, big data can become almost the primary source for providing information on the situation of unemployment and employment.

Finally, especially mobile phone data can help in providing more granular estimates of the employment/unemployment status in a fully consistent way with the aggregated data.

3.2.2 Additional information provided by big data sources

Thanks to the availability of big data sources, and especially mobile phone ones, it is also possible to derive very useful information on individual behavior in relation to the unemployment status Sundsøy et al. (2016). Finally Nomura et al (2017) show how using big data from online job search portals, in combination with data analytics tools can produce several aggregated and disaggregated indicators extremely useful in several areas related to labor market policies such as labor market monitoring and analysis, assessing demand for workforce skills, observing job-search behavior and improving skills matching, predictive analysis of skills demand and, finally experimental studies.

3.3 Sustainable Development Goals (SDGs)

In the context of the SDGs and related policy actions, the situation in terms of availability or traditional data source to support policy decision is much more complex than in the previous cases. Often, SDGs refer to complex phenomena such as poverty, well-being, social exclusion, etc., which are measured both by qualitative and quantitative indicators. Some weaknesses of the traditional information system and consequently of official statistics appeared since the beginning of the discussion of the SDGs. For this reason, the attention was moving to alternative data sources especially big data. This is one of the main reasons for which traditionally, the so-called big data revolution has been strongly associated to SDGs activities. Going into many details by considering individual goals and discussing how big data can contribute to their measurement and achievement goes largely out of the scope of this paper also taking into account space limitations.

Nevertheless, it is worth to say that, especially some ,big data categories, such us mobile phone calls and positioning, satellite images and IOT and social networks can be particularly relevant in the context of the SGDs while others such as financial market data , electronic payment data etc., are less helpful. Obviously, the lack of official statistics and the big data availability also strongly depend on the degree of country development especially with reference to underdeveloped and developing countries. Furthermore when looking at big data it is important to carefully analyze their spatial and cross-sectional coverage in order to avoid providing misleading information.

4. Conclusions

In this paper, we have briefly investigated the usefulness of various typologies of big data in answering to policy needs in different economic and socio economic areas. We can now provide some answers to the question formulated in section 1. We would like to stress that the answers provided here reflect personal opinions and experiences and won’t pretend to be generally accepted.

Why to use big data? Because they contain an incredible and still largely unexploited amount of information of which statisticians and policy makers could benefit.

When using big data? They could be used whenever possible with traditional data source to circumvent: unsatisfactory timeliness, the lack of coverage/relevance of traditional data sources, impossibility of measuring some phenomena via surveys/sampling.

How to use big data? They should be used within a methodological sound and robust framework using advanced tools and methods especially designed to deal with specific big data features. They have also to be used in a careful manner meaning that big data should be used being aware of their limitation and drawbacks deriving from the way in which they are collected.

What should be big data used for? Increase data timeliness by means of nowcasting and advanced estimates made available already during the reference period as well as to produce high frequency estimate such as daily or weekly frequency; produce more reliable and more granular estimates of given phenomena; construct new indicators measuring phenomena for which traditional data source are weak or unavailable; provide indicators of mood, sentiment or individual and collective behavior.

In our opinion, the outcome of this paper shows quite clear that big data have the potential to complement and supplement traditional data sources (not to replace them in the production of official statistics).

References

Aprigliano, V., Ardizzi, G., Monteforte, L. (2016). Using the payment system data to forecast the Italian GDP, *Bank of Italy, Working Paper*.

Baker, S.R., Bloom, N., Davis, S.J. (2015). Measuring Economic Policy Uncertainty". *NBER Working Paper Series, Working Paper* 21633.

Bacchini, F., Bontempi, M.E., Golinelli, R., Jona-Lasinio, C. (2017). Shortand long-run heterogeneous investment dynamics". *Empirical Economics*, DOI: 10.1007/s00181-016-1211-4.

Baldacci, E., Buono, D., Kapetanios, G., Krische, S., Marcellino, M., Mazzi, G-L., Papailias, F. (2016). Big Data and Macroeconomic Nowcasting: From data access to modelling, *EUROSTAT Statistical Working Paper collection*.

Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G.L., and Papailias, F. (2017), Big data types for macroeconomic nowcasting, *Eurona*, 94-145

Buono, D., Kapetanios, G., Marcellino, M., Mazzi, G.L., (2018) and Papailias big data ecnomotrecis nowcasting and early estimates, *forthcoming in Bidsa working paper series*.

D'Amuri, F., Marcucci, J. (2012). The Predictive Power of Google Searches in Predicting Unemployment. *Banca d'Italia Working Paper*, 891.

Doornik, J. A., Hendry, D. F. (2015). Statistical Model Selection with Big Data. *Cogent Economics & Finance*, 3(1), 2015.

Galbraith, J.W., Tkacz, G. (2007). Analyzing Economic E_ects of Extreme Events using Debit and Payments System Data". *CIRANO Scienti_c Series, Working Paper* 2011s-70.

Galbraith, J.W., Tkacz, G. (2011). Electronic Transactions as High-Frequency Indicators of Economic Activity". *Bank of Canada, Working Paper* 2007-58.

Giannone, D., Reichlin, L., Small, D. (2008). Nowcasting: The Real-Time Informational Content of Macroeconomic Data", *Journal of Monetary Economics*, 55, 665-676.

Kapetanios, G., Marcellino, M., Papailias, F. (2017). Big Data and Macroeconomic Nowcasting, *Eurostat Working Paper*, ESTAT No 11111.2013.001- 2015.278.

Koop, G., Onorante, L. (2013). \Macroeconomic Nowcasting Using Google Probabilities". *European Central Bank Presentation*.

Nomura, S., Imaizumi, S., Areias, A., Yamauchi, F. (2017). Toward Labor Market Policy 2.0: The Potential for Using Online Job-Portal: Big Data to Inform Labor Market Policies in India1, *The World Bank, Policy Research Working Paper* 7966.

Silver, M., Heravi, S. (2001). Scanner Data and the Measurement of Ination. *The Economic Journal*, 111, F383-F404.

Stock, J., Watson, M. (2002a). Forecasting Using Principal Components from a Large Number of Predictors, *Journal of the American Statistical Association*, 297, 1167-1179.

Sundsøy P., Bjelland J., Reme B., Jahani E., Wetter E., Bengtsson L. (2016). Estimating individual employment status using mobile phone network data. arXiv:1612.03870 [cs.SI] (Dec 2016).

Thorsrud, L.A. (2016). \Words are the new numbers: A newsy coincident index of business cycles". *Norges Bank Working Paper Series, Working Paper* 21-2016.

Toole, J.L., Lin, Y.-R., Muehlegger, E., Shoag, D., Gonzalez, M.C., Lazer, D. (2015). Tracking employment shocks using mobile phone data. *Journal of the Royal Society Interface*, 2015 12 20150185.

Tuhkuri, J. (2016). Forecasting unemployment with google searches. *ETLA Working Paper* No 35.

A Text-Based Framework for Dynamic Shopping-Cart Analysis

Kamakura, Wagner
Rice University, USA.

Abstract

Market Basket Analysis (MBA), also known as Association Rule Mining (ARM), is already widely known and utilized by traditional and online retailers. This practice has its origin in the data-mining literature, with the introduction of association-rule mining. Despite its popularity, MBA/ARM has been criticized for its assumption that joint occurrence implies complementarity. Moreover, despite being further optimized for large-scale implementation, MBA/ARM suffers from a "curse of dimensionality," with the problem size and data sparsity growing with the square of available items. A common solution is to first classify items into pre-defined categories and carry the analysis at the category level, considerably reducing problem size. However, this simplification is prone to problems because all items within each category are automatically assumed as perfect substitutes, and categories must be mutually-exclusive, preventing an item (e.g., almonds) from belonging to more than one category (snacks, baking goods and/or bulk sales).

The main purpose of our study is to incorporate a longitudinal component into Market Basket Analysis, looking at the sequential formation of the basket, rather than its final composition only, while also reducing the dimensionality of the problem down from the number of SKU's to the most common descriptors of a shopping cart, through the text-mining of all SKU descriptors. We demonstrate empirically how dynamic MBA provides valuable insights into how the purchase of one product leads to the purchase of another, which cannot always be properly inferred from the final basket compositions in traditional MBA. Given that sequential data on shopping-cart formation is now widely available to online retailers, there is no good reason for overlooking the additional insights embedded in purchase sequences.

Keywords: *Market Basket Analysis; Dynamic Shopping Cart Analysis; Text-Mining; Hidden Markov.*

Big Data Sources for Private Consumption Estimation: Evidence from Credit and Debit Card Records

Valdivieso-Anívarro, María; Vicente, María Rosalía
Department of Applied Economics, University of Oviedo, Spain.

Abstract

Recent research has shown that big data sources provide timely granular information of human behavior. Compared to traditional survey-based statistics which rely on individuals properly recalling and reporting their actions, big data allow to track human activity with no (or few) measurement errors and low costs. This paper explores the usefulness of credit and debit card data to estimate private consumption. Little research has been done in this field due to the difficult access to these records. Using data from one-year anonymized individual credit and debit card records in one the largest cities in Latin America, patterns of private consumption are analyzed. Data on credit/debit card transactions are disaggregated by user' socio-demographic profile (i.e., age, gender, income, residence), expenditure group (35 categories), and shopping district. Records also detail when the transaction took place (day/month/year). Properties of these data will be assessed and compared with traditional survey-based sources. Moreover, sample selection bias associated with this particular big data source will be analyzed.

Keywords: *Big data; credit card records; private consumption, selection bias.*

Measuring Retail Visual Cues Using Mobile Bio-metric Responses

Dishman, Paul; Groves, Joshua; Jolley, Dale
Marketing Department and the Vivint Neuromarketing Research SMARTLab, Utah Valley University, USA

Abstract

This research provides the results of a comprehensive in-store study that utilized eye tracking to determine the initial eye attractiveness of signage and displays used in a Toyota retail dealership. Potential car buyers (n = 24) walked randomly through the showroom for the first time, and were asked to view the various signs, displays, video monitors, decorations, and vehicles on display. Research was conducted while the dealership was open in order to include distractions from human interaction. Subjects' eye movements and the objects viewed were captured using Tobii II eye tracking glasses at 60Mhz. A typical showroom self-tour lasted approximately 4:31 minutes. Subjects were then shown their results and Retrospective Think Aloud interviews were conducted with the subjects to determine positive and negative reactions to the observed objects. Signage measured included those required by Toyota, as well as those created by the dealership. Types of signage measured included digital, video, posters, stand-up cards, and ads placed on the vehicle. Each potential eye attractive object was identified and classified by type (signage, décor, digital signage, vehicle information, etc.). Every subject's results were analyzed by the number of fixations and the time spent viewing each object. The study revealed that video or digital messaging was not any more effective than static signage, but that placement of the signage was a determining factor in the effectiveness of message receptivity. Many of the non-signage objects received more attention than did certain types of advertising signage. The various attributes of the objects and signs that received positive attention were analyzed as to their eye attactiveness characteristics. Although signage in a retail showroom is believed to be critical in providing advertising and product messages, this study (in its particular environment) demonstrated that signage is not viewed by customers as often as previouly thought.

Keywords: *In-Store Signage, Retail, Store Layout, Toyota, Automotive*

Identification of helpful and not helpful online reviews within an eWOM community using text-mining techniques

Olmedilla, Maria [a]; Martinez-Torres, Rocio [b] and Toral, Sergio [c]

[a] Research Center, Léonard de Vinci Pôle Universitaire, France, [b] Departmento Administración de Empresas y Marketing, Universidad de Sevilla, Spain. [c] Departamento de Ingeniería Electrónica, Universidad de Sevilla, Spain.

Abstract

Consumers represent today a significant source of information to learn about products and services quality thanks to the proliferation of user-generated content in the form of online reviews. It is thus of paramount to understand what makes online reviews helpful to consumers as this evaluation might affect their purchase decisions. In this regard, this research has applied text-mining techniques by extracting the characteristics from online reviews' texts of an eWOM community, and further utilized these characteristics to train a logistic classifier using three classes: helpful, neutral and not helpful. The aim is identifying which unique attributes determine whether an online review is helpful or not. Findings reveal that there are much more unique attributes classified as helpful than attributes classified as not helpful. Additionally, the unique attributes associated to helpful reviews exhibit more objective appraisal while those associated to not helpful reviews show more subjective appraisal. The proposed methodology can be used to predict the helpfulness of posted reviews and to obtain their unique attributes.

Keywords: *Text mining; unique attributes; objective and subjective appraisal; eWOM communities*

Gender discrimination in algorithmic decision-making

Andreeva, Galina [a] and Matuszyk, Anna [b,c]

[a] Business School, University of Edinburgh, UK, [b] Stern Business School, New York University, USA, [c] Warsaw School of Economics, Poland.

Abstract

Most countries prohibit the use of Gender when deciding whether to give credit to prospective borrowers or not. The increasing application of automated algorithmic-based decision-making raises series of questions as to how the discrimination may arise and how it can be avoided. In this paper we analyse a unique proprietary dataset on car loans from an EU bank with the objective to understand if the minority status of females amplifies gender bias, and if there are ways to mitigate it. The initial results show that Gender is statistically significant, and women show lower probability of default. However, if Gender is excluded from the model, women have lower chances to be accepted for credit as compared to the situation when it is included. Women constitute only a quarter of the sample, and we investigate if this may lead to a representation bias which could amplify the discrimination. We experiment with under- and over-sampling and explore the effect of balancing the training set on mitigating discrimination. Logistic regression is used as a benchmark with further plans to include random forests. The results are applicable to other situations where predictive models based on historical data are used for decision-making. The presentation will discuss initial results and work in progress.

Keywords: *credit scoring; gender disrimination; algorithmic decision-making.*

Estimating traffic disruption patterns with volunteer geographic information

Bright, Jonathan[a]; Camargo, Chico[a]; Hale, Scott[a]; McNeill, Graham[a] and Raman, Sridhar[b]

[a] Oxford Internet Institute, University of Oxford, UK, [b] Oxford Brookes University, Oxford, UK.

Abstract

Accurate understanding and forecasting of traffic conditions is a key contemporary problem for local policymakers. Road networks are increasingly congested, yet data on usage patterns is often scarce or expensive to obtain, meaning that informed policy decision-making is difficult. This paper explores the extent to which traffic disruption can be estimated from static features of the volunteer geographic information site OpenStreetMap [OSM]. Kernel Density Estimates of OSM features are used as predictors for a linear regression of counts of traffic incidents at 6,500 separate points within the Oxfordshire road traffic network. For highly granular points of just 10m^2, it is shown that more than half of variation in traffic outcomes can be explained with these static features alone. Furthermore, use of OSM's granular point of interest data improves considerably on more aggregate categories which are typically used in studies of transportation and land use. Although the estimations are by no means perfect, they offer a good baseline model considering the data is free to obtain and easy to process.

Keywords: *traffic networks; land use; social media; open data.*

This project was supported by funding from InnovateUK under grant number 52277-393176, the NERC under grant number NE/N00728X/1, and the Lloyd's Register Foundation.

The educational divide in e-privacy skills in Europe

Maineri, Angelica Maria[a] ; Achterberg, Peter[a] and Luijkx, Ruud[a,b]
[a]Department of Sociology, Tilburg University, The Netherlands, [b]Department of Sociology and Social Research, University of Trento, Italy.

Abstract

This work investigates the educational divide in e-privacy skills in Europe. We ask whether the gap exists at the level of the individuals, and subsequently we seek to frame it in the European context by using the reflexive modernization theory. By using data from the Flash Eurobarometer 443 and implementing multilevel linear regression models, we confirm the presence of an educational divide in Europe, although it is mediated by the frequency of Internet use. Furthermore, the enhancement of e-privacy protecting behaviors is more likely in highly reflexive countries. Yet, there are no differences in terms of the size of the educational divide between countries. The study contributes to the literature on the second-level digital divide by focusing on e-privacy issues. Furthermore, this paper is among the first in adopting a comparative perspective when studying e-privacy issues and shows that in highly reflexive countries the educational digital divide in e-privacy skills does not widen.

Keywords: *E-privacy; digital divide; reflexive modernization*

'Whatever it takes' to change beliefs: Evidence from Twitter

Stiefel, Michael [a] and Vivès, Rémi [b]
[a]University of Zurich, Switzerland [b] Aix-Marseille Univ., CNRS, EHESS, Centrale Marseille, AMSE, France

Abstract
The sovereign debt literature emphasizes the possibility of avoiding a self-fulfilling default crisis if markets anticipate the central bank to act as the lender of last resort. Motivated by the events of summer 2012 in the eurozone, this paper investigates the extent to which changes in beliefs about an intervention of the European Central Bank (ECB) explain the sudden reduction of government bond yields for the so-called PIIGS countries. To proxy beliefs, we study Twitter data from July to September 2012 and extract beliefs using machine learning techniques. Our results are consistent with the theoretical prediction that a central bank, which credibly commits to an intervention, can restore the fundamental "good" equilibrium.

Keywords: *Self-fulfilling default crisis, unconventional monetary policy, Twitter data.*

Automated Detection of Customer Experience through Social Platforms

Bustamante, Juan [a]; Kuffo, Leonardo [b]; Izquierdo, Edgar [a] and Vaca, Carmen [b]

[a]Graduate School of Management, ESPAE-ESPOL, Ecuador. [b]Facultad de Ingeniería en Electricidad y Computación, Escuela Superior Politécnica del Litoral (ESPOL), Ecuador,

Abstract

The emergence and acceptance of social media have become a crucial aspect of daily lives in the worldwide population. As a result of this phenomenon, it is not surprising that customers' buying patterns exhibit continuous change. For capturing the experience of consumers during their visit to a retail store, previous studies have proposed in-store customer experience (ISCX) scale from data captured through traditional methods like survey research. Accordingly, ISCX is conceived as a subjective internal response to and interaction with the physical retail environment.

The present study builds upon prior research and we take the concept of ISCX with the purpose of developing an automated model for capturing ISCX from data collected through a social network like Facebook. This approach offers a low-cost, real-time alternative to traditional elicitation methods. We gathered data from English written contents by Facebook users and collected approximately 1,6 million comments made in public sites belonging to 50 companies worldwide (e.g. Clothing and jewelry retailers, whole Box and electronics Stores), including IKEA, Samsung, Whole Foods, Walmart, Tiffany, Victoria Secret, and Dillards. Five reviewers manually checked the messages filtered by the automated model, resulting in a high accuracy, confirming the high effectiveness of the model in classifying Facebook written messages.

Keywords: *Customer Experience; Machine Learning; Data Classification; Text Mining.*

Transport-Health Equity Outcomes from mobile phone location data – a case study

Grant-Muller, Susan [a]; Hodgson, Frances [a]; Harrison, Gillian [a]; Malleson, Nick [b]; Redferen, Tom [b] and Snowball, Rob [c]

[a] Institute for Transport Studies, University of Leeds, UK , [b]School of Geography, University of Leeds, UK [c]Newcastle City Council, Newcastle, UK.

Abstract

Existing cross-sectoral models of transport–health interactions have been largely developed based on traditional (mainly aggregate) data sources e.g. travel diaries, traffic counts. Such modelling provides a 'snapshot' of the transport system at fixed locations and is prone to data related weaknesses. However dissagregate health equity consequences are needed for policy development and to direct mitigating financial resources.

This research harnesses new generation smartphone location data to collect day-to-day movement patterns of individuals via an app (with consent), including mode, route and other information i.e. Track and Trace data. A case study is presented from Newcastle City region (UK), for July 2017 onwards, (circa 2000 individuals). The characteristics of the dataset generated are found to depend on: branding and key features of the app, the subset of the population who engage, the subset that actively retain the app over time, the subset who engage in additional voluntary data collection, e.g. experience sampling, voluntarily offering personal data such as age and gender and voluntarily using features such as correction of mode where the app has mis-recorded. A computed Index of Multiple Deprivation highlights interfaces with established datasources and results show the distribution of transport-health outcomes from transport initiatives by sub-group.

Keywords: *Track-and-Trace; equity, transport-health impacts, sustainable transport, smartphone*

An Unconventional Example of Big Data: BIST-100 Banking Sub-Index of Turkey

Çelik, Sadullah [a] and İşbilen, Elif [b]
[a]Department of Economics (Eng.), Marmara University, Turkey, [b]Department of Economics, Marmara University, Turkey.

Abstract

This paper applies Big Data concept to an emerging economy stock exchange market by examining the relationship between price and volume of the Banking index in BIST-100. Stock markets have been commonly analyzed in big data studies as they are one of the main sources of rich data with recordings of hourly and minutely transactions. In this sense, nowcasting the economic outlook has been related to the fluctuations in the stock exchange market as news from companies open to public became important sources of changes in expectations for economic agents. However, most of the previous studies concentrated on the main stock market indices rather than the major sub-indices. This study covers the period 13 December 2017 – 12 March 2018, with minute data and approximately 31000 observations for each of the 11 bank stocks. The effects of stock market movements on exchange rates and interest rates are also examined. The methodologies used are frequency domain Granger causality of Breitung and Candelon (2006) and wavelet coherence of Grinsted et al. (2004). The main finding is the supremacy of the banking index as it seems to have great influence on economic fluctuations in Turkish economy through other high frequency variables and the households' expectations.

Keywords: *Big data; emerging market; banking stock market index; nowcasting.*

Google matrix analysis of worldwide football mercato

Loye, Justin; Coquidé, Célestin; Rollin, Guillaume; Lages, José
Theoretical physics and Astrophysics group, Institut UTINAM, Observatoire des Sciences de l'Univers THETA, CNRS, Université de Bourgogne Franche-Comté, Besançon, France.

Abstract

The worldwide football transfer market is analyzed as a directed complex network: the football clubs are the network nodes and the directed edges are weighted by the total amount of money transferred from a club to another. The Google matrix description allows to treat every club independently of their richness and allows to measure for a given club the efficiency of player sales and player acquisitions. The PageRank algorithm, developed initially for the World Wide Web, naturally characterizes the ability of a club to import players. The CheiRank algorithm, also developed to analyze large scale directed complex networks, characterizes the ability of a club to export players. The analysis in the two-dimensional PageRank-CheiRank plan permits to determine the transfer balance of the clubs in a more subtle manner than the traditional import-export scheme. We investigate the 2017-2018 mercato concerning 2296 clubs, 6698 player transfers, and 147 player nationalities. The transfer balance is determined globally for different types of player trades (defender, midfielder, forward, ...) and for different national football leagues. Although, on average, the network transfer flows from and to clubs are balanced, the discrimination by player type draws a specific portrait of each football club.

Keywords: *Football transfer market, Google matrix, Markov chains, Complex networks, PageRank, CheiRank.*

Editorial Universitat Politècnica de València

Measuring Technology Platforms impact with search data and web scraping

Blazquez, Desamparados; Domenech, Josep and García-Álvarez-Coque, José-María
Department of Economics and Social Sciences, Universitat Politècnica de València, Spain

Abstract

In recent years, European research policies and priorities in the agricultural sector have been developed through industry-based partnerships sponsored by the European Commission (EC). In 2004, the EC regulated a form of partnership called European Technology Platform (ETP) with the aim to define research agendas that would attract private investment.

Monitoring the impact and performance of public policies, such as the implementation of ETPs, is basic for policy-makers. However, assessing the performance of ETPs frequently result into costly efforts given the current lack of indicators to monitor their variety of activities. In addition, since most ETPs have been set up recently it is difficult to assess their results, which are typically revealed after some time and take a considerable amount of time to be captured and processed with traditional methods such as surveys.

In this study, we propose to assess the dynamics of ETPs through measures based on online information, given that it is fresh, available in real-time and is a publicly reflect of the activities of organizations. We firstly consider an ETP as an innovation intermediary and define its functions according to innovation literature. Then, we enumerate the particular activities within each function in which the ETP may be involved. To monitor such functions and activities, some indicators based on online data are proposed.

This conceptual basis has been put into practice with a particular case study based on the agri-food technology platform "TP Organics". Preliminary results show that the online-based indicators are able to measure the level of activity of the platform, if its scope is expanding or reducing, and how the importance of the different functions has evolved over time.

Keywords: *European Technology Platform; agri-food sector; online data; data mining techniques; search engines; TP Organics*

Editorial Universitat Politècnica de València

Fear, Deposit Insurance Schemes, and Deposit Reallocation in the German Banking System

Fecht, Falko [a]; Thum, Stefan [b] and Weber, Patrick [c]
[a]Frankfurt School of Finance and Management, Germany, [b]Deutsche Bundesbank, Division Securities and Money Market Statistics, Germany, [c]Deutsche Bundesbank, Division Securities and Money Market Statistics, Germany.

Abstract

Recent regulatory initiatives such as the European Deposit Insurance Scheme propose a change in the coverage and backing of deposit insurances. An assessment of these proposals requires a thorough understanding of what drives depositors' withdrawal decisions. We show that Google searches for 'deposit insurance' and related strings reflect depositors' fears and help to predict deposit shifts in the German banking sector from private banks to fully guaranteed public banks. After the introduction of blanket state guarantees for all deposits in the German banking system this fear driven reallocation of deposits stopped. Our findings highlight that a heterogeneous insurance of deposits can lead to a sudden, fear induced reallocation of deposits endangering the stability of the banking sector even in absence of redenomination risks.

Keywords: *Depositor expectations, Google, Deposit insurance, Competition for depositors, Bank runs.*

Macroeconomic Indicator Forecasting with Deep Neural Networks

Cook, Thomas R [ac] and Smalter Hall, Aaron[bc]

[a] Federal Reserve Bank of Kansas City, Kansas City, Missouri, USA, [b] Federal Reserve Bank of Kansas City, Kansas City, Missouri, USA [c] The views expressed in this article are those of the authors and do not necessarily reflect the views of the Federal Reserve Bank of Kansas City or the Federal Reserve System.

Abstract

Economic policymaking relies upon accurate forecasts of economic conditions. Current methods for unconditional forecasting are dominated by inherently linear models that exhibit model dependence and have high data demands. We explore deep neural networks as an opportunity to improve upon forecast accurac y with limited data and while remaining agnostic as to functional form. We focus on predicting civilian unemployment using models based on four different neural network architectures. Each of these models outperforms benchmark models at short time horizons. One model, based on an Encoder Decoder architecture outperforms benchmark models at every forecast horizon (up to four quarters).

Keywords: *C45- Neural Networks and Related Topics; C53 - Forecasting and Prediction Methods; C14 - Semiparametric and Nonparametric Methods: General.*

Financial Stability Governance and Communication

Londono, Juan M. [a]**; Claessens, Stijn** [b]**; Correa, Ricardo** [a]**; and Mislang, Nathan** [a]
[a]Federal Reserve Board [b]Bank of International Settlements

Abstract

We investigate how differences in governance frameworks across central banks explain their financial stability communication strategies and the effect of these strategies on the evolution of each country's financial cycle. To do so, we propose a simple conceptual framework that explains how central banks conduct their communication strategy, which eventually affects the evolution of financial conditions. To empirically validate our framework, we use a database with the financial stability governance characteristics of 24 central banks and the sentiment conveyed in the financial stability reports published by these central banks. We find that, after observing a deterioration of financial conditions, central banks participating in interagency financial stability committees or with an oversight role transmit a calmer message than banks without these characteristics. We also find that the effect of communication on the evolution of the financial cycle depends on each central bank's governance framework. In particular, communication by central banks participating in an interagency financial stability committee or with a financial supervisory role has an alleviating effect on the deterioration of financial conditions.

Keywords: *Financial stability, Central bank communications, Governance, Text analysis.*

Spread the Word: International Spillovers from Central Bank Communication

Armelius, Hanna [a]; Bertsch, Christoph [b]; Hull, Isaiah [b] and Zhang, Xin [b]
[a] Ministry of Finance, Sweden; [b]Research Division, Sveriges Riksbank

Abstract

We use computational linguistic methods and a novel dataset to measure the sentiment component of central bank communications in 23 countries over the 2002-2016 period. We first construct a Granger causality network to identify how sentiment is transmitted across central banks. The network structure suggests that comovement in sentiment is not reducible to comovement in output across countries. We also show that some central banks in the network, such as the Federal Reserve and the Bundesbank, tend to cause sentiment shifts in other central banks; whereas other central banks, such as the European Central Bank and the Bank of Japan, tend to be shifted by other central banks. Finally, we use a structural VAR to demonstrate that sentiment shocks generate cross-country spillovers in sentiment, policy rates, and real variables.

Keywords: *Central Bank Communication, Monetary Policy*

The Catalonian Crises through Google Searches: A Regional Perspective

Artola, Concha and Pérez, Javier J.
DG Econonomics and Statistics, Banco de España

Abstract

In this paper we focus in the period of political turmoil starting in September 2017 in Catalonia. Our research question is the following: can the Catalan crisis be tracked by the searches done by the public on different consumption items in the Internet? We do so by focusing in two set of consumption categories: Travel to Catalonia from the main international markets (France, Germany and United Kingdom) and searches on the main consumption categories done from Catalonia and from other five big regions (Madrid, Valencia, Aragón, Andalucía and Basque Country). The preliminary results show that the uncertainty in the political situation has translated unto a decline in searches on terms associated with tourism activities in Barcelona, one broad measure shows that searches for the term "Barcelona hotel" has declined by 12%, year on year for September 2017 to January 2018, by comparison searches for hotel in other comparable Spanish regions have increased slightly. When comparing searches done from Catalonia with other regions through simple time series models, a sizeable negative residual for Catalonia is present in October 2017 –the most difficult period in the Catalan conundrum- which is not observed in other geographical areas. This is the case for some search topics associated to durable goods and Catering and Accommodation services. The political turmoil in Catalonia had significant negative effects in two consumption categories: Theaters and Restaurants.

Keywords: *Google Trends, International Tourism, Catalonia, Private Consumption*

The Sentiment Hidden in Italian Texts Through the Lens of A New Dictionary[1]

Bruno, Giuseppe [a]; Marcucci, Juri [a]; Mattiocco, Attilio[a]; Scarnò, Marco [b] and Sforzini, Donatella [b]

[a] DG Econonomics and Statistics. Bank of Italy, Rome, Italy; [b] Cineca, Rome, Italy

Abstract

The aim of this work is to propose a strategy to classify texts (or parts of them) in an ordinal emotional scale to gauge a sentiment indicator in every domain. In particular, we develop a new dictionary for the Italian language which is built using an objective method where the polarities of synonyms and antonyms are accounted for in an iterative process. To build our sentiment indicator negations and intensifiers are also used, thus considering the context in which the single word is written. We apply our new dictionary to extract the sentiment from a set of around 40 issues of the Bank of Italy quarterly Economic Bulletin. Our results show that our strategy is able to correctly identify the sentiment expressed in the Bulletins, which is correlated to the main macroeconomic variables (such as national GDP, investment, consumption or unemployment rate). Our analysis shows that sentiment represents not only an evaluation of the stylistic way in which texts are written, but also a valid synthesis of all the external factors analysed in the same document.

Keywords: *Text analysis, Sentiment analysis*

[1] The views are those of the authors only and do not imply those of the Bank of Italy.

X11-like Seasonal Adjustment of Daily Data

Ladiray, Dominique [a]; Mazzi, GianLuigi [b]
[a]INSEE, France; [b]GOPA, Luxembourg.

Abstract

High frequency data, i.e. data observed at infra-monthly intervals, have been used for decades by statisticians and econometricians in the financial and industrial worlds. Weekly data were already used in the 20's by official statisticians to assess the short-term evolution of the Economy. For example, Crum (1927) studied the series of weekly bank debits outside New York city from 1919 to 1026 and proposed a method to seasonally adjust these data based on the median-link-relative method developed by Persons (1919).

Nowadays, these data are ubiquitous and concern almost all sectors of the Economy. Numerous variables are collected weekly, daily or even hourly, that could bring valuable information to official statisticians in their evaluation of the state and short-term evolution of the Economy. But these data also bring challenges with them: they are very volatiles and show more outliers and breaks; they present multiple and non integer periodicities and their correct modeling implies numerous regressors: calendar effects, outliers, harmonics.

The current statistician's traditional toolbox, methods and algorithms, has been developed mainly for monthly and quarterly series; how should these tools be adapted to handle time series of thousands observations with specific characteristics and dynamics efficiently?

We present some ideas to adapt the main seasonal adjustment methods, and especially "the X11 family" i.e. methods based on moving averages like X11, X11-ARIMA, X12-ARIMA and X-13ARIMA-SEATS. We also make some recommendations about the most appropriate methods for pretreatment and filtering of daily and weekly data.

Keywords: *Seasonal adjustment, high-frequency data, ruptures, calendar effects.*

Empirical examples of using Big Internet Data for Macroeconomic Nowcasting

Kapetanios, George [a]; Marcellino, Massimiliano [b] and Papailias, Fotis [a]
[a] King's Business School, King's College London, [b] Bocconi University, Italy

Abstract

In this paper we present results for nowcasting and one-month-ahead forecasting three key macroeconomic variables: inflation (measured by the month on month growth rate in the Harmonized Index of Consumer Prices), retail sales (measured by the Retail Trade Index), and the Unemployment Rate. The exercise is conducted recursively in a pseudo out of sample framework, using monthly data for three economies: Germany, Italy and the UK. We assess the relative performance of Big Data (proxied via weekly Google Trends) and standard indicators (based on a large set of economic and financial variables). We also evaluate the role of several econometric methods and alternative specifications for each of them (with or without big data), for a total of 279 models and model combinations. In general, we find that Google Trends tend to slightly improve the forecasts of factor models and penalized regressions. Furthermore, a data-driven automated model selection strategy, where the forecasts from a set of best performing models over the recent past are pooled, performs particularly well, with Big Data present in about 65% of the pooled models (on average across the cases where the strategy is the best model).

Keywords: *Big Data; Macroeconomic Nowcasting; Unstructured data*

Using big data at Istat: forecasting consumption

Bacchini, Fabio; Iannaccone, Roberto and Zurlo, Davide
Istat, Econometric Studies and Economic Forecasting Division, Rome, Italy

Abstract

In our paper we discuss the possibility to use Big Data on payment instruments to improve the short-term forecast of Italian household consumption.

The literature on forecasting has evolved rapidly in the last few years. Several papers have been focused on the use of variable selection methods on large dataset of economic indicators that can potentially improve the forecasting of the main macroeconomic variables. The variable selection methods implemented anyway are always based on economic indicator (soft or hard) released by the Statistical Offices.

More recently given the presence of several sources of data on real-time economic activity available from Google, MasterCard, Facebook and many others, the use of Big data for macroeconomic forecasting has started to be exploited. With respect to Official Statistics Big data could provide potentially important complementary information based on different information sets. Moreover, compared to economic indicators, Big data are timely available and, generally, not subject to any revision process. Between different Big data possible sources, data on payment instruments (cheques, credit transfers, direct debits, payment cards) represent a relevant source of information for short-term forecasting of the main macroeconomic variables.

Concerning consumption, they capture a wide range of spending activities and are available on a very timely basis. One of the issue of Big data anyway is to structure them in a statistical form. To reach this aim data needs to be controlled for outliers and then seasonally adjusted. The ability of retail payment data to forecast the short-term development of household consumption (both for durable and non-durable goods) has been tested compared to traditional benchmark.

Keywords: *Big Data; Househouled Consumption*

Mining Big Data in statistical systems of the monetary financial institutions (MFIs)

Ashofteh, Afshin
Nova University of Lisbon, Portugal.

Abstract

The financial crisis prompted a number of statutory and supervisory initiatives that require great disclosure by financial firms of their data to a central system. Recently, core banking and payment systems data as a main big data sources of monetary financial Institutions (MFI's) have been used to monitor different kind of risks, however distress situations for MFI's are relatively infrequent events and as the same time under the pressure of rapid changes in compliance and rules. The very limited information for distinguishing dynamic fraud from genuine customer or monetary and financial institution behavior in an extremely sparse and imbalanced big data environment with probable change points in data distribution is making the instant and effective fraud detection and banking Big Data management more and more difficult and challenging. Being still a recent discipline, few research has been conducted on imbalanced classification for Big Data. The reasons behind this are mainly the difficulties in adapting standard techniques to the MapReduce programming style and inner problems of imbalanced data, namely lack of data, small disjuncts and data distribution changes. These are accentuated during the data partitioning to fit the MapReduce programming style and data mining process. This paper is going to summarize some technical problems of imbalanced data and artificial data for the adjustment of big data for MFI's and to investigate how it can be made ready for implementation.

***Keywords**: Big Data, Artificial data, Imbalanced classification, Monetary financial institutions.*