Noname manuscript No.
(will be inserted by the editor)

Monitoring E-commerce Adoption from Online Data

Desamparados Blazquez · Josep Domenech · Jose A. Gil · Ana Pont

Received: date / Accepted: date

Abstract The purpose of this paper is to propose an intelligent system to automatically monitor the firms' engagement in e-commerce by analyzing online data retrieved from their corporate websites. The design of the proposed system combines web content mining and scraping techniques with learning methods for Big Data. Corporate websites are scraped to extract more than 150 features related to the e-commerce adoption, such as the presence of some keywords or a private area. Then, these features are taken as input by a classification model that includes dimensionality reduction techniques. The system is evaluated with a data set consisting of 426 corporate websites of firms based in France and Spain. The system successfully classified most of the firms into those that adopted e-commerce and those that did not, reaching a classification accuracy of 90.6%. This demonstrates the feasibility of monitoring e-commerce adoption from online data. Moreover, the proposed system represents a cost-effective alternative to surveys as method for collecting e-commerce information from companies, and is capable of providing

This work has been partially supported by the Spanish Ministry of Economy and Competitiveness with grant TIN2013-43913-R, and by the Spanish Ministry of Education with grant FPU14/02386.

Desamparados Blazquez
Department of Economics and Social Sciences, Universitat Politècnica de València, 46022 Valencia (Spain)
E-mail: mdeblzso@upvnet.upv.es

Josep Domenech
Department of Economics and Social Sciences, Universitat Politècnica de València, 46022 Valencia (Spain)
E-mail: jdomenech@upvnet.upv.es

Jose A. Gil
Department of Computer Engineering, Universitat Politècnica de València, 46022 Valencia (Spain)
E-mail: jagil@disca.upv.es

Ana Pont
Department of Computer Engineering, Universitat Politècnica de València, 46022 Valencia (Spain)
E-mail: apont@disca.upv.es

more frequent information than surveys and avoids the non-response errors. This is the first research work to design and evaluate an intelligent system to automatically detect e-commerce engagement from online data. This proposal opens up the opportunity to monitor e-commerce adoption at a large scale, with highly granular information that otherwise would require every firm to complete a survey. In addition, it makes it possible to track the evolution of this activity in real time, so that governments and institutions could make informed decisions earlier.

Keywords Corporate websites · online data · e-commerce indicator · short-term monitoring · web scraping · Big Data learning methods

1 Introduction

Internet and the WWW (the Web) have emerged as the main drivers of the transition into the digital society and economy, transforming day by day our lifestyle and habits. This is particularly relevant and challenging for companies, which are enforced to adapt to the new way of doing business that this digital context implies. Indeed, with the empowering influence of the Internet, the Web has turned into much more than a window for firms to show their products and services to the world. Nowadays, firms can use their websites as commercial platforms by engaging in electronic commerce (e-commerce).

More and more firms are adopting e-commerce because of the great advantages and new opportunities that it involves. For instance, it helps firms to reduce costs, be closer to the clients and provide a more customized service [53]. In addition, engaging in e-commerce encourages the adoption of other innovations, such as electronic exchange data systems or automated inventory management [44, 52]. For all these reasons, the global B2C e-commerce market is in expansion. In 2015, B2C e-commerce sales worldwide reached up to nearly €2.3 trillion, which represents an increase of about 20% with respect to the previous year and a contribution of 3.1% to the global gross domestic product [16].

To promote and monitor the evolution of the digital economy, governments require detailed and updated information about the level of adoption of e-commerce in firms, grouped by different economic sectors and geographic areas. The information about the current e-commerce adoption, evolution and trends help them to better define strategic plans for the economy and enact laws for regulating this activity [42, 50].

In fact, private and public institutions are aware of the value of these data so that they are making efforts to monitor e-commerce evolution. The national and supranational statistics offices and other e-commerce observers currently obtain these data using surveys [20, 31, 43]. This traditional method that turns to primary sources bring truthful information, but the procedure implies a number of disadvantages [23, 45, 10]. First, the processing costs are high considering the human resources involved. Second, the generally low response rate complicates the characterization of the variables under study, potentially introduces bias and leads to higher survey costs and complexity in terms of design, implementation and data processing. The time taken by this data processing makes surveys inappropriate to conduct a short-term monitoring of the economy, which in the digital era when changes happen fast is more important than ever.

To deal with the shortcomings mentioned above, there is an increasing tendency to use online data which, appropriately extracted and processed, can result in accurate and prompt indicators for a variety of economic topics, ranging from unemployment to car sales or export orientation [13, 60, 7]. Online data show some advantages compared to traditional sources of information when generating economic indicators, as pointed out by Edelman [17] and Einav and Levin [18]. These include: very fast processing times because of their electronic nature; a high level of granularity; lower collection costs; and the availability of an enormous quantity of fresh information because people, business and governments generate and share information online every minute of the day.

This digitized and huge amount of information, increasingly known as "Big Data", imply challenges for computation and statistics, such as data storage and processing scalability, noise aggregation or spurious correlations [21]. For these reasons, Big Data requires specifically-developed computational and statistical techniques ("Big Analytics") that allow their appropriate exploitation, in real time and more efficiently than traditional methods would [18, 47, 6].

In this vein, this paper proposes a Big Data approach to generate an indicator for e-commerce adoption from the analysis of information provided by corporate websites. In order to make the information retrieval process from these sources quick and feasible, we have designed and implemented the System for Automatically Monitoring E-commerce Adoption (SAME), which is an intelligent system aimed at automatically capturing and processing economic related data from websites by making use of web scraping techniques and learning methods for Big Data. The final output of the system is the e-commerce adoption indicator, which is produced by means of a classification model in which more than 150 features extracted from websites were included.

The system performance has been evaluated with a sample of 426 corporate websites of manufacturing firms based in France and Spain, which were manually classified. Results show that SAME manages to predict the availability of e-commerce activity with a precision about 91%. A more detailed analysis evidenced that, as expected, websites with e-commerce tend to include some specific keywords and have a private area. The automatic nature of the proposal enables a large scale monitoring of the economy, providing prompt and actual information to governments and organizations.

The remainder of this paper is organized as follows. Section 2 reviews the literature on the generation of economic indicators from online data. Section 3 describes the architecture of the system developed for monitoring e-commerce availability on corporate websites. Section 4 describes the experimental results, including an overview of the data and a comprehensive analysis of the classification model performance. Finally, Section 5 draws some concluding remarks.

2 Related work

The generation of economic indicators from Internet activity is an incipient research topic that is receiving increasing attention due to the potential relation between online data and offline phenomena [33]. Although the first attempts in this direction date back to 2009, it is in the last couple of years when the potential of online information for monitoring the economy is being revealed. Indeed, Einav

and Levin [18] suggest that economic indicators from automatically gathered online data may already be more reliable than government survey measures in some countries. To deal with such amount of digitized data, the application of Big Data retrieval and analysis techniques is being increasingly required [59, 47].

There exist some different research lines that explore the generation of economic indicators from Internet data. One of these lines is focused in relating the popularity of some keyword searches (generally obtained from *Google Trends* reports) to the evolution of specific economic magnitudes. In this vein, the seminal work by Choi and Varian [12] analyzed how some search categories in Google were related to car and home sales and to income tourists. Later, the same authors proposed a model for nowcasting the initial claims for unemployment in the US labor market by including the popularity of some specific category searches in the prediction model [13]. Similar approaches have been considered for predicting unemployment-related variables in other countries, such as France [22] and Spain [60], as well as for predicting other economic-related variables [24]. However, search engines are not the only online source for Internet-based economic indicators. For instance, social network activities have been used as a predictor of the evolution of the stock market [8, 1], box office revenues [37, 34] or telecom sales [9].

Corporate websites are also a rich source of information for monitoring what is happening in the economy, since companies usually reflect new products and intended strategies on them. Furthermore, the digital and public nature of the Web makes it possible to automatically analyze websites and generate economic indicators from them [15]. However, websites contain a large volume of unstructured information which requires from Big Data retrieval and analysis methods such as web scraping and penalized regressions. Web scraping is a retrieval technique that consists in collecting and processing information from the Web, so that it can be used for further analysis [41]. Then, some computational statistics techniques should be applied in order to reduce the information dimensionality and produce accurate economic indicators. These include the penalized regression LASSO (which will be described later), classification and regression trees (CART), neural nets or support vector machines [27, 59]. In this context, Youtie et al [61] and Arora et al [2] applied web scraping and content analysis techniques on corporate websites to track company strategies on emerging technology sectors, while Li et al [39] tracked firms' sales growth. Similarly, Blazquez and Domenech [5] described how website contents are connected to the firm export orientation, and how this information can be used to automatically monitor the export orientation of an economy [7].

The permanent interest of statistics offices in monitoring the adoption of e-commerce [20, 31, 43] demonstrates that its development is one of the aspects of the digital economy that receives important attention from policymakers. Indeed, the work conducted by the Italian National Statistics Institute (Istat) constitutes a precedent on the detection of e-commerce and other web functionalities by analyzing websites. Their proposal [3, 4] relies on content analysis on scraped websites to detect e-commerce availability (B2B and B2C indistinctly), obtaining an accuracy about 85%. With respect to this initial study, our proposal constitutes an alternative approach in which not only text content is analyzed, but also the HTML source, the HTTP headers and the website structure are considered to detect and quantify a number of features potentially related to e-commerce (e.g. having a private area or outlinks to external sites such as payment gateways).

This way, not only the accuracy could be improved (up to 91%), but also our proposal contributes to shed light on which website features are indeed related to e-commerce implementation.

Currently, most of the research efforts on e-commerce are mainly focused on developing the technology to run these sites. For instance, user behavior patterns are being explored with web usage mining techniques to make e-commerce sites more profitable by including recommendation systems [29, 51, 63] or improving site responsiveness [46, 25, 55].

E-commerce sites have also been studied to detect its success factors. In this context, by doing a manual analysis of the sites or conducting surveys to the managers, characteristics such as the design [26], website quality [38], navigability [30] and ease of purchase processing [62] are found to be determinant for the successful adoption of an e-commerce strategy.

A different approach is followed by Thorleuchter and Van den Poel [57], who apply text mining techniques to build a prediction model on the success of e-commerce companies. Similarly, Stoll and Hepp [54] proposes a technique that analyzes HTML tag attributes of e-commerce sites to discover which of the six most popular e-commerce systems is being used. Unfortunately, all the reviewed works depart from the existence of an e-commerce site, thus being inappropriate for monitoring the adoption of e-commerce by any firm.

3 SAME: A system for detecting and monitoring e-commerce adoption

This section describes the architecture of SAME, the intelligent system developed for automatically detecting and monitoring e-commerce availability. SAME follows a web content mining approach (as defined by Cooley et al [14]) to extract and analyze data from the selected corporate websites, and finally compute the e-commerce adoption indicator. Figure 1 shows the architecture of SAME, which consists of three main modules, each one implementing one of the web mining tasks described by Kosala and Blockeel [35]. These are: the capture module (resource finding), the analysis module (information selection and pre-processing), and the production module (generalization). Below, we describe how these modules were implemented.

3.1 The Capture Module

The **capture module** is the part of the system that is in charge of accessing, downloading and storing the corporate websites of the firms under analysis, which are provided as input. These websites provides us with three types of information that will be used to feed the analysis module. The first is the text content included in the HTML (or Adobe Flash SWF) resources, since companies usually describe there the activities carried out by them. The second is the HTML (or Adobe Flash SWF) code itself, as long as it includes important information about the structure and organization of the website. This includes, for instance, the URLs in the links and anchor elements or the forms to access the ordering process. Relevant keywords may appear both in the text content and in parts of the HTML attributes. The third type of information is the HTTP headers issued by the web server in the

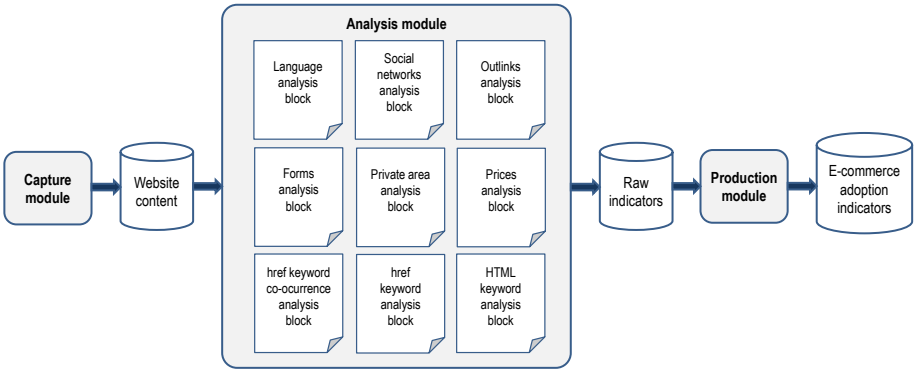

Fig. 1 Architecture of SAME.

responses of the communication protocol, because they contain information about how to interpret the resource (e.g. Content-Type, Content-Language, etc.).

This module acts basically scraping the Web and it is implemented as a modified version of HTTrack [49], which is an open source web crawler that downloads a website by recursively following the links found on the resources. HTTrack provides certain support for discovering and interpreting hyperlinks in the *javascript* code, as well as for parsing Adobe Flash SWF content, both of which are relatively common technologies in corporate websites. This allows us to follow some links that are not included in the HTML anchor tags, thus enabling a more comprehensive crawling and analysis. This module is capable of dealing with redirections, both at the HTTP and HTML level. And, also, it permits to store the response headers sent by the web server in the HTTP interactions.

In addition, this module is respectful with the companies' web servers. It respects the directives of the robots exclusion standard (i.e., *robots.txt*) so that only those websites which give permission to be crawled are actually downloaded and stored. The crawler is also configured to use a limited bandwidth and download time on each website to avoid overloading the servers because of our research.

Finally, this module is equipped with a version control system implemented with *git* in order to store snapshots of the websites at different moments of time. Although this feature is not used in the experiments presented below, it is potentially useful, as it would allow us to track firm behavior changes through time by comparing different versions of the same website.

3.2 The Analysis Module

The **analysis module** examines and processes the content previously downloaded by the capture model to produce multiple raw features that are expected to be related to the firm's e-commerce availability. This module is composed of several independent blocks that parallely compute raw website features potentially related to e-commerce. Blocks are implemented as independent scripts, each one using selected technologies for their purpose, providing flexibility to the imple-

mentation and making it easy to parellelize and distribute the tasks. Most blocks are implemented by using python and shell scripting, complemented with other command-line tools such as *sed*. Parallelization was carried out by means of the Distributed Parallel Processing Shell Script (PPSS). These blocks are the following:

- **Forms analysis block:** It is in charge of detecting the presence and number of forms, both in HTML and XHTML format, in a website. It is implemented by detecting the `<form>` element in the source file.
- **Private area analysis block:** It is set to find website areas restricted to registered users only. To do so, the block focuses on discovering forms with password fields inside them.
- **Prices analysis block:** It is designed to detect the presence of currency symbols (e.g., £, $, €) or abbreviations (e.g., GBP, USD, EUR) close to numeric characters, since this is usually indicative of the presence of prices in a website, and thus closely related to e-commerce services.
- **HTML keyword analysis block:** This block takes a list of e-commerce related keywords and counts the number of occurrences of each keyword in the text of the website. This list included terms such as: *tienda*, shop, *carro*, *panier*, cart or purchase. The block provides counting not only for strict matching (i.e., exact coincidence), but also for wide matching, which is performed by applying a word stemmer to the website text contents. Therefore, the block can also detect derived words departing from the stem of the given keywords or other words related to them. The derived words may be also helpful to detect e-commerce activity.
- **Href keyword analysis block:** It works in a way similar to the HTML keyword detector block with strict matching, but applied to the link in the *href* of an HTML anchor tag. E-commerce sites usually include words related to the ordering process in the links.
- **Href keyword co-ocurrence analysis block:** This block counts the number of terms related to e-commerce that appear in the *href* property of an HTML anchor tag. In this way, the e-commerce sites that use processing links intensively are detected.
- **Social networks analysis block:** It is in charge of detecting whether the sites include links to some of the most widespread social network sites, including Facebook, Twitter, Google+, LinkedIn, Youtube, Vimeo, Flickr and Reddit. Its implementation analyzes the `<a>` elements in the source files.
- **Outlinks analysis block:** It analyzes the HTML resources of each website in order to find references that link the contents to other external sites, such as banks, payment gateways and so on.
- **Language analysis block:** It detects the language in which the HTML resources of the corporate websites are written. Its output is the number of HTML resources in each language, being French, Spanish and English the most relevant ones in the experiments presented below.

3.3 The Production Module

Finally, the **production module** takes as input all the raw features generated by the analysis module to compute the web-based indicator for e-commerce avail-

ability. This module implements some learning methods to build the classification model for detecting e-commerce availability and to properly treat the training data. The implementation was done using the free statistical software R [48].

To build the classification model, the LASSO (Least Absolute Shrinkage and Selection Operator) regularization was applied to a logistic regression model that takes all the raw features (more than 150) as input. Logistic regression is a linear classifier that models the probability of the response variable (binary or with multiple categories) taking a particular value, and generates predictions based on the fitted probabilities. [32, 36]. In the case of this study, the logistic regression models the probability that a firm is enrolled in e-commerce. The LASSO, developed by Tibshirani [58], is a regularization method for regression models which is used to find more parsimonious models, that is, to reduce the number of variables without losing predictive performance. To do so, it includes a shrinkage parameter (λ) that makes some of the variable coefficients take value zero, thus allowing variable selection. In this study, this parameter is adjusted by means of a 10-fold cross-validation and it is selected following the "one standard error rule" [28]. This rule allows to select the most parsimonious model whose error is within one standard error of the best model's error. The LASSO is generally applied when the number of predictors is large and/or when some of them are highly correlated, given its ability to identify the most important variables and select among redundant predictors. As a result, not only more parsimonious models can be built, but also multicollinearity can be feasibly limited [58, 27]. Given its ability to reduce the dimensionality of information, it is a particularly useful technique for analyzing the huge volume of online data that is being generated day by day in a Big Data context [59].

About the training process of the classification model in this module, it is important to remark that it should be done with a balanced sample otherwise it would compromise the learning process [40]. A sample is balanced when each of the categories (or classes) of the variable under study (i.e., e-commerce availability in this case) is present in the same proportion. This way, the model is trained to be equally successful for detecting e-commerce presence or absence. The production module balances the training data by employing a method which generates new observations of the minority class (usually, websites with e-commerce) and undersamples the majority class.

4 Experimental results

4.1 Data

The evaluation of the system was performed by applying the predictive model to 426 corporate websites from manufacturing companies[1] based in France and Spain. The list of corporate websites was randomly retrieved from the SABI and ORBIS databases, which are provided by the company Bureau van Dijk. To perform the supervised learning process, we firstly did a manual revision of each website in order to detect the presence or absence of an e-commerce platform. This feature,

[1] Companies with codes 10-33 in the Statistical Classification of Economic Activities in the European Community NACE Rev. 2 [19].

which is the dependent variable in the predictive model, was coded as a binary variable with value 1 if the website had an e-commerce platform available. Then, the list of websites was provided as input for the capture module of SAME, which retrieved 21.9 GB representing 741,350 resources. After that, the analysis module processed the 426 websites to generate for each of them all the raw features (161 in total), from which the classification model for e-commerce availability selects the more relevant and is built. For descriptive purposes, they have been assembled as follows:

- **Forms:** Feature that is coded as a binary variable with value 1 if the website contained any HTML or XHTML form, and 0 otherwise.
- **Private area:** Feature that is coded as a binary variable with value 1 if there was a private area available and 0 otherwise.
- **Prices:** Feature that is coded as a binary variable with value 1 if there were product prices available on the website and 0 otherwise.
- **HTML Keywords:** This group of features make reference to the detection of keywords related to e-commerce on the HTML documents of the corporate website. A list of about 45 keywords was prepared and then searched for by SAME on the HTML documents. For each keyword and match type (which included strict and wide matching), the system coded a binary variable with value 1 if there was at least one coincidence on the website HTMLs and 0 otherwise. As a result, about 80 variables were obtained.
- **Href Keywords:** This group of features make reference to the detection of keywords related to e-commerce on the href attribute of the anchor tags of the corporate website HTMLs. The above-mentioned list of keywords was searched for by SAME on the website links, taking into account only the strict matching. For each keyword, a binary variable with value 1 was coded if there was at least one coincidence on the website links and 0 otherwise. As a result, about 45 binary variables were obtained.
- **Href Keyword Co-occurence:** This feature is related to the intensity in which the keywords appear in the HTML anchor tags. It is coded as a binary variable with value 1 if the website contained links whose href included at least two keywords of the above-mentioned list, and 0 otherwise.
- **Social networks:** Feature that refers to the presence or absence of links to any of the social networks explored by the "social networks analysis block" in the capture module of SAME. It is coded as a binary variable with value 1 if the website contained any link to a social network and 0 otherwise.
- **Outlinks:** Feature that is coded as a binary variable with value 1 if the website contained any link to external sources and 0 otherwise.
- **Language availability:** This group of features refer to the different language versions of the website that are available. The languages that have been considered are Spanish and French, because they are the native languages of the firms in the sample, and English because it is the most common language for international transactions. For this reason, this group is composed of three particular features, which are coded as three binary variables with value 1 if the website had any HTML in Spanish, French or English, respectively, and 0 otherwise.

4.2 Results

This section firstly shows some descriptive statistics that were obtained to provide a general view of the presence of the selected features on both groups of websites (with and without e-commerce). Second, the predictive model for the e-commerce availability is built by using the learning methods previously discussed. Finally, some graphical representations of the results are provided to illustrate how the model works.

4.2.1 Overview

A first approach is to explore whether or not the presence of some features differed substantially between the corporate websites with and without e-commerce. These differences are illustrated in Table 1, which also shows that the sample is unbalanced, since 60 firms offered e-commerce services while the remaining 366 did not. Regarding the presence of features, the majority of them were more frequently found in the websites of firms with some e-commerce capabilities, as expected.

A closer look to the differences reveals interesting patterns that will be useful in the classification model. For instance, it is shown that almost every website with e-commerce includes at least one HTML form, being also more frequent on them compared to those websites without e-commerce (95% vs. 69.9%). This could be related to the fact that forms are usually involved in online ordering, although they can also be used as a simple contact method. The private area exhibits a similar pattern (58.3% vs. 15.3%) since many e-commerce websites require customers to log in to gain access to the ordering or product browsing functionalities. Analogously, product prices are more frequently detected when websites include e-commerce services (60% vs. 16.7%).

About the features related to the presence of keywords, they have been presented in Table 1 under three main topic areas ("Ordering process", "Products or services offered" and "Other actions, such as payment, delivery and refund"), and also divided into content matching (CK, which include the HMTL keywords) and link matching (LK, which include the href keywords). As expected, the presence of all these groups of features was greater in the set of websites with e-commerce than in those without it. Notice that these groups were prepared for illustrative purposes only and they are not inputs for the classification model introduced below.

This descriptive analysis has helped to confirm that some of the considered features are effectively related to e-commerce and the majority of their values meet our initial expectations. These features appear, in general more frequently in websites that offer e-commerce facilities with respect to those that do not offer them. In order to determine which features are more useful to detect the presence of e-commerce, tests of statistical inference have been conducted. These tests are shown in the next section.

4.2.2 The classification model

After exploring the features that mainly differ between e-commerce and non e-commerce websites, this section evaluates the classification performance of SAME after training the production module as described above.

Table 1 Presence of key features on e-commerce and non e-commerce websites.

Feature	Mean E-commerce=1 (N=60)	Mean E-commerce=0 (N=366)
Forms	0.950	0.699
Private area	0.583	0.153
Prices	0.600	0.167
CK Ordering	0.983	0.861
CK Products	0.750	0.402
CK Other processes	0.833	0.664
LK Ordering	0.850	0.385
LK Products	0.850	0.680
LK Other processes	0.317	0.022
LK Co-occurrence	0.517	0.139
Social networks	0.550	0.257
Outlinks	0.983	0.869
French version	0.767	0.440
Spanish version	0.317	0.642
English version	0.400	0.413

Note: Keyword-related features have been grouped by topic and target area (content keywords, CK; or link keywords, LK).

At this point, it is important to remark that to properly evaluate the model, it is recommended to conduct a holdout process where the sample is split into two parts: a "training" set and a "test" set. The first set is used to train the model, so that it learns how to work, while the second is used to evaluate it (including the classification accuracy and any other result). The evaluation of the model with a different set from the one which is used to build it brings a much more reliable approach to the real performance of the model, making it possible to generalize the results [27]. From the different split ratios that are proposed in the literature, in this study it was applied the 75-25 one, in which 75% of the observations set up the training set while the remaining 25% form the test set.

To prepare the training set, 75% of the observations (i.e., 320 websites) were randomly hold out from the sample. Following the distribution of the initial sample, in which only 14% of the websites have e-commerce, this set is also unbalanced (45 websites with e-commerce vs. 275 without). That is, one of the categories of the variable under study is over-represented. Since building a robust predictive model requires a balanced training sample, the SMOTE method [11] was applied to the training set. This method artificially generates new observations of the minority class using their nearest neighbors and under-samples the majority class to obtain a balanced set. A perfectly balanced training set was obtained by generating 5 new cases for each of the 45 observations with e-commerce (270 cases in total), and randomly selecting 270 out of 275 cases without e-commerce. This way, the final training set included 540 cases. Regarding the test set, it included the remaining 25% of the initial observations (i.e., 106 websites), which were unequally distributed among both classes (15 websites with e-commerce vs. 91 websites without).

Once the two data sets are prepared, the SAME production module is trained following the procedure detailed above. Accordingly, the λ associated to the coefficient penalization parameter of the LASSO was estimated by means of a 10-fold cross-validation procedure, whose results are shown in Figure 2. By applying the

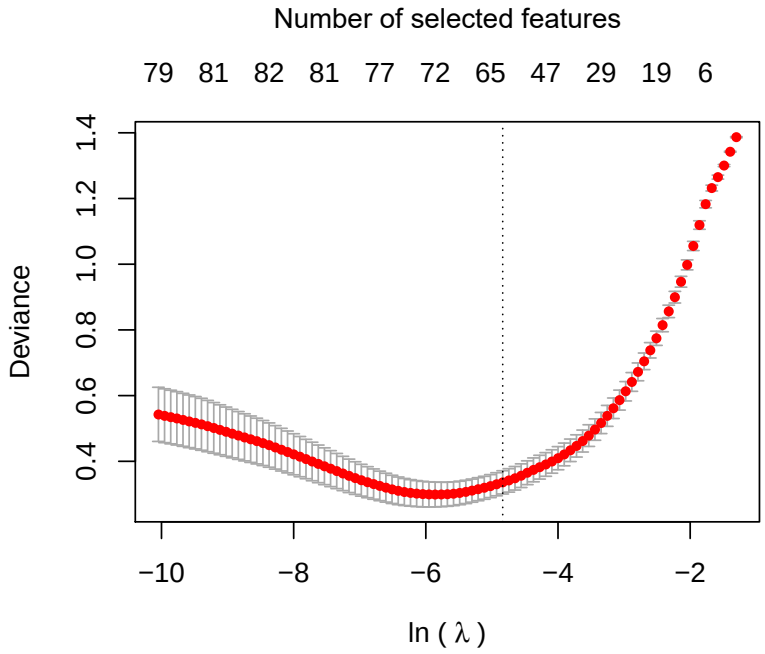

Fig. 2 Cross-validation train error (with 5% confidence intervals) for a range of λ parameter values of the LASSO method for computing the automatic e-commerce indicator. The vertical dotted line indicates the value of λ in which the error is no more than one standard error above of the minimum, following the "one standard error rule"

LASSO, a binomial logistic regression with 60 predictors (equivalent to a $\ln(\lambda)$ value of -4.837, where the value of (λ) corresponds to the largest possible such that the cross-validation error is within one standard error of the minimum) was selected out of the more than 150 raw features generated by the analysis module, as indicated by the dotted line in Figure 2. Among the selected features, we found the private area and a large number of keywords.

After training the production module, the model performance was evaluated by means of the test set. Table 2 shows the results in form of a confusion matrix. As one can observe, the classification accuracy reached 90.6% on the test set, which is considerably high. After running the experiment 100 times with different train and test sets, the 95% confidence interval for the model's accuracy ranged from 83.0% to 93.4%.

The proportion of false positives and negatives among the misclassified cases is similar. False positives correspond to websites classified as having e-commerce that actually do not have this functionality, while false negatives correspond to websites whose e-commerce functionality was not detected by SAME. In the next section, a more detailed analysis of these cases is provided.

To give a wider view of the performance of SAME, Figure 3 shows the ROC curve for the predictions on the test set. As one can observe, the model line (solid)

Table 2 Confusion matrix for the test set

		Real e-commerce availability	
		0	**1**
Predicted e-commerce availability	**0**	80.19%	3.77%
	1	5.66%	10.38%

is far from the diagonal that would represent a random model. The area under the curve (AUC) also evidences the discriminant power of the model; with a value of 0.9132 and a 95% confidence interval that ranges from 0.8333 to 0.9597, it is largely above the threshold of 0.7 for being considered very discriminant [56].

The performance of the LASSO was compared against other classifiers in order to check if, as hypothesized, it was the best for the case of this study. Results confirm that the LASSO achieve higher values for all metrics considered (see Appendix).

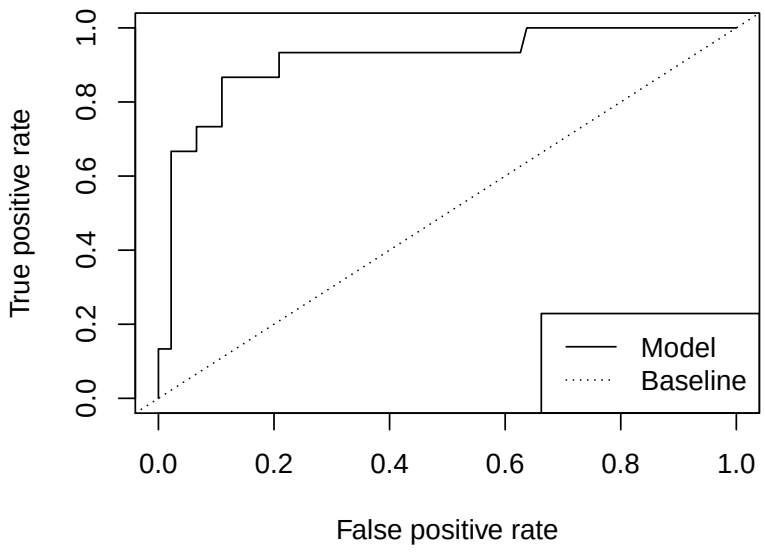

Fig. 3 ROC curve corresponding to the test set

4.2.3 More insights on the model performance

Finally, to provide more insights on how the classification model works, the predicted probabilities are calculated for the original sample of 426 corporate websites and represented against some of their features. This way, it is easy to check that as hypothesized, a greater number of features and keywords is related to a greater probability of having e-commerce. This also permits to identify which cases are outliers in order to analyze them individually and obtain information that could help to improve the performance of SAME in the future.

Figure 4 shows the relationship between the probabilities calculated by the production module and the number of assembled features reflected in Table 1 that are present in each website. The majority of websites with e-commerce have been assigned a probability greater than 0.5, so they are properly classified. In addition, it can be observed that most of these websites include a great number of the features, which is in line to what we expected in the research. The association of high probabilities with a large number of features makes that these cases are mainly located on the top right side of the plot. A more detailed analysis on the false negatives, which correspond to websites that offer e-commerce but are located near to the left side of Figure 4, revealed that most of them are caused by an incomplete crawling of the site by the capture module. This usually happens when parts of the website are developed with Flash or *javascript* technologies, given that HTTrack provides a limited support to them.

About the websites without e-commerce, most of them have been assigned a probability lower than 0.5, so they are classified as negative cases. This confirms that the model is identifying these cases properly, although the range in the number of features detected on the websites is wider in this case. Notwithstanding this, it can also be appreciated that they are mainly located on the bottom left side of the figure, which is what we expected. A more profound analysis on the false positives, which correspond to websites without e-commerce that are located near to the right side of Figure 4, revealed that, although evidences about the presence of e-commerce were found on the websites, the e-commerce functionality was available in a different URL with respect to the one analyzed, therefore it was manually coded as not offering e-commerce following the criteria employed in this study. Additionally, in some cases a variety of keywords selected as predictors by the LASSO are found on these sites, which means that these terms can be used in different contexts and are making reference to an activity which is not e-commerce.

Figure 5 shows the relationship between the probabilities calculated by the production module and the number of keyword matches (either HTML or href keywords) per website. The results of the model in this case are similar to those observed in Figure 4. Most of the websites with e-commerce are located on the right side of the graphic. As the cut point considered is 0.5, this indicates that they have been assigned probabilities greater than 0.5 by the model and so have been well predicted. The bulk of cases is concentrated among 10 and 40 keywords, meaning that the presence of e-commerce can be detected by a number of keywords.

Regarding the cases without e-commerce, most of them are located on the bottom left side of the plot. On the one hand, this indicates that they have been assigned probabilities lower than 0.5, thus being correctly classified. On the other hand, it indicates that these keywords are usually absent when the website do

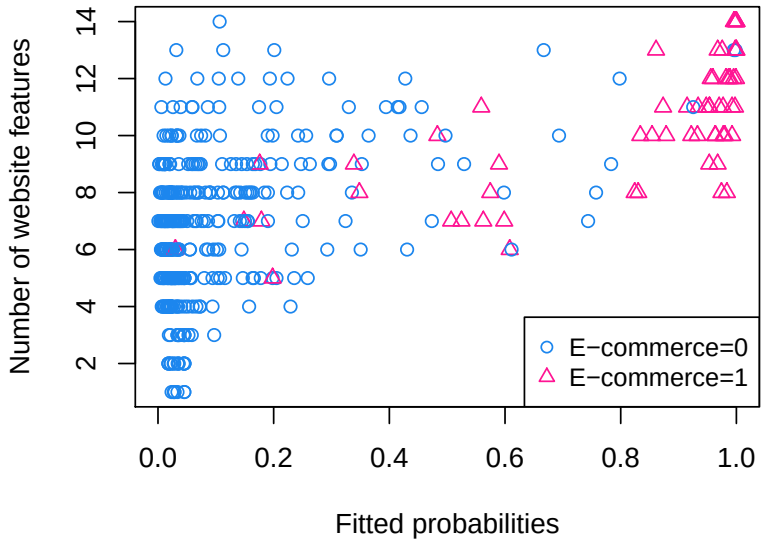

Fig. 4 Relationship between the fitted probabilities of the production module and the number of features available per website

not have e-commerce, although some of them could appear (typically including between 0 and 20 keywords).

A detailed analysis of the misclassified cases point out some ideas on how to improve the system. False negatives due to an incomplete crawling suggest that these cases could be reduced by better dealing with sites that use javascript and Flash intensively. Other lines to explore how to improve performance include considering other HTML elements potentially related to e-commerce, or extending the selection of keywords to consider other words commonly used in e-commerce websites.

5 Conclusions

The current digital environment in which the Internet provide us with fresh data in abundance offers the chance to interpret and give value to all this information in real time. To this end, Big Data analysis has emerged as a particularly useful tool that could allow companies and governments to real-time monitoring key aspects of the economy and thus implement more effective strategies.

Focusing on the growing importance of e-commerce, this paper has proposed and developed SAME, which is an intelligent system to automatically retrieve and analyze data from corporate websites to discover the adoption of e-commerce.

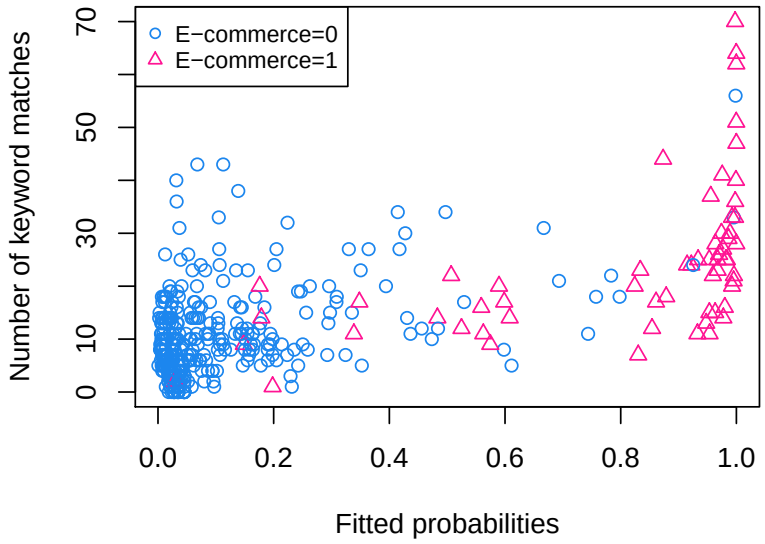

Fig. 5 Relationship between the fitted probabilities of the production module and the number of keywords (HTML and Href) available per website

Our proposal relies on a classification model that monitors e-commerce adoption in manufacturing firms by simply considering a selection of features from their websites. To do so, techniques such as web content mining and learning methods for Big Data were employed.

SAME has been tested and evaluated with the information retrieved from 426 corporate websites of firms located in France and Spain. After extracting the most relevant features from them and training the production module, the evaluation of the proposed system evidenced its accuracy for classifying firms into those that adopted e-commerce and those that did not (90.6% classification accuracy).

Detecting the presence of e-commerce on corporate websites from the automatic analysis of online data opens a new way for real-time and inexpensive monitoring. This implies a number of advantages in comparison to traditional monitoring methods based in surveys.

First, given its automatic nature, it opens up the opportunity to monitor e-commerce adoption at a large scale, thus obtaining highly granular information that otherwise would have required every firm to complete a survey. For the same reason, SAME makes it possible to track the evolution of this activity in real time. Thus, it is capable of providing us with more frequent information related to e-commerce than official surveys, which are usually conducted and processed annually.

In this way, government and institutions could make informed decisions earlier and, for instance, assess the impact of programs to promote digital sales in the short term, disaggregated by geographic area and economic sector. Moreover, business managers could also take advantage of the fresh information about e-commerce adoption in their activity sector in order to anticipate changes and reorientate the strategies of companies.

Second, it constitutes a good complement to surveys. Since the information about e-commerce adoption can be discovered directly from the Web, without specifically asking firms, this frees some space in the questionnaire to include more complex matters that necessarily involve personal intervention.

Third, SAME provide researchers with a new source of information that could be very valuable when combined with other sources. For instance, when analyzing the factors which lie behind the adoption of e-commerce; or to focus a research only in the group of adopters.

As future work, we plan to improve SAME to increase the level of detail on the provided output. That is, to generate not only an indicator for e-commerce adoption, but also to give details on its implementation, such as the integration with other information systems of the company (e.g., ERP) or the connection with a payment gateway, as well as to determine which features contribute to this implementation being successful.

In this way, SAME could provide us with information about the degree in which the e-commerce is being used by the firms, to which extent it is included in the company strategies, and how to successfully implement it.

References

1. Arias M, Arratia A, Xuriguera R (2013) Forecasting with Twitter data. ACM Transactions on Intelligent Systems and Technology 5:1 – 24, DOI 10.1145/2542182.2542190
2. Arora SK, Youtie J, Shapira P, Gao L, Ma T (2013) Entry strategies in an emerging technology: A pilot web-based study of graphene firms. Scientometrics 95:1189 – 1207, DOI 10.1007/s11192-013-0950-7
3. Barcaroli G, Nurra A, Scarnò M, Summa D (2014) Use of web scraping and text mining techniques in the istat survey on information and communication technology in enterprises. In: Proceedings of Quality Conference, pp 33–38
4. Barcaroli G, Nurra A, Salamone S, Scannapieco M, Scarnò M, Summa D (2015) Internet as data source in the istat survey on ict in enterprises. Austrian Journal of Statistics 44:31, DOI 10.17713/ajs.v44i2.53
5. Blazquez D, Domenech J (2014) Inferring export orientation from corporate websites. Applied Economics Letters 21:509 – 512, DOI 10.1080/13504851.2013.872752
6. Blazquez D, Domenech J (2017) Big data sources and methods for social and economic analyses. Technological Forecasting and Social Change Online, DOI 10.1016/j.techfore.2017.07.027
7. Blazquez D, Domenech J (2017) Web data mining for monitoring business export orientation. Technological and Economic Development of Economy Online:1 – 23, DOI 10.3846/20294913.2016.1213193

8. Bollen J, Mao H, Zeng X (2011) Twitter mood predicts the stock market. Journal of Computational Science 2:1 – 8, DOI 10.1016/j.jocs.2010.12.007
9. Bughin J (2015) Google searches and twitter mood: nowcasting telecom sales performance. NETNOMICS: Economic Research and Electronic Networking 16:87 – 105, DOI 10.1007/s11066-015-9096-5
10. Bulligan G, Marcellino M, Venditti F (2015) Forecasting economic activity with targeted predictors. International Journal of Forecasting 31:188 – 206, DOI 10.1016/j.ijforecast.2014.03.004
11. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP (2002) Smote: Synthetic minority over-sampling technique. Journal of artificial intelligence research 16:321–357
12. Choi H, Varian H (2009) Predicting the present with Google Trends. URL http://static.googleusercontent.com/external_content/untrusted_dlcp/www.google.com/en//googleblogs/pdfs/google_predicting_the_present.pdf
13. Choi H, Varian H (2012) Predicting the present with Google Trends. Economic Record 88:2 – 9, DOI 10.1111/j.1475-4932.2012.00809.x
14. Cooley R, Mobasher B, Srivastava J (1997) Web mining: information and pattern discovery on the world wide web. In: Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence, IEEE Comput. Soc, IEEE Comput. Soc, Newport Beach, CA, USA, pp 558 – 567, DOI 10.1109/TAI.1997.632303
15. Domenech J, de la Ossa B, Pont A, Gil JA, Martinez M, Rubio A (2012) An intelligent system for retrieving economic information from corporate websites. In: IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), Macau, China, pp 573 – 578, DOI 10.1109/WI-IAT.2012.92
16. Ecommerce Foundation (2016) Global B2C E-commerce Report 2016
17. Edelman B (2012) Using Internet data for economic research. Journal of Economic Perspectives 26:189 – 206, DOI 10.1257/jep.26.2.189
18. Einav L, Levin J (2014) The data revolution and economic analysis. Innovation Policy and the Economy 14:1–24, DOI 10.1086/674019
19. Eurostat (2008) NACE Rev. 2 Statistical classification of economic activities in the European Communities. EUROSTAT Methodologies and Working papers, Office for Official Publications of the European Communities, Luxembourg
20. Eurostat (2016) ICT usage and e-commerce in enterprises. URL http://ec.europa.eu/eurostat/statistics-explained/index.php/E-commerce$_s tatistics(accessed 12 December 2016)$
21. Fan J, Han F, Liu H (2014) Challenges of Big Data analysis. National Science Review 1:293 – 314, DOI 10.1093/nsr/nwt032
22. Fondeur Y, Karamé F (2013) Can Google data help predict French youth unemployment? Economic Modelling 30:117 – 125, DOI 10.1016/j.econmod.2012.07.017
23. Griffis SE, Goldsby TJ, Cooper M (2003) Web-based and mail surveys: A comparison of response, data, and cost. Journal of Business Logistics 24:237 – 258, DOI 10.1002/j.2158-1592.2003.tb00053.x
24. Hand C, Judge G (2012) Searching for the picture: forecasting uk cinema admissions using google trends data. Applied Economics Letters 19:1051 – 1055, DOI 10.1080/13504851.2011.613744

25. Hao W, Walden J, Trenkamp C (2013) Accelerating e-commerce sites in the cloud. In: 10th Anual Consumer Communications and Networking Conference (CCNC), IEEE, IEEE, p 605–608
26. Hasan B (2016) Perceived irritation in online shopping: The impact of website design characteristics. Computers in Human Behavior 54:224 – 230, DOI 10.1016/j.chb.2015.07.056
27. Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: Data mining, inference and prediction, 2nd edn. Springer
28. Hastie T, Tibshirani R, Friedman J (2013) The elements of statistical learning: Data mining, inference and prediction, 3rd edn. Springer Series in Statistics, Springer
29. He LJ (2012) The Application of Web Mining Ontology System in E-Commerce Based on FCA, vol 149, Springer Berlin Heidelberg, Berlin, Heidelberg, pp 429 – 432. DOI 10.1007/978-3-642-28658-2_65
30. Hernández B, Jiménez J, Martín MJ (2009) Key website factors in e-business strategy. International Journal of Information Management 29:362 – 371, DOI 10.1016/j.ijinfomgt.2008.12.006
31. INE (2016) Encuesta de uso de TIC y Comercio Electrónico en las empresas 2015-2016. http://ine.es/dynt3/inebase/?path=/t09/e02/a2015-2016, URL http://ine.es/dynt3/inebase/?path=/t09/e02/a2015-2016 (accessed 9 October 2016)
32. James G, Witten D, Hastie T, Tibshirani R (2013) An Introduction to Statistical Learning, Springer Texts in Statistics, vol 112. Springer New York, New York, NY
33. Jungherr A, Jürgens P (2013) Forecasting the pulse. Internet Research 23:589 – 607, DOI 10.1108/IntR-06-2012-0115
34. Kim T, Hong J, Kang P (2015) Box office forecasting using machine learning algorithms based on SNS data. International Journal of Forecasting 31:364 – 390, DOI 10.1016/j.ijforecast.2014.05.006
35. Kosala R, Blockeel H (2000) Web mining research. ACM SIGKDD Explorations Newsletter 2:1 – 15, DOI 10.1145/360402.360406
36. Kuhn M, Johnson K (2013) Applied predictive modeling, vol 810. Springer
37. Kulkarni G, Kannan P, Moe W (2012) Using online search data to forecast new product sales. Decision Support Systems 52:604 – 611, DOI 10.1016/j.dss.2011.10.017
38. Lee Y, Kozar KA (2006) Investigating the effect of website quality on e-business success: An analytic hierarchy process (ahp) approach. Decision Support Systems 42:1383 – 1401, DOI 10.1016/j.dss.2005.11.005
39. Li Y, Arora S, Youtie J, Shapira P (2016) Using web mining to explore Triple Helix influences on growth in small and mid-size firms. Technovation Online:1 – 12, DOI http://dx.doi.org/10.1016/j.technovation.2016.01.002
40. Menardi G, Torelli N (2014) Training and assessing classification rules with imbalanced data. Data Mining and Knowledge Discovery 28:92 – 122, DOI 10.1007/s10618-012-0295-5
41. Munzert S, Rubba C, Meißner P, Nyhuis D (2015) Automated Data Collection with R: A Practical Guide to Web Scraping and Text Mining. John Wiley & Sons, Chichester, UK
42. Oliveira T, Martins MF (2010) Understanding e-business adoption across industries in european countries. Industrial Management & Data Systems

110:1337 – 1354, DOI 10.1108/02635571011087428
43. ONS (2016) E-commerce and ICT Activity: 2015. URL https://www.ons.gov.uk/businessindustryandtrade/itandinternetindustry/bulletins/ecommerceandictactivity/2015 (accessed 5 December 2016)
44. Ordanini A, Rubera G (2010) How does the application of an it service innovation affect firm performance? a theoretical framework and empirical analysis on e-commerce. Information & Management 47:60 – 67, DOI 10.1016/j.im.2009.10.003
45. Peytchev A (2013) Consequences of survey nonresponse. The ANNALS of the American Academy of Political and Social Science 645:88 – 111, DOI 10.1177/0002716212461748
46. Poggi N, Carrera D, Gavaldà R, Ayguadé E, Torres J (2014) A methodology for the evaluation of high response time on e-commerce users and sales. Information Systems Frontiers 16:867–885, DOI 10.1007/s10796-012-9387-4
47. Pokorný J, Škoda P, Zelinka I, Bednárek D, Zavoral F, Kruliš M, Šaloun P (2015) Big Data Movement: A Challenge in Data Processing, Studies in Big Data, vol 9. Springer International Publishing, Cham, DOI 10.1007/978-3-319-11056-1_2
48. R Core Team (2015) R: A language and environment for statistical computing. Vienna, Austria, URL https://www.R-project.org/
49. Roche X (2014) HTTrack. URL http://www.httrack.com
50. Rodríguez-Ardura I, Meseguer-Artola A (2010) Toward a longitudinal model of e-commerce: Environmental, technological, and organizational drivers of B2C adoption. The Information Society 26:209 – 227, DOI 10.1080/01972241003712264
51. Rosaci D, Sarnè G (2014) Multi-agent technology and ontologies to support personalization in B2C E-Commerce. Electronic Commerce Research and Applications 13:13 – 23, DOI 10.1016/j.elerap.2013.07.003
52. Shih HY (2012) The dynamics of local and interactive effects on innovation adoption: The case of electronic commerce. Journal of Engineering and Technology Management 29:434 – 452, DOI 10.1016/j.jengtecman.2012.06.001
53. Sohrabi B, Mahmoudian P, Raeesi I (2012) A framework for improving e-commerce websites usability using a hybrid genetic algorithm and neural network system. Neural Computing and Applications 21:1017 – 1029, DOI 10.1007/s00521-011-0674-7
54. Stoll KU, Hepp M (2013) Detection of e-commerce systems with sparse features and supervised classification. In: 10th International Conference on e-Business Engineering (ICEBE), IEEE, IEEE, Coventry, United Kingdom, pp 199 – 206, DOI 10.1109/ICEBE.2013.30
55. Suchacka G, Borzemski L (2013) Simulation-based performance study of e-commerce Web server system–results for FIFO scheduling, Springer, p 249–259
56. Swets J (1988) Measuring the accuracy of diagnostic systems. Science 240:1285 – 1293, DOI 10.1126/science.3287615
57. Thorleuchter D, Van den Poel D (2012) Predicting e-commerce company success by mining the text of its publicly-accessible website. Expert Systems with Applications 39:13,026 – 13,034, DOI 10.1016/j.eswa.2012.05.096
58. Tibshirani R (1996) Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B (Methodological) 58:267 – 288

59. Varian HR (2014) Big Data: New tricks for econometrics. Journal of Economic Perspectives 28:3 – 28, DOI 10.1257/jep.28.2.3
60. Vicente MR, López-Menéndez AJ, 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 92:132 – 139, DOI 10.1016/j.techfore.2014.12.005
61. Youtie J, Hicks D, Shapira P, Horsley T (2012) Pathways from discovery to commercialisation: Using web sources to track small and medium-sized enterprise strategies in emerging nanotechnologies. Technology Analysis & Strategic Management 24:981 – 995, DOI 10.1080/09537325.2012.724163
62. Zhang Y, Fang Y, Wei KK, Ramsey E, McCole P, Chen H (2011) Repurchase intention in B2C e-commerce — a relationship quality perspective. Information & Management 48:192 – 200, DOI 10.1016/j.im.2011.05.003
63. Zhao WX, Li S, He Y, Wang L, Wen JR, Li X (2016) Exploring demographic information in social media for product recommendation. Knowledge and Information Systems 49:61–89

Appendix

Table 3 Comparison of classifiers performance.

Metric	LASSO Logistic Regression	Classification Tree	Random Forests	SVM
Accuracy	0.9057	0.8585	0.8491	0.8679
Sensitivity	0.9341	0.8901	0.8791	0.9011
Specificity	0.7333	0.6667	0.6667	0.6667
AUC	0.9132	0.7520	0.7833	0.7719

Table 3 reflects the performance of four different classifiers: The LASSO logistic regression, which is the classifier used to develop the study; a Classification Tree; Random Forests; and a Support Vector Machine (SVM). It includes results regarding the metrics of accuracy, sensitivity, specificity and AUC, which are the most common for comparing how well a classifier perform.