3rd International Conference on Advanced Research Methods and Analytics (CARMA2020)
Universitat Politècnica de València, València, 2020
DOI: http://dx.doi.org/10.4995/CARMA2020.2020.11322

Mining News Data for the Measurement and Prediction of Inflation Expectations

Diana Gabrielyan[1], Jaan Masso[1], Lenno Uusküla[2]
[1]University of Tartu, Tartu, Estonia, [2]Bank of Estonia, Tallinn, Estonia.

Abstract

In this paper we use high frequency multidimensional textual news data and propose an index of inflation news. We utilize the power of text mining and its ability to convert large collections of text from unstructured to structured form for in-depth quantitative analysis of online news data. The significant relationship between the household's infla-tion expectations and news topics is documented and the forecasting performance of news-based indices is evaluated for different horizons and model variations. Results sug-gest that with optimal number of topics a machine learning model is able to forecast the inflation expectations with greater accuracy than the simple autoregressive models. Addi-tional results from forecasting headline inflation indicate that the overall forecasting accu-racy is at a good level. Findings in this paper support the view in the literature that the news are good indicators of inflation and are able to capture inflation expectations well.

Keywords: *inflation; inflation expectations; news data; natural language processing; topic modelling.*

1. Introduction

Household surveys of inflation often indicate that the perception of the current inflation differs substantially from the actual values of inflation. Similarly, expectations about the future expectations differ strongly from the surveys of professional forecasters and the implied inflation rates of financial markets (for evidence see e-g- Coibion et al. 2018). Potential reason for the difference is that households and firms obtain only very partial information while doing everyday purchases and aggregating the information is very costly. Imperfect information in turn affects adversely the formation of expectations. Subjective inflation nowcasts and expectations are built through personal experiences, prior memories of inflation, and various other sources of information. One primary source of information is public media and it is well established that consumers rely largely on it when thinking about overall price changes (Blinder and Krueger 2004, Curtin 2007). Media covers a lot of news on prices and price developments.

In this paper we explore online news as novel data source for measuring inflation perception and forecasting inflation expectations by utilizing the power of text mining and its ability to convert large collections of text from unstructured to structured form. We propose a novel index of inflation news that provides a real-time indication of the price developments. Such index of inflation news captures and summarizes well the information used in the formation of expectations[1]. Available survey-based inflation expectations have low frequency and the high-frequency market-based forecasts involve risk premia and may be uncertain[2]. Our main contribution is therefore using the novel source of information to prove that online news can provide a real-time and accurate indication of consumer's expectations on inflation.

Machine learning methods are considered to be very promising avenue for academic and applied research. Although its applications are already actively used in many disciplines and research areas, it is still relatively new to economics. One modern strand of machine learning is text mining – the computational approach to processing and summarizing large amounts of text, which would be far more difficult to read, even impossible, for any single person. Extracting information from novel sources of data, such as social media (e.g. Twitter, Google) or public media (e.g. online news, communication reports) allows analysis and

[1] As Nimark and Pitschner (2018) note, since no agent has resources to monitor all events potentially relevant for his decisions, news are preferred delegates for information choice to monitor the world on their behalf. And since news mainly reports selection of events, typically major ones, coverage becomes more homogenous across different outlets.

[2] Market-based expectations are available daily but include risk premia. Survey-based expectations are published monthly. For example, for the United Kingdom, the quarterly Consumer trends data are typically published around 90 days after the end of the quarter. See https://www.ons.gov.uk/economy/nationalaccounts/satelliteaccounts/bulletins/consumertrends/apriltojune2019

different kind of understanding of economics relationships, e.g. consumer behaviour, therefore contributing to policy making and forecasting. See for example, Tuhkuri (2016), D'Amuri and Marcucci (2017), Yu et al (2018), Nyman et al (2015).

Another contribution of this work is to forecast the inflation in real-time using machine learning methods. The importance of inflation forecasting for rational decision making is well established in the literature along with the common knowledge that improving upon simple models is quite challenging. According to Medeiros et al (2019), most of this literature however ignores the recent machine learning advances. In their work they show that with machine leaning and data-rich models improving inflation forecasts is possible. Their LASSO and Random Forest models are able to produce more accurate forecasts than the standard benchmark models, e.g. autoregressive models. Similarly, Garcia, Medeiros and Vasconcelos (2017) find that high dimensional models perform very well in inflation forecasting in data rich environments. Our findings from LASSO regressions support these findings: for inflation expectations the short-term forecast errors are smaller than those of the autoregressive models. The analysis also identifies the optimal number of news topics for predicting up to five quarters ahead inflation expectations to be either four or five suggesting that the LASSO regression using optimal number of topics and best value of regularization parameter results in simpler model, which doesn't compromise the model performance. These results are, however, not robust for longer forecasting horizons and for different values of the regularization parameter. In additional results, when forecasting headline inflation, we find that the LASSO models fail to improve upon the benchmark models but demonstrate similar forecasting accuracy.

The rest of the paper is organized as follows. Section 2 describes the data sources and methodology. Section 3 and 4 provide results and an application in forecasting respectively. Section 5 concludes.

2. Data and Methodology

For official statistics, we use the Bank of England Inflation Attitude Survey data and actual UK inflation statstics. Our novel inflation news indicator is built from the article data of one of the UK leading newspaper's , Guardian, business section over the last 15 years. The choice of the news outlet is due relevance to our research in terms of content and readership, as well as the availability of open source data. As such, we chose Guardian news data for our analysis. Any news in Guardian is public and readable by anyone by default. Overall, we collected around 20,000 documents and 32 million terms from January 2004 to January 2019, which is sufficient amount of data to conduct our analysis. We only fetch articles from the business section, since this is the most relevant section for economic topics in general. In addition, articles were also filtered based on subjectively chosen key-words, which in our

opinion are relevant to inflation expectations topic. Namely, they are price, price increase, expensive, cheaper, cost, expense, bill, payment, oil, petrol, gas, diesel. The data comes in unstructured form, that is, the data is in a text form and does not have a given structure. Overall, our news corpus consists of around 100000 English language articles with well above 20 million words from January 2004 to January 2019, which is sufficient amount of data to conduct our analysis. However, the amount of data, also makes statistical computations challenge. We therefore apply data pre-processing steps suggested by Bholat and co-authors (2015) at the same time adding more steps and more developed methods. We use the text mining's bag of word approach in the text, which means all words are analysed as a single token and their structure, grammar or part of lexicon does not matter. Pre-processing results in a document term matrix, which includes all occurrences of the words in the corpus and their respective frequencies. At this step, the dimensionality of the corpus is reduced, and we get more understandable results. Frequency counts of the top 31 words in their stemmed form, that is the number of times those words appear in the final sample, are plotted in Figure 1.

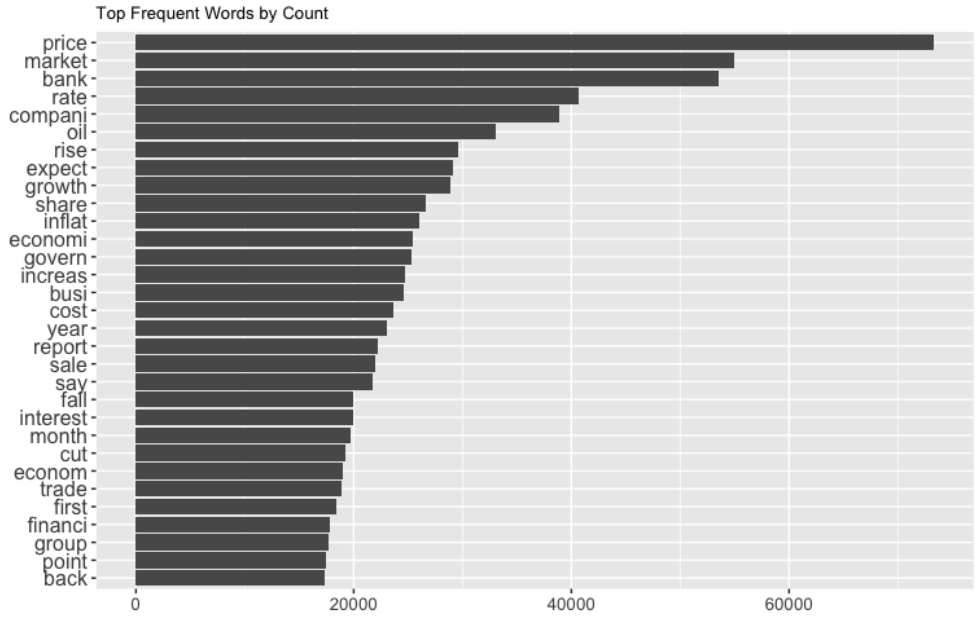

Figure 1. Top frequent words and their counts. The words are presented in stemmed form.

To proceed to building the index, we proceed with topic modelling. Since any document can be assigned to several topics at a time, the probability distribution across topics for each

document is therefore needed. Latent Dirichlet Allocation (LDA) [3] is a statistical model that identifies each document as a mixture of topics (related to multiple topics) and attributes each word to one of the document's topics, therefore, clustering words into topics. With LDA method, it is possible to derive their probability distribution by assigning probabilities to each word and document. Assigning words and documents to multiple topics also has the advantage of semantic flexibility (ex. the word 'rate' can relate both to inflation and unemployment topic). Thorstrud (2018) notes that LDA shares many features with Gaussian factor models, with the difference being that factors here are topics and are fed through a multinomial likelihood. In LDA, each document is given a probability distribution and for each word in each document, a topic assignment is made.

3. Results

For each document within a day, five most popular words are identified, and their daily frequency is calculated. This allows counting also the frequency of each topic for a given day. At this step, our results of topic decompositions and distribution are used to build the new high frequency index that will capture the intensity of inflation expectations. The index is built for every day, that is, we build daily time series using Guardian's business articles for each day. To do so, we first collect together all articles for a given day into one document, grouping them into one plain text for each day. Next, based on the first ten most frequent words in each topic the article's daily frequency is calculated. In other words, the frequency is calculated for the given day as the raw count of frequencies with which the most common words in each topic appear in that day. The news volume *I(t)* of given topic *z* is given by

$$I_z(\mathrm{t}) = \sum_{d \in I(t)} \sum_w N(d, w, z), \tag{2}$$

where $N(d, w, z)$ is the frequency with which the word *w* tagged with topic *z* appears in document d. These time series $I_z(\mathrm{t})$ are measures of volume, that is, they measure the intensity of given topic for given time period, that is for given day.

We find that some of index series are non-stationary and consequently transform them to stationary series by differencing. Augmented Dickey Fuller test is used to determine the presence of unit root and hence understand if the series are stationary or not. As such, some of the indices are evaluated as non-stationary and are transformed to by differencing.

[3] Detailed description of the LDA approach is provided in Blei, Ng and Jordan (2003).

4. Application in Forecasting

The first task is to filter information from the list of variables and select more relevant components. It is highly inefficient to use all the topic indices for predicting in such a rich dataset, as some of the regressors may be imparting redundant information. Therefore, number of topics *N* is too high and there is a definite multicollinearity present among the topic indices, as can also be observed from Figure 4. To reduce dimensionality and tackle the issue of multicollinearity[4], we use another machine learning method for variable selection. LASSO (Least Absolute Shrinkage and Selection Operator) method automates variable selection by reducing the coefficients of some features to zero, while keeping those that have the most impact on the dependent variable. LASSO's main goal is finding β that minimizes (3) with constraint $\sum_{j=1}^{p}|\beta_i| \leq t$.

$$\sum_{i=1}^{M} (y_i - \sum_{j=1}^{p} \beta_i x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_i| \quad (3)$$

λ is the shrinkage parameter and controls the strength of penalty finding the model with the smallest number of predictors that also gives a good accuracy. Therefore, the number of variables to be removed is decided by the shrinkage parameter λ, which is chosen using cross validation. Once the topic indices are selected, we forecast the inflation expectations by building a model using a direct forecast approach as given in equation (4).

$$\pi_t^{t+h} = \alpha + a * \pi_{t-1}^{t-1+h-1} + \sum_{n=1}^{N} b_n * x_{n,t-1} + u_t, \quad (4)$$

where π_t^{t+h} is the inflation (expectations) for the next *h* quarters at period *t* and $\pi_{t-1}^{t-1+h-1}$ the lagged value for the same horizon. *N* is the number of indices built from news data, b_n are vectors of unknown parameters, $x_{n,t}$ are the lagged indices and u_t is the forecasting error. We call the Equation (3) a news-based model (NBM). It is common practice to fit a model using training data, and then to evaluate its performance on a test data set. Forecast horizon *h* is also the length of the out-of-sample period (i.e. fitted values on the training set) and will be varied from 1 to 12 to compare the forecasts at different horizons and find the 'optimal' horizon defined by the lowest forecasting error. Since all of the data in this analysis is quarterly, *h* is measured in quarters. For benchmarking we use naïve AR (1) model on inflation expectations and compare the root mean squared errors (RMSE).

Table 1 reports the normalized results of estimating (3) and an AR (2) with different forecast horizons relative to simple AR (1) model. The first column of the table shows the forecast horizon, the second column (*n_var*) shows the number of variables (topics) selected by LASSO regression, and the last two columns show the root mean squared errors (RMSE) for each of the applied models. It can be seen that generally, the RMSEs are small, varying from

[4] LASSO is very robust against multicollinearity, see Friedmanet al. (2001).

0.02 to 0.76), while the forecast errors are the lowest when forecasting the next one or two period expectations using the news data. In this case the LASSO model outperforms both the naïve AR (1) and AR (2) forecasts in terms of accuracy.

Table 1. RMSEs of h-period inflation expectations forecasts using LASSO and AR (2) models. Errors are normalized relative to AR (1) benchmark.

h	n_min	RMSE_LASSO_MIN	RMSE_AR2
1	5	0.6	6
2	5	0.7	1.9
3	6	0.9	1.8
4	5	0.8	1.8
5	5	0.8	1.8
6	5	1	1.6
7	5	1.1	1.7
8	5	1	2.1
9	5	1.2	2.9
10	4	1.6	1.9
11	3	1.3	1.8
12	3	1.5	1.8

Several interesting observations can be made from Table 1. Firstly, LASSO models select different number of topics that are relevant for inflation expectations prediction for different forecast horizons. Out of our fifty topics compiled by the LDA method, LASSO selects three to six topics depending on the forecast horizon. Lagged value of the inflation expectations is always included among selected regressors and is always significant. The adjusted R-squared statistic is informative and for some horizons is as high as 70%. Thus, the selected news topic, as well as the past values of inflation expectations explain a relatively large fraction of the variation in the household's inflation expectations. One to two quarters ahead expectations can be forecasted with five topic indices as regressors, while the longer forecasts of eleven and twelve quarters can be forecasted with the best accuracy when only three

relevant topics are employed in the regression. It can also be observed that the longer the forecast horizon, the lower the forecast accuracy, which is intuitive.

These results were not robust when controlling and comparing different values of regularization parameter in the LASSO regression. There are different ways to choose the optimal value of lambda by cross-validation. Our main results in Table 1, where based on the smallest value of lambda from the cross-validation results. Table 2 compares the accuracy obtained with LASSO regression using different values of lambda shrinkage parameters against the benchmark autoregressive models. First column is the forecast horizon, while following 3 columns report the number of regressors selected by LASSO for different values of lambda. Among selected topics for all three variations of lambda, first lag of inflation expectations is selected. Column RMSE_LASSO_MIN uses the value of lambda that is equal to the minimum value of lambda chosen by cross-validation, while column RMSE_LASSO_LSE is based on the model where lambda is within one standard error. Column RMSE_LASSO_BIC is based on the lambda which is chosen using information criterion. Last two columns show the errors for benchmark AR (1) model AR (2) model, normalized relative to AR (1). Given the sparsity across normalized errors for different forecast horizons, as well as in the number of topics selected by LASSO, it can be noted that LASSO models other than that based on its minimum value are less accurate and fail to outperform the naïve models.

Table 2. RMSEs of h-period inflation expectations forecasts using different values of lambda in LASSO model, as well as AR (1) and AR (2) models. All values are normalized relative to AR (1) benchmark.

h	n_min	n_lse	n_bic	RMSE_LASSO_MIN	RMSE_LASSO_LSE	RMSE_LASSO_BIC	RMSE_AR1	RMSE_AR2
1	5	2	4	0.6	3.4	0.4	1	6
2	5	2	4	0.7	1.8	0.7	1	1.9
3	6	2	48	0.9	3.5	5.9	1	1.8
4	5	2	50	0.8	3.7	6.1	1	1.8
5	5	2	4	0.8	4.5	0.7	1	1.8
6	5	2	44	1	4.6	7.1	1	1.6
7	5	2	47	1.1	5.1	7.4	1	1.7
8	5	2	41	1	5	7.5	1	2.1
9	5	2	41	1.2	4.9	6.8	1	2.9
10	4	3	2	1.6	1.3	1.5	1	1.9
11	3	2	2	1.3	1.3	1.5	1	1.8
12	3	2	2	1.5	1.5	1.6	1	1.8

The model obtained from RMSE_LASSO_LSE includes less topics but shows poor forecasting performance. Similarly, the model from RMSE_LASSO_BIC includes even more predictors, particularly in the intermediate horizons, however, shows even worse performance. In the shorter forecasting horizons, the number of chosen topics is four, which is closer to five from the minimum lambda model, and the forecast accuracy improves. These analyses demonstrate that the optimal number of topics to predict inflation expectations up to five quarters ahead are between four and five. This also suggests that the LASSO regression, using minimum lambda as the best lambda, results to simpler model without compromising much the model performance on the test data.

It is also of interest to look how the same news data and model can be used to predict the headline inflation. We computed forecast errors for different horizons and models compared to benchmark AR (1) for annual rate of inflation and its quarterly rate. Results, not included in this chapter, but available from authors upon request suggest that while the LASSO model built using pre-selected news topics does not outperform the benchmark models, it can however be used as a forecasting model with similar forecast accuracy as those naïve models. This means that the model obtained with LASSO regression does at least as good a job fitting the information in the data as the more complicated one.

References

Alan S. Blinder & Alan B. Krueger . What Does the Public Know about Eco-nomic Policy, and How Does It Know It? *Brookings Papers on Economic Activity, Economic Studies Program, The Brookings Institution* 35(1), 327-397

Francesco D'Amuri & Juri Marcucci (2017): The predictive power of Google searches in forecasting US unemployment, *International Journal of Forecasting*, Vol. 33, No. 4, pages 801-816.

David M. Blei, Andrew Y. Ng & Michael I. Jordan (2003). Latent Dirichlet Allocation. *The Journal of Machine Learning Research* 3, 993–1022.

Erik Cambria & Bebo White: Jumping NLP Curves (2014). *A Review of Natural Language Processing Research, proceedings of Research Review Article IEEE Computational intelligence magazine*, 9, pp. 48.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning*. Vol. 1. Springer series in statistics Springer, Berlin.

Kristin J. Forbes, Lewis Kirkham & Konstantinos Theodoridis (2017). A Trendy Ap-proach to UK Inflation Dynamics. *Bank of England Working Paper* 49.

Neil Gerstein, Bart Hobijn, Fernanda Nechio & Adam H. Shapiro (2019). The Brexit price Strike. *FRBSF Economic Letter*, Federal Bank of San Francisco.

Leif A. Thorsrud (2018). Words are the new numbers: A newsy coincident index of business cycles, *Journal of Business & Economic Statistics*.

Marcelo C. Medeiros, Gabriel F. R. Vasconcelos, Álvaro Veiga & Eduardo Zil-berman (2019). Forecasting Inflation in a Data-Rich Environment: The Benefits of Machine Learning Methods, *Journal of Business & Economic Statistics*.

Kristoffer P. Nimark and Stefan Pitschner (2018). News Media and Delegated Infor-mation Choice. CEPR Discussion Papers 11323, *C.E.P.R. Discussion Papers*.

Joonas Tuhkuri: Forecasting Unemployment with Google Searches (2016). *ETLA Working Papers*, No 35.

Yu L, Zhao Y, Tang L and Yang (2018). Online big data-driven oil consumption fore-casting with Google trends, *International Journal of Forecasting*.

Rickard Nyman, David Gregory, Sujit Kapadia, Paul Ormerod, David Tuckett & Robert Smith (2015). News and narratives in financial systems: exploiting big data for systemic risk assessment, mimeo.