IFC-Bank of Italy Workshop on "Data Science in Central Banking: Applications and tools"

14-17 February 2022

News and banks' equities: do words have predictive power?[1]

Valerio Astuti, Giuseppe Bruno, Sabina Marchetti and Juri Marcucci,
Bank of Italy

[1] This contribution was prepared for the workshop. The views expressed are those of the authors and do not necessarily reflect the views of the Bank of Italy, the BIS, the IFC or the other central banks and institutions represented at the event.

News and banks' equities: do words have predictive power?

Valerio Astuti[*], **Giuseppe Bruno**[*], **Sabina Marchetti**[*] **and Juri Marcucci** [*]

[*] Bank of Italy, DG for Economics, Statistics and Research, Via Nazionale 91, 00184 Rome, Italy.

Abstract

The employment of textual data from Italian newspapers can bring useful and timely insights into the economic conditions of banks and financial intermediaries. In this work we collect textual data from the most important national newspapers, we extract news sentiment and topics and investigate their role in predicting and explaining banking market variables such as stock volumes, yield and volatility. Different full and out-of-sample experiments show that many topics have predictive power for key banking market indicators. We show this by studying the performance of our model with respect to a simple autoregressive benchmark. Our model has smaller prediction errors over the period studied, and in addition it automatically selects topic and sentiment variables as useful for the predictions. Here our goal is twofold: on one hand we provide an empirical methodology to evaluate the polarity of newspaper articles written in Italian and secondly we establish a sound statistical framework to measure the causal links between sentiment and stock market series. We deem quite relevant a quantitative evaluation of the impact of the sentiment on financial markets in order to increase the timely awareness of the regulating institutions with respect to potentially critical microeconomic conditions.

JEL classification: C83, D84, E32.
Keywords: Latent Dirichlet Allocation (LDA), news aggregation, Topic Analysis, Sentiment Analysis.

December 2021
giuseppe.bruno@bancaditalia.it

1 Introduction and motivation

Veritas numquam perit.

Extracting useful signals from textual data taken from media outlets is an important topic in the field of Artificial Intelligence. The sheer amount of detailed online information streaming from social networks is increasingly attracting the attention of many kinds of researchers and practitioners. The linguistic analysis of social media, in different languages, has become a hot topic even for applied research [1, 2]. Detection of sentiments and opinions in social media is now a critical tool for monitoring such platforms. While the idea of news-driven economics forecasts is rather simple, evaluating its relevance could be quite challenging.

In this paper, we will focus on articles extracted from a comprehensive set of Italian newspapers starting at a different time period depending on the time an agreement between Dow Jones and the newspaper's editor was established. Among these newspapers we have included all of those mentioned in the Audiweb Internet ranking total November 2017. [1]

Our analysis adds to the literature along two lines. The first one is the development of a sentiment analysis dictionary for the Banking-financial sector. The second consists of the definition and evaluation of a model for gauging the effects of news sentiments on stocks for the financial intermediation sector. Using only news sentiments, we achieved a mean directional accuracy of 80% in predicting the trends in short-term stock price movement.

The paper is arranged in the following way. In Section 2 we show how we assembled our *corpus* of Banking news. Section 3 describes the analysis for finding the number of topics and extracting them from our banking *corpus*. Section 4 explains the methodology and rules employed to carry on a sentiment analysis and extracting a banking sentiment index on the chosen *corpus*. Section 5 presents the results of the forecasting exercises for some balance sheet items, based on the sentiment index computed in the previous paragraph. Finally, Section 6 provides some concluding remarks and suggests some possible threads for future research.

[1] https://www.agcom.it/documents/10179/10214149/Studio-Ricerca+13-04-2018/4f2f5a5f-b76b-40f5-b07c-cb89359edecb?version=1.1

2 Building a *corpus* of Banking news

To build our *corpus*, we have considered the features of Dow Jones Data, News and Analytics (DNA) information aggregator platform. For the purpose of our research we have designed a query (see the appendix for details) aimed at extracting most of the articles on banks and banking agglomerates in Italian language, and appeared on the major Italian newspapers in the period ranging from September 1996 to May 2018. The different newspapers from which the articles were extracted have different coverage over time: among the paper editions the oldest available newspaper is "La Stampa", starting in 1996, whereas the most recent is "La Repubblica", starting in 2005; the online editions started much later, the most recent being "Il Sole 24 Ore Online", which started in 2013. The detail of the time coverage of the different newspapers can be found in Figure 2.

The result of the query submitted to the DNA platform consists of a total amount close to 220,000 articles, to which we applied some cleaning to remove duplicates and articles not suitable to textual analysis. More precisely, we removed duplicates and articles consisting mainly of tables and numerical data. For this purpose we employed a simple threshold based discrimination. For each article we computed the ratio R_d between the number of digits and alphabetic characters. A document is deemed suitable for our analysis when $R_d \leq .1$ In this way we excluded any document having more than 1 digit every 10 alphabetic characters. In addition we found and removed a small set of articles whose publication date was unknown. This preliminary filtering activity left us with a *corpus* of around 215,000 articles, one hundred million words and around 0.3 million of unique tokens.

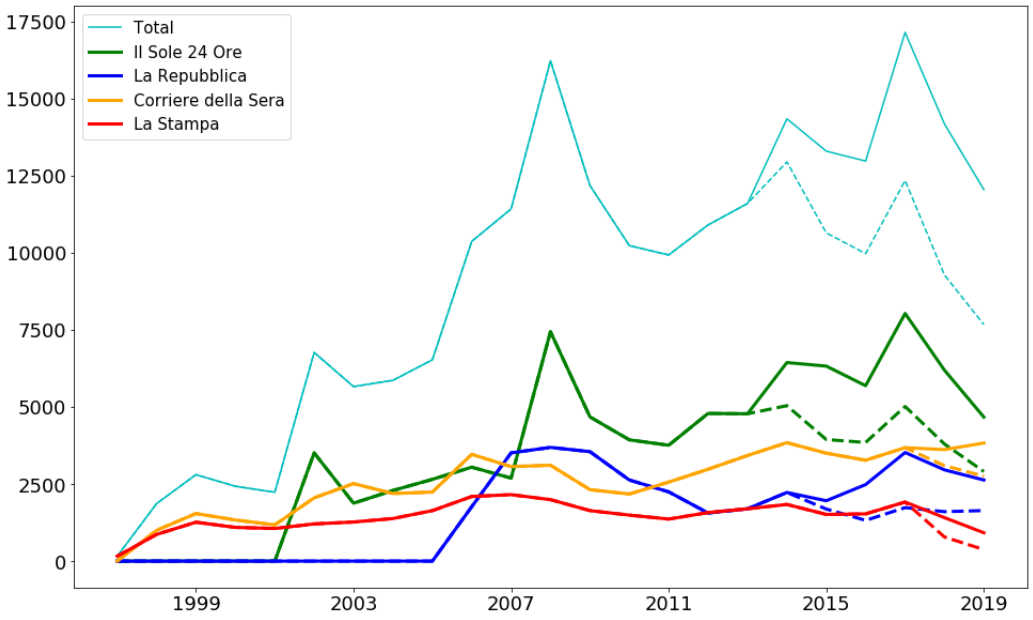


Figure 2.1: Number of articles per year (dashed lines are numbers excluding online editions)

As we will see in the following, the prediction algorithm we developed was applied only to a subset of articles, because the target variables we studied were available only from 2015 onward. This reduced the number of articles fed to the algorithm to 62,000.

The text mining preprocessing consisted of the following steps:

- **case normalization**: all capital letters were removed to discard difference between word given just by their position in the sentences;
- **punctuation removal**: non-alphanumeric characters were removed;
- **tokenization**: documents were transformed from whole strings to lists of words, treated as indepen-

dent objects. Extra spaces between words were removed;

- **stop-words removal**: a list of words carrying few informative content was selected and removed from all articles. Examples are articles and prepositions;
- **word stemming**: to reduce unjustified redundancies word-stemming was removed, identifying words like "prestito" and "prestiti" ("loan" and "loans");
- **n-grams formation**: to retain some information about co-occurrences of words we considered not only single-word tokens, but also bi-grams: after the punctuation and stop-word removal we built the list of all couples of contiguous words;
- **bag-of-words analysis**: the daily content of the articles has been analyzed as an aggregate, so the final step of the text analysis was the formation of a bag-of-words (BoW) for any given day considered.

The previous steps are standard in any text mining analysis, and functional to any natural language processing application [3].

In order to perform the topic analysis detailed in the next section, articles were *vectorized*. Vectorization consists in the representation of every article as an element in a high dimensional vector space. We used one of the simplest mappings available, the so-called *term frequency* representation. This is a type of bag-of-words representation, that only accounts for the frequency of a given token in a document, discarding information about the position and the order of words in it. The first step is the construction of a vocabulary derived from our *corpus*. In principle we could include every word with at least an appearance in a document of the *corpus*; in this way however we would be forced to consider also typos or very rare and too common words. We will see that some kind of filter in the construction of the vocabulary can be employed to retain more useful information. Once a vocabulary is completed, every word contained in it will define a dimension of the vector space in which our articles are represented. In this representation every article is a vector having as many elements as the number of words in the vocabulary. Each component of this vector is the number of times the corresponding word appears in the statement or the whole document. As an example, consider the vocabulary:

$$V = \{\text{dog}, \text{sofa}, \text{cat}, \text{chair}, \text{table}\}$$

and the following sentences:

$$\begin{aligned} s_1 &= \text{The dog chased the cat over the sofa.} \\ s_2 &= \text{We should buy a sofa for us and a sofa for our dog.} \end{aligned}$$

The two sentences have the *term frequency* representation:

$$\begin{aligned} s_{1V} &= (1, 1, 1, 0, 0) \\ s_{2V} &= (1, 2, 0, 0, 0) \end{aligned}$$

In this representation every word in the vocabulary is an additional dimension in the documents vector space, so increasing the size of the vocabulary implies raising the complexity of the document representation. For this reason setting some filters in the construction of the vocabulary can improve the document representation. In particular words with very few appearances in the whole *corpus* - being them typos or very uncommon words - are not very important in the description of documents, the associated component being null in all but few documents. Conversely, very common words will be found in almost all the documents, so the associated components will not be useful in discriminating them.
We built our vocabulary discarding words appearing in more than 90% of the articles, and in less than 10% of the articles. Moreover we put a maximum limit on the size of the vocabulary, keeping only the 300'000 most frequent words. This latter selection discards less common words, so its effect goes in the same direction of putting a lower threshold on the number of documents in which a word has to appear so as to be considered.

3 Topic Distribution

Many different methodologies are available to quantify the content of news articles. Our application starts from a set of heterogeneous newspapers (some of them mostly focused on economic and financial news, other are generalist ones). We first pinned down the most relevant topics in the *corpus*.
Hierarchical mixture modeling is one among the most powerful methodology available to find patterns and structure in large collections of data. The main reason for which these models are useful is the large dimensionality reduction they achieve. Without any particular model every document would be described by the set of words by which it is composed. This would make each document a point in a space having dimension equal to the cardinality of the vocabulary. In our *corpus* we have an order of 10^5 words, so every article would be very sparse point in a huge dimensional space.

An interesting offspring of mixture modeling is *topic modeling*, where the data under study are large collections of documents. In this circumstance mixture modeling algorithms find the underlying patterns of words that are embedded in the collection. When using topics to identify documents, they become points in a space of dimension usually around the order of 10. In addition, with respect to other dimensionality reduction methods, the results obtained with topic modelling are more interpretable by humans. Pinning down these patterns - called in this setting *topics* - allows for effective clustering, searching, exploring, predicting and summarizing large corpora of documents.

To describe the topic content of the articles in our *corpus* we use Latent Dirichlet Allocation (LDA) [4], an unsupervised method which can describe at the same time the topic content of an article and the word content of a topic. LDA is a two level generative process, in which documents are linked to topic distributions and the *corpus* is modeled as a Dirichlet distribution on these latent topics. Given a vocabulary of V words this model assumes that each document in the *corpus* is generated by the following process:

1. The number of topics K, a K-dimensional vector α and a $K \times V$ matrix $\boldsymbol{\beta}$ are assumed as parameters of the model;
2. A K-dimensional random variable $\boldsymbol{\theta}$ is selected from the $(K-1)$-simplex with a Dirichlet probability density having parameters $\boldsymbol{\alpha}$:

$$p\left(\boldsymbol{\theta}|\boldsymbol{\alpha}\right) = \frac{\Gamma\left(\sum_{i=1}^{K}\alpha_i\right)}{\prod_{i=1}^{K}\Gamma\left(\alpha_i\right)}\theta_1^{\alpha_1-1}\ldots\theta_K^{\alpha_K-1} \tag{1}$$

   where $\Gamma(x)$ is the Gamma function;
3. Vector $\boldsymbol{\theta}$ is used as a parameter for a multinomial distribution, used to pick the topic z from the K available;
4. Topic z is used to condition another multinomial distribution with parameter $\boldsymbol{\beta}$, which is used to extract a word w from the vocabulary.

Along this process the K dimensional vector $\boldsymbol{\alpha}$ has components $\alpha_i > 0$ and $\boldsymbol{\beta}$ is the probability matrix of selecting the word w_i once a topic z_j is chosen: $\beta_{ij} = p\left(w_i|z_j\right)$.

Given the parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we obtain a joint probability distribution for the set of generated words $\boldsymbol{w}$ and the latent variables characterizing the topic $\boldsymbol{z}$. Let the document be composed of N words, we obtain joint probability:

$$p\left(\boldsymbol{\theta},\boldsymbol{z},\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}\right) = p\left(\boldsymbol{\theta}|\boldsymbol{\alpha}\right)\Pi_{n=1}^{N}p\left(z_{k_n}|\boldsymbol{\theta}\right)p\left(w_n|z_{k_n},\boldsymbol{\beta}\right) \tag{2}$$

where z_{k_n} denotes the topic selected for the word w_n. Integration over the latent variables yields the probability distribution for the set of N words $\boldsymbol{w}$:

$$p\left(\boldsymbol{w}|\boldsymbol{\alpha},\boldsymbol{\beta}\right) = \int d\boldsymbol{\theta}\, p\left(\boldsymbol{\theta}|\boldsymbol{\alpha}\right)\Pi_{n=1}^{N}\sum_{k_n=1}^{K}p\left(z_{k_n}|\boldsymbol{\theta}\right)p\left(w_n|z_{k_n},\boldsymbol{\beta}\right)\ . \tag{3}$$

LDA posits a fixed number of topics in a document collection and assumes that each document reflects a combination of those topics. The process can be reverted: in particular we are interested, given a set of documents, in deriving the topic distribution more suitable to synthesize their content. In principle we are interested in the distribution of the latent variables given the observed *corpus*:

$$p\left(\boldsymbol{\theta}, \boldsymbol{z}|\boldsymbol{w}, \boldsymbol{\alpha}, \boldsymbol{\beta}\right) = \frac{p\left(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}\right)}{p\left(\boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\beta}\right)} \tag{4}$$

Even if the last expression is in general analytically intractable, a number of approximate inference algorithms can be used to find a solution. When a document collection is analyzed under these assumptions, these inference algorithms reveal an embedded thematic structure. With this structure, LDA provides a way to quickly summarize, explore, and search massive document collections.

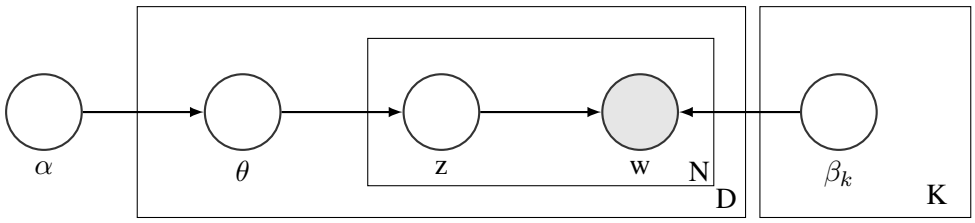


Figure 3.1: Graphical model representation of LDA.

The probabilistic graphical model in 3.1 reveals the nested structure of the LDA assumptions. LDA is composed of a hierarchy of mixture models. Each document is modeled with a finite mixture model, where the mixture proportions (i.e. the topic proportions) are drawn uniquely for each document but the mixture components (i.e. the topics) are shared across the collection. This is known as a grade of membership or mixed membership model in statistical theory [25]. LDA builds on seminal work in psychology [23] and machine learning [35]. It has close links to classical principal component analysis [18].

The most common way to evaluate a topic model is to compute the log-likelihood of a hold-out test set. This is usually done by splitting the dataset in two parts: one for training, the other for testing. For LDA, a test set is a collection of unseen documents $\{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_d}\}$, and the model is described by the topic matrix $\boldsymbol{\beta}$ and the topic weights $\{\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, \dots, \boldsymbol{\theta_d}\}$ for every document. We need to evaluate the log-likelihood:

$$\mathcal{L}(\{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_d}\}) = \log p(\{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_d}\}|\boldsymbol{\beta}, \boldsymbol{\alpha}) = \sum_{\boldsymbol{j}=1}^{\boldsymbol{d}} \log p(\boldsymbol{w_j}|\boldsymbol{\beta}, \boldsymbol{\alpha}) \tag{5}$$

The measure traditionally used for gauging the goodness of fit of topic models is the *perplexity* of the held-out documents, defined as:

$$perplexity\left(\{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_d}\}\right) = \exp\{-\frac{\mathcal{L}(\{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_d}\})}{\sum_{j=1}^{d} N_j}\} \tag{6}$$

where N_j is the number of words contained in the j-th document. The perplexity is a monotonic decreasing function of the log-likelihood $\mathcal{L}(\{\boldsymbol{w_1}, \boldsymbol{w_2}, \dots, \boldsymbol{w_d}\})$ of the unseen documents, such that minimizing the perplexity is tantamount to maximizing the likelihood function.

4 Italian Dictionary for Sentiment Analysis

One of the most basic information contained in a text is the sentiment expressed about a given subject. This is often also among the most interesting feature one would want to capture in any text analysis. For these reasons *sentiment analysis* is one of the most common applications of natural language processing [5, 6].

In its simplest formulation the goal of a sentiment analysis application is to transform any piece of text into a value on an ordered scale, representing the amount of positivity or negativity carried by the text. The usual

formats of the output are of two kinds: continuous, for example a number in the interval $[0, 1]$, or discrete, like a number of stars between 1 and 5.

The main approaches to sentiment analysis are of two kinds: rule-based algorithms and machine learning models. In the first category of applications the researcher identifies a set of keywords or features carrying the sentiment and maps each of them to a score value. One can then sum or average the scores for every identified feature to obtain a value representing the whole document. Some of the benefits of this approach - also called *vocabulary based* - are its transparency and ease of use: given that the vocabulary is built by the user he has complete control over it and can justify why each feature has a given score. In addition it can be very generic: keywords like "excellent" and "very good" have a positive connotation in most context, so a score based on such keywords will be mostly domain-independent. Another substantial advantage of a rule-based method over a machine learning model is that the latter requires a previously labeled set of documents to be trained, while the former can assign a score to documents without any previous knowledge (apart from the researcher's knowledge).

Machine learning models leave the identification of features and the assignment of scores to an automated algorithm, trained to reproduce the scores of a given set of previously labeled documents. For example one could use as training set a sample of twitter feeds associated with one of the two hashtags *#good* or *#bad*, and train a supervised algorithm to replicate the given labeling. As already mentioned the use of this class of methods is bound to the availability of a large set of previously labeled documents, and the result can be domain specific. This can be of course both an advantage or an obstacle: the automated algorithm can pick nuances of the language which cannot be decoded in a vocabulary, but the interpretation of these nuances is usually dependent on the context in which they are used, and therefore on the document set used to train the model. A machine learning approach is useful only whenever the model can be trained on a class of documents similar to the ones the model has to label.

For our analysis we used a rule-based approach, employing a vocabulary built in [7], where the authors aimed at defining a vocabulary specialized on economic and financial language. This vocabulary is based on a self consistent algorithm which takes into account scores for synonyms and antonyms of any given word, in order to enforce a coherent score assignment. The authors evalued the performance of their dictionary using as benchmark the Open Polarity Enhanced Named Entity Recognition (OpeNER) vocabulary [8], obtaining better results on every test performed.

5 Forecasting and Benchmarking Banking performances

Gauging the relevance of the information contained in news articles in explaining economic fluctuations of banking variables is of the utmost importance at both macro and micro-economic level. The sole yard stick we take into account is the ability of news to improve the predictive power toward balance sheet variables. We synthesize the information contained in the articles using the tools introduced in the previous sections: topic and sentiment analysis. With this approach we obtain a handful of variables representing the topic content and sentiment expressed in any given article, and these variables can be used as predictors for the behavior of indices related to four important Italian banks and the Italian stock index FTSE MIB.

To study the predictive power of the news variables we analyzed the period from the beginning of 2015 to May 2019[2].

Over this period we applied a 3-month wide moving window width to carry out an out-of-sample analysis: we performed an LDA analysis over the set of articles published in a first three month window, extracted the topics, and studied the weights of these topics in articles appearing in the following 3-month interval. Those weights, along with the sentiment extracted from the articles, are used to perform prediction of daily trading volumes and volatility.

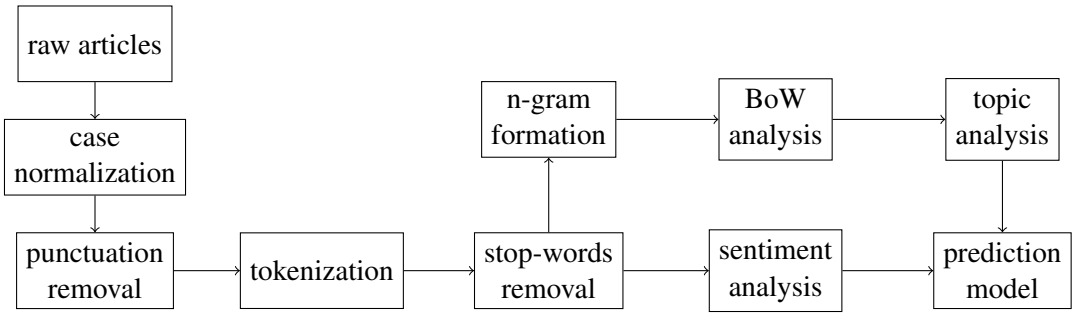
More in detail, the process consisted in dividing the whole period T going from Q1-2015 to Q2-2019 in

[2]This is the longest period for which all the data analyzed are available.

3-month windows m_j, $j \in \{1, 2, ..., 18\}$ (we have 18 quarters in the period considered). Then, for every j going from 2 to 18, we applied the following steps:

- we fit the LDA model on the *corpus* of articles appearing in the window m_{j-1}, to obtain the most important topics in the period;
- we projected every article appearing in the window m_j on the topics obtained in the last step, to obtain a topic distribution for every article;
- we performed the vocabulary based sentiment analysis on every article in the window m_j;
- we pooled all the articles relative to a single day, taking into consideration the average of the sentiment score and of the topic distribution for that day;
- we used the daily topic weights and sentiment score as predictor variables for the trading volumes and volatilities in the next day.

The following flow diagram shows a picture of the above described process, together with the preprocessing steps:



The topic analysis was performed with different choices of parameters, in order to minimize the perplexity score, enhance the interpretability of the topics, and increase the predictive power of the model. Removing the word stems in the preprocessing phase do not decrease sensibly the perplexity of the LDA, but in turn creates much less intelligible topics. For this reason we decided to skip the stemming step in the preprocessing phase.

We have a comparable number of articles in every time window, so we opted to select a constant number of topics over the whole period T. The minimum number of topics giving a satisfying perplexity was $n_{topics} = 3$, while the maximum was $n_{topics} = 5$, above this value the perplexity did not show any appreciable decrease. We took in consideration two types of models: one with single-word tokens, and a second in which bi-grams were considered. Both of them gave satisfactory results.

After the LDA and sentiment analysis the output variables were used as predictors in an enhanced autoregressive model, to test their forecasting accuracy using a benchmark autoregressive model of order one. The forecasting accuracy of the news was analyzed by running a battery of out of sample tests for the outcome variables volatility and trading volume for four of the most relevant Italian banks and the Italian stock index FTSE MIB. In a first phase both full-sample and out-of-sample predictive power were tested. The former approach however suffers from a possible look-ahead bias: the topics are obtained from an LDA performed on articles taken from the whole period, thus considering also articles appeared after the prediction date. In principle the topics obtained in this way could contain information about events in the future with respect of the day on which we want to make a forecast, thus introducing a possible look-ahead bias. Moreover the topics computed over the whole period T have another flaw: with a static description each topic has to describe some aspect of news appearing over the span of years. As such, these topics can only capture generic features, appearing in a large portion of the period T. Any "local" topic pattern would be lost with this approach, being them not persistent enough to be noticeable; this kind of information however is the most interesting for our goals, carrying the most predictive power for events in the short time range. For these reasons we focused on the results obtained with the out-of-sample method explained above in which

topics are defined using only a 3-month period preceding the prediction date.

In order to evaluate the forecasting power of the different selected topics we have compared the following regression models:

$$y_t = a + b\, y_{t-1} + c\, Sent_t + \sum_{j=1}^{K} d_j\, z_{j,t} \tag{7}$$

where the variables $z_{j,t}$ are the weights of each topic resulting from the LDA and $Sent_t$ is the average sentiment score for the day t. This equation has been estimated by the LASSO penalized regression [9, 10], in order to retain only variables with sufficient explanatory power. This method provides automatic variable selection, assigning a cost to any variable considered in the regression and thus forcing the selection of only the most useful ones. In the estimation step we fed the model with all the topic and sentiment variables, and then retained only the ones with predictive power as given by the LASSO regression. The forecasting accuracy of equation 7 has been compared with an AutoRegressive benchmark:

$$y_t = a + by_{t-1} \tag{8}$$

If any of the variables among sentiment and topics are retained in the regression with some degree of statistical significance, we conclude they are good predictors of stock market movements for the day following the news coverage.

The increase of forecasting accuracy between the benchmark autoregression and our model including topics and sentiment has been evaluated using three classical indices. Let $y_1, \ldots, y_T$ denote *actual* values, and $\hat{y}_1, \ldots, \hat{y}_T$ be the corresponding predictions:

1. Root Mean Square Error (RMSE)
$$\sqrt{\frac{1}{T}\sum_{t=1}^{T}(\hat{y}_t - y_t)^2}\,;$$

2. Mean Absolute Percentage Error (MAPE):
$$\frac{1}{T}\sum_{t=1}^{T}\left|\frac{\hat{y}_t - y_t}{y_t}\right|\,;$$

3. Mean Directional Accuracy (MDA):
$$\frac{1}{T}\sum_{t=1}^{T}[\![\mathbf{1}_{\text{sign}\,(y_t - y_{t-1}) = \text{sign}\,(\hat{y}_t - y_{t-1})}]\!]\,.$$

Additionally, we tested the hypothesis of equal forecasting accuracy between our models and the benchmark against the alternative of improved performance of the first with Diebold-Mariano (DM) test.

In table 1 and table 2 we show the previously mentioned statistical measures of the significance of the topics and sentiment coefficients in the regressions. As we can see the Diebold-Mariano test has always indicated a rejection of the null hypothesis of no difference between the forecasts produced by the AR model and those of the competing models which included the topic and the weighted sentiment:

Table 1: Comparison of forecasting accuracy for the Volume

	Topics	n-gram	MAPE	Rel. RMSE	MDA	DM
BMPS	4	1	45.97%	0.50	0.73	2.82 **
	5	1	46.18%	0.51	0.73	2.77 **
	3	2	48.22%	0.50	0.75	2.56 **
	4	2	46.40%	0.50	0.73	2.80 **
FTSE MIB	4	1	19.18%	0.76	0.87	3.93 ***
	5	1	19.10%	0.76	0.87	3.95 ***
	3	2	19.29%	0.76	0.86	4.06 ***
	4	2	19.60%	0.77	0.88	3.47 ***
ISP	4	1	26.25%	0.78	0.84	4.56 ***
	5	1	26.22%	0.78	0.84	4.55 ***
	3	2	27.27%	0.77	0.86	4.05 ***
	4	2	27.49%	0.81	0.87	3.41 ***
UBI	4	1	30.18%	0.72	0.85	3.65 ***
	5	1	29.57%	0.71	0.83	4.11 ***
	3	2	31.16%	0.71	0.85	3.55 ***
	4	2	30.35%	0.72	0.86	4.22 ***
UCG	4	1	28.23%	0.78	0.85	4.78 ***
	5	1	28.26%	0.79	0.86	4.01 ***
	3	2	29.21%	0.79	0.87	4.01 ***
	4	2	28.55%	0.79	0.86	4.22 ***

Table 2: Comparison of forecasting accuracy for the Volatility

	Topics	n-gram	MAPE	Rel. RMSE	MDA	DM
BMPS	4	1	41.35%	0.55	0.40	3.04 **
	5	1	41.93%	0.56	0.41	2.97 **
	3	2	44.68%	0.56	0.44	2.90 **
	4	2	44.28%	0.56	0.43	2.74 **
FTSE MIB	4	1	33.94%	0.79	0.42	6.35 ***
	5	1	33.95%	0.80	0.40	6.19 ***
	3	2	35.55%	0.80	0.41	5.87 ***
	4	2	36.20%	0.83	0.41	5.03 ***
ISP	4	1	35.19%	0.82	0.39	6.27 ***
	5	1	35.68%	0.83	0.41	5.56 ***
	3	2	37.59%	0.84	0.42	5.11 ***
	4	2	37.02%	0.83	0.40	5.23 ***
UBI	4	1	33.11%	0.74	0.44	6,47 ***
	5	1	33.98%	0.75	0.41	5.86 ***
	3	2	36.23%	0.75	0.40	5.43 ***
	4	2	35.02%	0.75	0.42	5.62 ***
UCG	4	1	34.90%	0.82	0.40	5.27 ***
	5	1	34.98%	0.84	0.40	4.73 ***
	3	2	35.72%	0.82	0.41	5.09 ***
	4	2	35.99%	0.82	0.40	4.73 ***

This means that in our model at least one among the topic and sentiment variables has a significant predictive power both for the volumes and volatilities over the period taken into consideration.

Another measure of the enhanced predictive power of our model over the AR benchmark are the *cumulative sum of squared error differences* (CSSED). They show how over the period considered the errors of the topic-sentiment model are consistenly lower than the ones of the benchmark model:

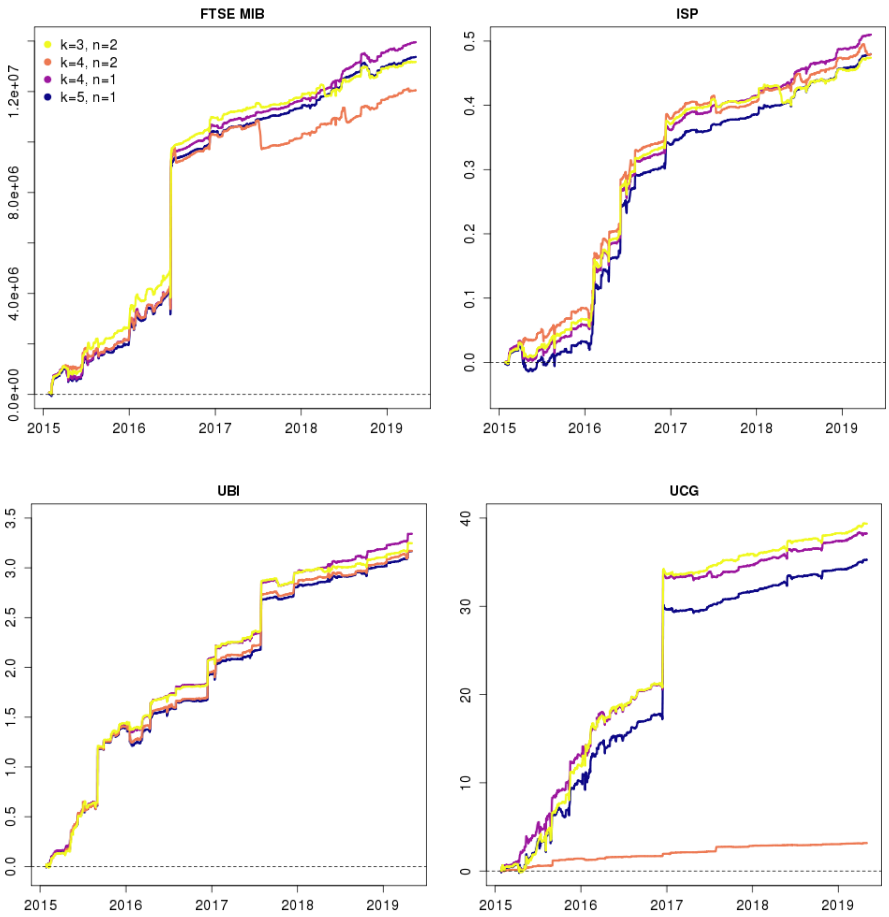


Figure 5.1: CSSED: Volatility

In the plots we show the CSSED for the volatility predictions for the class of models with $k = 3, 4$ and 5 as number of topics considered in the regression. The model with $n = 1$ is the one in which the tokens are single words, while in the model with $n = 2$ bigrams are considered in the topic definition. We show the results for the FTSE MIB index and three of the four banks studied [3]. As we can see the models comprising topic and sentiment variables always perform better than the benchmark on average, and performances are better in most of the sub-periods considered.

Finally, two of the most useful pieces of information we can extract from our model are the content of the topics with most predictive power, and the co-occurrences of variables in any given period. These information tell us what is really predicting the evolution of the target variables, allowing to make a connection between news content and market variables movements. As an example, in fig. 5.2, we show for the FTSE MIB index and the four banks considered a co-occurrence graph and a word-cloud with the content of the most predictive topic variable:

[3]The fourth was temporarily suspended from trading, hence the comparison is not available for the full period.

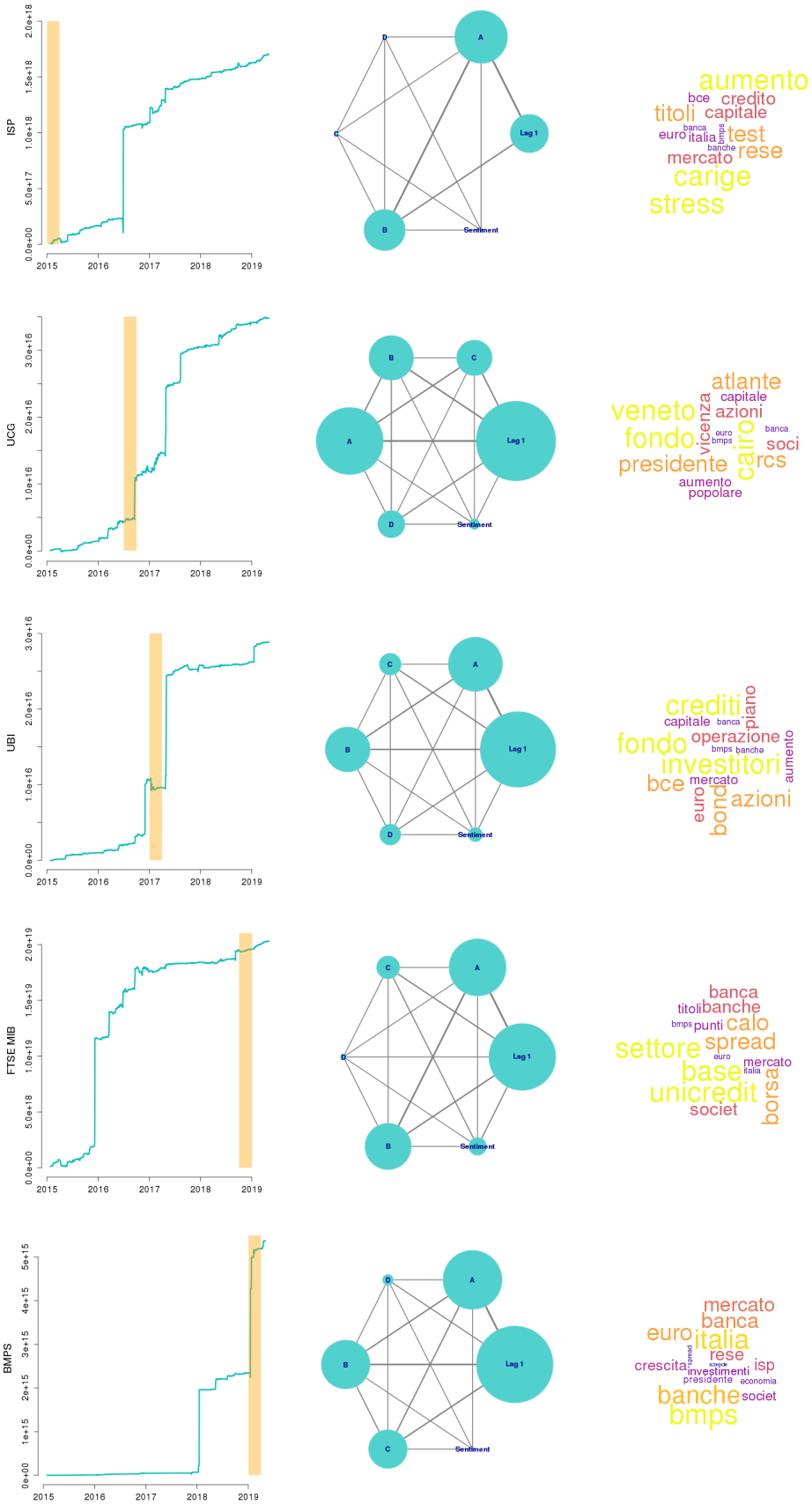


Figure 5.2: Variables co-occurrences and topics word-clouds

In the first column of the table the quarter in which the predictions are being made is highlighted, and the effect of the topic and sentiment variables on the predictive power is made clear by the CSSED. In the second column the variables with most predictive power and their connections are shown: the diameter of a circle is proportional to the number of days of the quarter in which the variable in question was retained by the LASSO, hence the number of days in which its presence was useful to make predictions. The thickness of the edges of the graph shows how often a couple of variables was retained in the model together: variables considered in the model often together are connected by wider links. We can see that though the autoregressive variable is kept in all the quarters showed, usually it comes together with other, news-related, variables.

Finally the third column of the table shows the words contained in the most predictive topic for that quarter, which allows us to check the relevant news content in the predictions. We can see that the model is able to pick and describe topics "trending" in the media during the periods considered. This confirms the fact that our model is able to exploit the most important news, as described by the sources considered, in order to perform better predictions.

6 Concluding Remarks

In this work we have shown how to extract relevant signals from textual data useful to forecast the main variables for the banking market in Italy.
We have compared the forecasting performance for the volatility, rate of return and exchanged volumes for four sistemically important banks and for the FTSE-MIB index for the Italian stock market. [4] In all the examples, we have adopted an AR model whose order has been selected on the base of an Information criteria.
For the volatility and the rate of returns we have systematically achieved improvements for the MAPE and the Relative RMSE. The improvement is significant, though not dominant for the MDA.
When choosing 4 topics, we have seen that on about 90% of the times the Lasso regression selects at least one topic as significantly relevant. The sentiment results significant around the 30% of the shown regressions.
Robustness of our results has been checked by running the Diebold-Mariano test which has always indicated a rejection of the null hypothesis of no difference between the forecasts produced by the AR model and those of the competing models which included the topic and the weighted sentiment.
We aim to extend our work to explore also the role of the sentiment on the relevant variables composing the banks' balance sheet. In addition the sentiment probed by our approach was extracted from articles appearing in specialized financial journals. Given the nature of the source the sentiment extracted is rarely strongly polarized. A more expressive sentiment variable could be obtained studying the public discourse regarding a given bank on social networks. This further direction is left for future studies.

[4] FTSE MIB is the benchmark stock market index for Italian national stock exchange. The index consists of the 40 most-traded stock classes on the exchange.

References

[1] Paul C. Tetlock. "Giving content to investor sentiment: The role of media in the stock market". In: *The Journal of finance* 62.3 (2007), pp. 1139–1168.

[2] Tim Loughran and Bill McDonald. "When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks". In: *The Journal of finance* 66.1 (2011), pp. 35–65.

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.", 2009.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[5] Minqing Hu and Bing Liu. "Mining and summarizing customer reviews". In: *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. 2004, pp. 168–177.

[6] Bo Pang, Lillian Lee, et al. "Opinion mining and sentiment analysis". In: *Foundations and Trends® in Information Retrieval* 2.1–2 (2008), pp. 1–135.

[7] Giuseppe Bruno et al. "The Sentiment Hidden in Italian Texts Through the Lens of A New Dictionary". In: *2nd International Conference on Advanced Research Methods and Analytics (CARMA 2018). Proceedings*. 2018.

[8] Rodrigo Agerri et al. "OpeNER: Open polarity enhanced named entity recognition". In: *Procesamiento del Lenguaje Natural* 51 (2013), pp. 215–218.

[9] Fadil Santosa and William W Symes. "Linear inversion of band-limited reflection seismograms". In: *SIAM Journal on Scientific and Statistical Computing* 7.4 (1986), pp. 1307–1330.

[10] Robert Tibshirani. "Regression shrinkage and selection via the lasso". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

News and banks' equities: do words have predictive power?

Valerio Astuti, Giuseppe Bruno, Sabina Marchetti & Juri Marcucci[1]

Bank of Italy

IFC-Bank of Italy workshop on "Data science in central banking" – Part 2: Data Science in Central Banking: Applications and tools

February 15, 2022

[1]The views expressed in this presentation are the authors' only and do not necessarily reflect those of the Bank of Italy.

Motivations and Main steps

Motivations

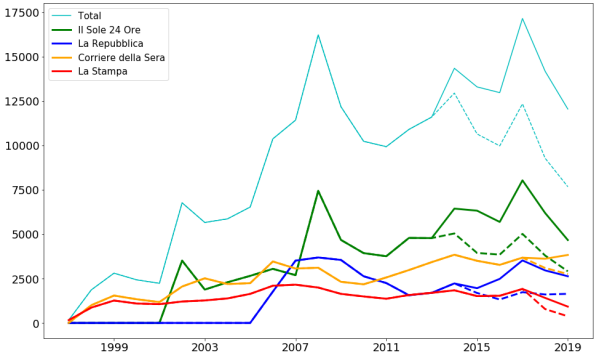
- Can we extract useful quantitative information from narrative content in newspaper?
- Are news a predictive factor in banks' equities trends?
- Is there any advantage in putting together text and classical balance sheet indicators?

Main Steps

- Starting point: building an archive of newspaper articles talking about the main Italian banks
- Pre-processing of articles, topic and sentiment analysis
- Application to predictive models for banks' trading volumes
- Memory-intensive tasks: Python and PySpark

Our Database

Number of articles per source



- From **Dow Jones Factiva** articles on the **100 most important Italian banks**
- *From September 1996 to May 2019* (article number not uniformly distributed over time)
- **Sources**: “Il Sole 24 Ore”, “La Stampa”, “La Repubblica”, “Corriere della Sera” plus online editions (long-dash lines!)
- **Corpus**: 217K articles, 100M words, 0.33M unique tokens (Zipf’s law)
- Case normalization, tokenization, stop-words removal, stemming
- Cut-off on minimum number of appearances of a term

Latent Dirichlet Allocation (LDA)

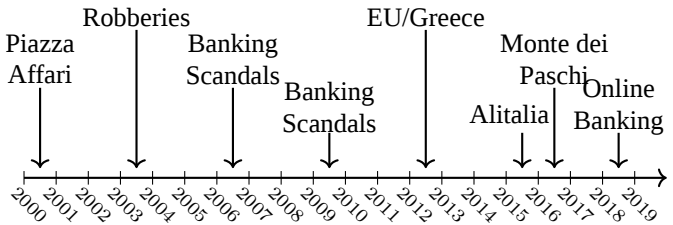
- Unsupervised, hierarchical probabilistic model to decompose a document in its most salient topics (probability distribution over words)
- Full sample (synchronic) and rolling subsample (diachronic): one defined over whole period, the other limited to three-year spans, rolling yearly. The number of topics is chosen to minimize perplexity
- Rolling sample was used to sidestep two possible problems: look-ahead bias and coarsening of topics over a 20-year span
- Full sample: minimum perplexity = 7.93, **number of topics = 15**
- Rolling sample: average perplexity = 7.83, **average number of topics = 8.75**

LDA Results: Main Topics $\longrightarrow$ Full sample vs Rolling subsamples

Main Topics in Full Sample

- Economy-Politics
- Investigations
- Industry
- Stock Exchange
- Local Activities
- English articles
- News Reports
- Financial Activities
- Italian Groups
- Public
- Balance/Capital
- Growth and Taxes
- Investments
- Stock Market Trends
- Boards

Main Topics in Rolling Sub-samples

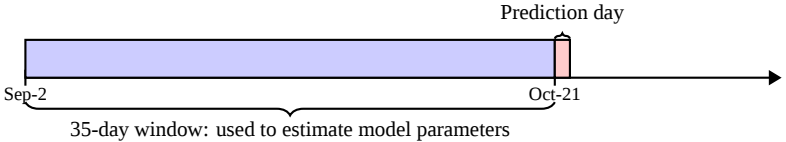


Model

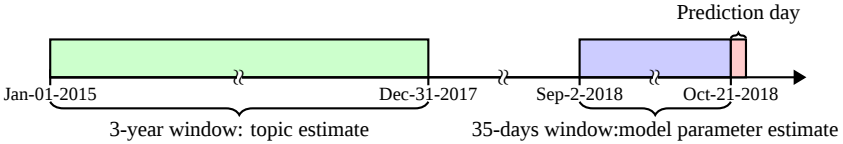
- Topics predictive power tested on stock market indices of 4 Italian banks and the Italian stock index FTSE MIB
- We analyzed returns, volatilities, and volumes. Volumes are the most reactive variables to news, and the ones topics forecast better
- Applied topic distributions with both static and rolling samples, with different results
- Our model is a LASSO with an adaptive number of topics k_t updated daily and the possibility to keep up to three lagged variables
- The benchmark model is an $AR\left(p_t\right)$ with p_t selected to minimize the BIC at each t

Topic distribution

- The number of variables used as predictors in a given day is selected by the LASSO methodology over the previous 35-day period;
- Possibility of look-ahead bias using topic estimated with future articles but on the other hand topics much coarser given the definition on a longer timespan;



- In the topic model with rolling sample the predictive variables are estimated over the 3 calendar years preceding the prediction day (weighted with the daily sentiment).
- Having defined the topics, the regression coefficients are estimated in the 35 days before the forecast.

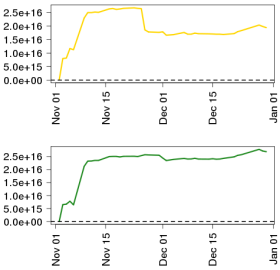


Out-of-sample performance: Cumulative Sum of Squared Error Differences (CSSED)

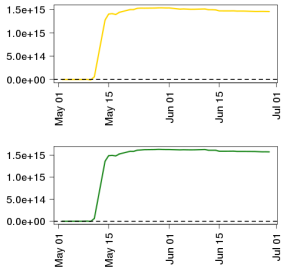
The performance of our models is evaluated through the difference between the CSSED of our models and the AR benchmark (if positive, our models are on average more accurate)

Upper panel → full sample
Lower panel → Rolling subsample

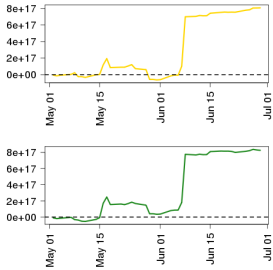
Bank 1, 2015



Bank 2, 2018



FTSE, 2018



Conclusions

- Topic analysis effectively captures relevant information content of newspaper articles
- Topics can be used as predictor variables for banks' equities
- Using the topics to make predictions, our models perform on average better than the AutoRegressive benchmark
- Tiny differences in terms of CSSED between using topics from the full sample and from rolling sub-samples

Thank You very much for Your Attention!