Central Bank Review 21 (2021) 141–153

Contents lists available at ScienceDirect

Central Bank Review

journal homepage: http://www.journals.elsevier.com/central-bank-review/

Can central bank speeches predict financial market turbulence? Evidence from an adaptive NLP sentiment index analysis using XGBoost machine learning technique

Anastasios Petropoulos, Vasilis Siakoulis*

Bank of Greece, Amerikis 3, Athens, 10564, Greece

ARTICLE INFO

Article history:
Received 4 January 2021
Received in revised form
1 December 2021
Accepted 3 December 2021
Available online 13 December 2021

JEL classification:
G01
G21
C53

Keywords:
Sentiment analysis
Machine learning
XGboost
Text mining

ABSTRACT

Central Bank speeches usually function as aggregators of internal quantitative and qualitative analysis of the institutions regarding the macro economy, the monetary policy and the health of the financial systems. Speeches usually function as a summary of the current status of a countries economic health, the undergoing trends and some future perspectives of the global economy. In this study departing from classical econometrics we employ natural language processing technologies in combination with machine learning techniques in order to filter out the most important signals in the corpus of speeches and translate into a sentiment index for forecasting the future financial markets behaviour. In our analysis, it is evident that central banker's expectations on economy tend to exhibit a predictive ability for financial markets turmoil. Using a combination of dictionaries which are either predefined or build based on historical speeches of the corpus we train an Extreme Gradient Boosting model that generates a sentiment index which signals turmoil with acceptable accuracy when passing a specific threshold.

1. Introduction

Central bank communication either through official monetary statement or through central bank speeches, promotes transparency on the main central banking operations while at the same time aims at increasing financial stability by informing the markets on economic topics. Through this channel, central bankers often convey their medium-term predictions about the global and the local economies by sharing their macro-economic projections. In the recent decade, several studies have provided empirical evidence that these types of direct and indirect central bank communications, have either an impact in the financial markets especially exchange rate and interest rates or can provide signals about the evolution of the stock market indices. .

Central bank speeches usually function as aggregators of internal quantitative and qualitative analysis performed by various internal divisions, such as economy research, monetary policy etc. Thus, central bank speeches usually include terms or phrases that convey significant messages regarding the evolution of the economy if filtered with the appropriate text mining tools. This is because central bank speeches are often general in nature, making use a broad statement and revealing less technical details about financial measures and indicators. They usual focus on the general trends and terms about future prospects of the health of the economies providing optimism for financial stability. To this end, to extract valuable information an advanced Natural Language Process (henceforth NLP) approach is necessary to filter out the noise and the vagueness in central banks statements included in the speeches.

In the recent years with the evolution of machine learning and the appearance of big data Natural Language Processing (NLP) is gaining ground as a tool for performing financial analysis. Following this trend, in this study we depart from a conventional macroeconomic analysis, which uses an input macro variables historical time series and replace it with a set of text documents, consisting of central bank speeches. Although a number of studies are concentrated on specific jurisdictions, in our analysis we aim to

* Corresponding author.
E-mail addresses: apetropoulos@bankofgreece.gr (A. Petropoulos), Vsiakoulis@bankofgreece.gr (V. Siakoulis).
Peer review under responsibility of the Central Bank of the Republic of Turkey.

https://doi.org/10.1016/j.cbrev.2021.12.002

create a global sentiment index aggregating speeches from all central banks around the world, providing a global surveillance tool for monitoring the evolution of the global economy. Our approach provides further evidence of the effectiveness of oral intervention by providing cross-validation through the simultaneous examination of all countries' central banks. In this way, our analysis provides a holistic view combining all countries' economic specificities, expressed through their respective central bank communications, into a global sentiment index aggregating economic trends across all jurisdictions.

Our contribution to the literature about central banks communications and their connection to the evolution of financial markets is the following. At first, we provide a methodological contribution by directly quantifying the information in communication using NLP and machine learning. Secondly, we provide evidence of the impact of speeches by the central banks across a set of different jurisdictions. Third from a technical perspective, we build a pipeline of models and techniques that fully automate the process of text data cleansing, filter speeches for relevance and provide an adaptive dictionary terms in an automatic way by using machine-learning models that assign relevance and sentiment score. Finally, though our analysis, we provide several sub sentiment indices for capturing several aspects of speech using different dictionaries, and combine them in a global sentiment index using a set of different Machine Learning and comparing across them in order to find the best one.

The remainder structure of the paper is the following:. In Chapter. 2, we review the literature on central bank communication with a particular emphasis on the research that focuses on the connection with financial markets and the application of NLP methods for text analysis. Chapter 3 describes the data used whereas Chapter 4 provides some analytical background on the machine learning techniques, which are employed for combining signals into a sentiment index. Chapter 5 provides the steps of the methodology employed and the final results are presented on Chapter 6. These results focus on the predictive ability of the NLP machine learning approach employed and performance measured both using in sample and out sample data. Chapter 7 concludes the paper and gives recommendations for further research.

2. Literature review

Previous research studies have focus on the relation between Central bank communication and the level of exchange rates and the term structure of interest rates. Furthermore, in recent years academia explored also the connection of official monetary statements with the stock market performance (Bennani 2019, Sapphasak 2019). To this end, numerous studies have performed text-mining analysis in order to define trends in the official central bank's communications and the respective financial market variables.

Bruno and Giuseppe (2016) reviewed some of the main methodologies employed in text mining and for the extraction of sentiment and emotions from textual sources and provided an empirical application on Bank of Italy Governors concluding remarks from 1996 to 2015. Their results show the feasibility in extracting the main topics from the considered corpus. Moreover, they show how to check for positive and negative terms in order to gauge the polarity of statements and whole documents. Hubert and Labondace (2017) explored empirically the theoretical prediction that waves of optimism or pessimism conveyed by ECB and Federal Open Market Committee (FOMC) policymakers in their statements affect the term structure of private short-term interest rate expectations. They quantified central bank tone using a computational linguistics approach and by identifying sentiment as exogenous shocks, they estimated the impact of sentiment on private agents' expectations about future short-term interest rates. Their main finding was that sentiment shocks increase private interest rate expectations around maturities of one and two years in a non-linear pattern.

Maqsood et al. (2020) explored the effect of different major events occurred during 2012–2016 on stock markets by using the Twitter dataset to calculate the sentiment analysis for each of these events. Their results indicate that performance improves by using the sentiment component in the analysis. Bennani (2019) tests whether the People's Bank of China's communication affects expectations of market participants and matters as a monetary policy instrument relying also on a computational linguistic tool to measure the tone of PBC speeches and second and using a high frequency methodology to estimate the effect of tone on stock prices. Their results show that positive changes of the tone affect stock prices in the Shanghai and the Shenzhen stocks markets. Su et al. (2019) also investigated to what extent that the People Bank of China's (PBC) communication influences the Chinese money market, finding that the PBC's communication has a significant effect on the country's money market.

Mathur and Rajeswari (2019) quantitatively analyzed monetary policy statements of the Reserve Bank of India (RBI) from 1998 to 2017. Using natural language processing tools, they constructed measures of linguistic and structural complexity that capture governor-specific trends in communication. They found that lengthier and less readable statements are linked to both higher trading volumes and higher returns volatility in the equity markets, though the effects are not persistent.

Apel et al. (2019) focused on the information content of the Federal Open Market Committee minutes and transcripts by performing deep transfer learning, a technique that involves training a deep learning model on one set of documents — U.S. congressional debates — and then making predictions on another. Their findings suggest that transcripts are more informative than minutes and heightened committee agreement typically precedes policy rate increases. Kahveci and Aysun (2016) examine the monetary policy statements of three central banks in the pre-crisis years of 2002 to the subsequent period of 2015 and compare the specifics of the language used both pre- and post-crisis. Their results convey that optimistic tone of Fed has decreased over the crisis period we studied while certainty tone has increased. Finally Masawi et al. (2018) found evidence that the Bank of Canada's (BOC's) speeches reduce the mean exchange rate returns whereas Shapiro et al. (2019) used text analysis to estimate central bank objectives focusing on Federal Open Market Committee (FOMC) preferences.

In this study, following a different venue we explore the predictability of central bank speeches using NLP sentiment analysis combined with machine learning algorithms. Although Natural Language processing has been applied already in other studies in specific jurisdictions, our analysis is based on a corpus of documents that extend to all jurisdictions globally. Our main innovation lie in the use of advance Machine Learning techniques in order to provide an adaptive automatic way to create and update sentiment dictionaries and combine the signals provided by different

dictionaries into an overall sentiment index. We afterwards use this index to forecast crisis events defined as S&P falls by more than 8% in the following three months period.

3. Data collection

Our dataset includes data from 2008 to 2019, a period that includes a series of financial market turbulence events including the recent financial crisis of the Lehman Brothers collapse. It covers also the time that Central Banks changed significantly their monetary policy and became more transparent in their communication to the public.

Our sentiment index analysis is based on a corpus of central bank speeches collected from the official BIS website, ranging from 2008 to 2019. The final dataset is comprised of more than 10000 speeches from central banks officials. Their content is rather broad: referring to financial and economic themes or internal issues that involve aspects of the respective central bank entity (Insurance schemes, organization issues) or even speeches that are performed in universities. The database also includes the date of performance, the speaker, the central bank that the speaker is employed, the length as measured by the number of pages and the title of the speech. These static data are used to differentiate between different dictionaries accounting also for the length of the speeches and the central bank origin. The construction of the dictionaries is described in detail below.

The database collected is split into two subsets for development purposes: Speeches spanning from 2008 to 2014 constitute the train sample, while speeches from 2015 to 2019 belong to the test sample. The test sample is used to test the generalization of the proposed sentiment index framework. Some descriptive attributes of the Central Bank speeches collected (Appendix) show that our dataset has global jurisdiction coverage whereas Fig. 1 confirms that speeches numbers are almost uniformly distributed across all the analysis period.

The dataset is enriched with historical time series of S&P 500 and VIX volatility index used for both building sentiment dictionaries and evaluating the predicting performance of our model. The forecast horizon for the sentiment index is set to 3 months. We define a threshold of 8% commutative 3-month drop in the S&P stock market index to characterize an upcoming event of increased market turbulence. In Table 1, we present some additional descriptive characteristics of the Corpus. Specifically it includes the number of relevant speeches by year, the average number of pages by year, the average sentiment using the generally accepted dictionaries in the literature (namely AFINN, NRC, BING and LOUGHRAN) and the 3-month change of the S&P500 and VIX. In all, the sentiment scoring exhibits an increase in optimism by year with exception of the first two years in the sample and 2015.

4. Machine learning techniques employed - background

In this study, we depart from previous studies, which focus on traditional econometric theory, by examining the complementary use of Machine Learning methods in two different domains. Firstly, the Extreme Gradient Boosting algorithm is employed in the creation of a self-evolving dictionary based on set of words (1000) qualitatively allocated a positive or negative meaning and by try to predict this meaning following a set of predefined word criteria (e.g. LDA score, syntax type). The dictionary is self-evolving in the sense that if a new word appears it will be scored as positive or negative based on the XGBoost parameterization. In a second stage, we employed a set of Machine Learning Algorithms in order to combine the signals provided by different dictionaries into an overall sentiment index, useful for the prediction of financial crisis events. The Machine Learning Techniques we employ are Random Forests, Extreme Gradient Boosting (XGBoost), Support Vector Machines and Deep Neural Networks.

Random Forests are an algorithm based on the random generation of a number of classification trees, which is the so-called Forest. Tree generation is randomly performed in an iterative mode so in each iteration, a random subsample of the included features is selected from the dataset by means of bootstrap. Then, a tree is generated from using the CART algorithm, which contains a relatively limited number of features. After constructing the random trees, prediction is performed using Bagging. Each input enters through each decision tree in the forest and produces a forecast. Then, all predictions of each tree are aggregated as either a (weighted) average or majority vote, depending whether the underlying problem is a regression or a classification, respectively.

Random Forests (RF) have recently received considerable attention in various financial research fields (Breiman, 2001; Breiman et al., 1984). Our implementation of RFs is based on the randomForest package of R. To perform optimal selection of the maximum number of trees in the forest, we perform cross-validation using the available validation set; we optimize the RF

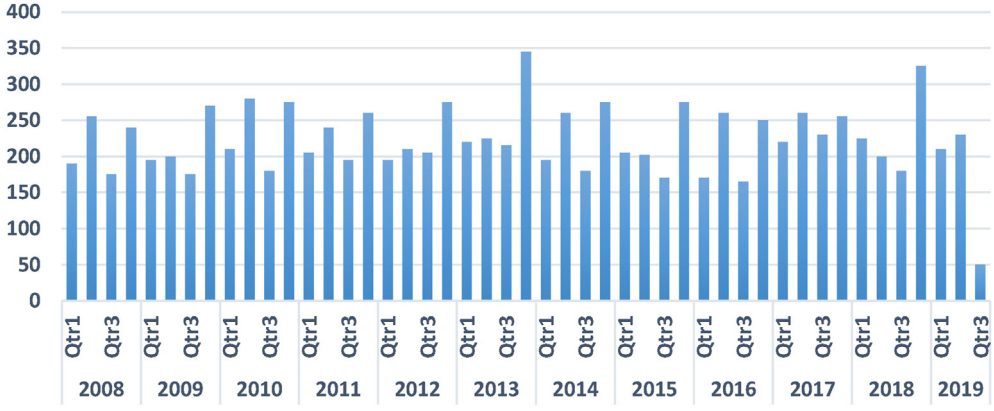

Fig. 1. Distribution of number of speeches include in the sample across the analysis period.

Table 1
Statistics of speeches across the analysis period by year.

Year	No of Speeches	No Pages	AFINN	NRC	BING	LOUGHRAN	3 Month Return S&P 500	3 Month Change in VIX
2008	682	6.2	37.4	107.7	7.5	−24.9	−11%	5%
2009	713	6.4	2.8	91.7	−2.8	−43	8%	−20%
2010	763	6.5	11.7	96.4	4.1	−36.3	3%	−1%
2011	808	6.9	11.8	98.4	2.3	−36.6	1%	0%
2012	734	6.9	14.4	101.5	3.4	−36.8	3%	−6%
2013	808	6.5	25.4	104.1	11.1	−28.3	5%	3%
2014	764	6.4	34.8	106.8	12.3	−25.5	3%	4%
2015	700	6.5	32.9	96.1	7.8	−27.8	−2%	9%
2016	686	6.7	36	97.5	9.2	−26.2	5%	−15%
2017	762	8.3	41	110.8	16.6	−20.4	4%	9%
2018	718	8.6	41	109.9	16.1	−18.5	0%	−2%
2019	371	8.3	37.3	108.1	13.1	−18.1	3%	3%

using various setups in number of trees and the number of random variables to consider for node split as well the number of minimum observations in the leaf nodes.

The XGBoost (Chen, Tianqi, 2016) is a boosting tree algorithm that is an enhancement over tree bagging methodologies, such as Random Forests (Breiman, 2000), which have gained significant ground and are frequently used in many machine learning applications across various fields of the academic community. The basic philosophy of bagging is based on combining three concepts: I) Creation of multiple datasets; II) building of multiple trees and III) bootstrap aggregation or bagging. It adopts a divide-and-conquer approach to capture non-linearities in the data and perform pattern recognition. Its core principle is that a group of “weak learners” combined, can form a “strong predictor” model.

Gradient Boosting trees model is proposed by Friedman (2001) and has the advantage of reducing both variance and bias. It reduces variance because multiple models are used (bagging), whereas it additionally reduces bias in training the subsequent model by telling him what errors the previous models made (boosting). In gradient boosting, each subsequent model is trained using the residuals (the difference between the predicted and true values) of previous models. XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithm, offering increased efficiency, accuracy and scalability over simple bagging algorithms. It supports fitting various kinds of objective functions, including regression, classification and ranking. XGBoost offers increased flexibility, since optimization is performed on an extended set of hyper-parameters, while it fully supports online training.

We developed Random Forest and XGBoost in the context of our study by utilizing the randomforest and XGBoost R package. We performed an extensive cross-validation procedure to select a series of entailed hyper-parameters, including the maximum depth of trees generated, the minimum leaf nodes size to perform a split, and the size of sub-sampling for building the classification trees and the variables considered in each split. The objective function used for the current problem was logistic due to the binary nature of the dependent variable while the area under the curve (AUROC) metric was used for model selection in the context of cross-validation. The AUROC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance. In practice, the value of AUROC varies between 0.5 and 1, with a value above 0.8 denoting a very good performance of the algorithm. To reduce overfitting tendencies, we tuned the γ hyper parameter, which controls model complexity by imposing the requirements that node splits should yield a minimum reduction in the loss function, as well as the α and λ hyper parameters, which perform regularization of model weights similar to shrinkage techniques such as LASSO.

Boosting and Bagging algorithms, even though they are computation intensive, have the relative advantage that they are not “black boxes” regarding the factors affecting the result, since they provide a module for calculating variable importance measures through reshuffling. In other words after predicting with the benchmark model the reshuffling technique predicts hundreds of times for each variable in the model while randomizing that variable. If the variable randomized negatively affects the model's benchmark score, then it ranks high as an important variable.

Finally, for benchmarking further the XGboost model performance we applied also two complementary algorithms namely Support Vector Machines and Neural Networks. We evaluate soft-margin SVM (Vapnik, 1998) classifiers using linear, radial basis function (RBF), polynomial, and sigmoid kernels, and retain the model configuration yielding optimal performance. For selecting the proper kernel and hyper-parameters of the evaluated kernels as well as the cost hyper parameter of the SVM we exploit the available validation set. For fine-tuning the parameters of the final SVM model, a grid search was performed with various combinations of values for the cost and gamma hyper parameters. We implemented this model in R using the kernlab package along with the grid-search functionality included in the e1071 package. The final SVM selected had a Radial Basis "Gaussian" kernel.

We explored also the use of multilayer feedforward deep neural network (Goodfellow, 2016) in order to classify the sentiment indices signals and predict severe financial shocks. Deep learning applications in the domain of finance is rather limited but it has been an active field of research in the recent years, as it has achieved significant breakthroughs in the fields of computer vision and language understanding. Deep Neural Networks are built based on nonlinear activation functions typical choices of which are the logistic sigmoid, hyperbolic tangent, and rectified linear unit (ReLU). The activation layers increase the ability and flexibility of a DNN to capture non-linear relationships in the training dataset but on the other hand, the huge number of trainable parameters could lead to overfitting. Therefore, the use of simple, effective, and efficient regularization techniques is necessary to avoid poor out of sample performance. Dropout is the most popular regularization technique for DNNs. We postulated deep networks that are up to five-hidden layers deep and comprise various numbers of neurons. We perform cross-validation to select the number of hidden layers and their component-hidden neurons, exploiting the available validation set. Training and optimization of the deep neural networks was performed in Python using TensorFlow package.

5. NLP model development

The steps we implement on the corpus of speeches are shown in

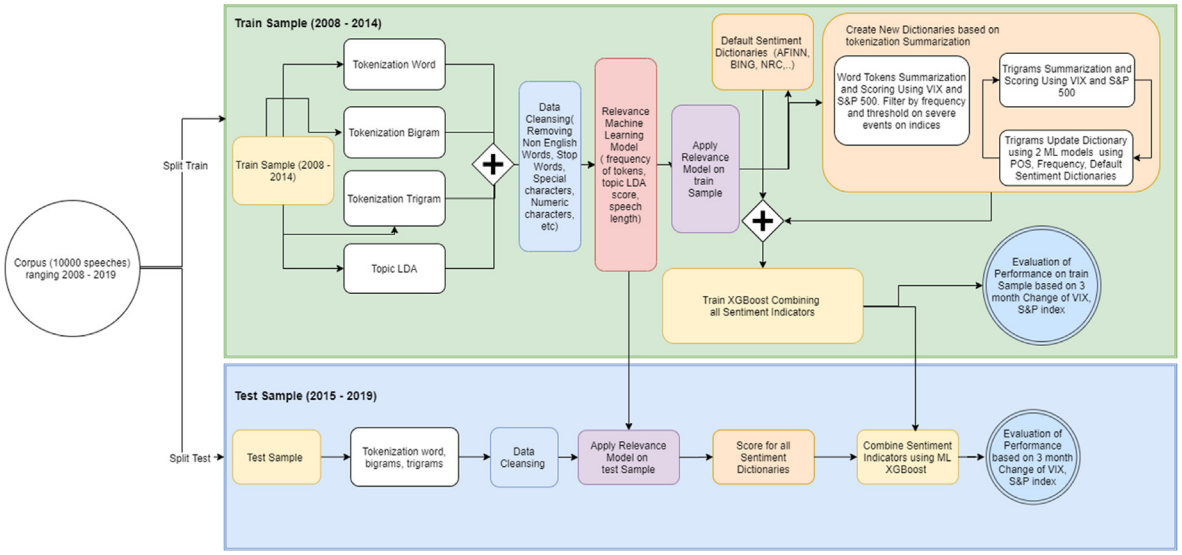

Fig. 2. Overview of the NLP processor for the creation of a Global Sentiment Index.

Fig. 2. In the first stage, we split the dataset between train and test. The train dataset is composed of all speeches that dated between 2008 and 2014. All the sentiment dictionaries developed are based solely on the train sample. The test sample is used in the last stage for performance evaluation when combining the signals we receive from the dictionaries into a sentiment indicator.[1] . The NLP processor for the creation of a Global Sentiment Index and a financial crisis Early Warning system is segregated into three discreet steps.

- Step 1: Tokenization
- Step 2: Relevance topic filter Modelling
- Step 3: Creation of dictionaries to perform individual sentiment analysis.
- Step 4: Scoring of speeches and combination of all individual indicators in a Global Sentiment Index used as an input in an Early Warning System for distress events defined in binary form as « S&P falls by more than 8% in the three month period ahead».

5.1. Step 1: Tokenization of central bank speeches

The first step in the textual analysis of corpus (set of documents) is the Tokenization. Under the tokenization process, each document speech is mapped to its “ingredients” like words, bigrams, and trigrams for further NLP analysis. In our analysis, we focus on monograms, bigrams and trigrams since the last play an important role, as they refer to phrases where sentiment is clearer in the context, they used in the speeches.

As expected, speeches are free text documents, without any structure thus extensive data cleansing is performed on a next step to remove noise. Thus all tokens were filtered in order to remove words that do not belong in the English dictionary. In addition, under this stage we remove special characters, symbols, alpha-numeric characters, numeric characters, urls, emails and most importantly stop words like: and, or, the etc. which tend to not curry significant meaning of the relevant speech. In a final stage, tokens are filtered for names, like the speaker's surname and words in the header and footer, which tend to appear repeatedly without any sentiment meaning.

Tokenization is performed for supporting both topic relevance modelling (Step 2) and creation of dictionaries containing the most frequent monograms, bigrams, and trigrams (Step 3).

5.2. Step 2: relevance topic filter modelling

Before proceeding to sentiment analysis, we build a model for automatic filtering relevant speeches with financial and economic content, as irrelevant speeches would impede the forecasting ability of a financial crisis early warning model. For succeeding in scoring its document for relevance we combine several indicators based on word frequencies and topic modelling probabilities. More specifically, we employ.

- **Frequency scoring**: For words (monograms) in order to estimate a more reliable frequency we apply first the statistic term frequency–inverse document frequency (TF-IDF). Based on that we measure how important a word is to a document in a collection in order to exclude unimportant words that usually increase the noise in text topic classification. Regarding bigrams and trigrams we estimate the log(frequency), based on which we allocate a relevance importance indicator.
- **Topic modelling**: We use Latent Dirichlet Allocation (LDA) to calculate the probability for each word to belong to two different topics recognized. The higher the probability to belong in a topic the higher the relevance.

Logistic Regression model: Using the inputs from frequency scoring and topic modelling we build a regression model for relevance measurement. Subsequently, we score each speech based on its content. Speeches that relate to macro economy, monetary policy and global financial outlook score higher against speeches that were given in other occasions. The following paragraphs provide the technical details of the relevance topic modelling.

5.2.1. Frequency scoring

During the previous step all speeches/documents were decomposed in monograms bigrams and trigrams. In order to gain

[1] Text analysis was performed using the tidyr package in r.

an importance of each term we estimate their frequency in the in-sample of the corpus. For example, for a specific trigram e.g. "financial market turbulence" we count the time it appears across all speeches. The more time it appears the more relevant to the whole corpus content the term is. Subsequently to a speech that contains a high frequent trigram, a higher score for relevance is allocated.

Especially for monograms words in order to estimate a more reliable frequency we apply the statistic TF-IDF which is intended to measure how important a word is to a document in a collection (or corpus) of documents, in our case in a specific speech across the whole collection of central bank speeches. It is a rule-of-thumb or heuristic quantity; while it has proved useful in text mining, search engines, etc. The formula for the estimation of TF-IDF statistic is given below: It is a combination of two measures, the Term Frequency (tf) and the Inverse Document Frequency (idf). The term frequency is a count of the occurrence of term t in document d.The inverse document frequency is the natural log of the total number of documents (N) divided by the number of documents containing the term t. The TF-IDF is the multiplication of the two measures.

$tf(t,d) = f_{t,d}$, number of occurrences of word t in document d of Corpus D.

$idf(t,D) = \log \frac{N}{|\{d \in D : t \in d\}|}$ Where N is the number of documents in the corpus D. and the denominator is the number of documents that contain word t in the corpus.

$$tf - idf(t,d,D) = tf(t,d) \cdot idf(t,D)$$

We use tidytext package in R to approach TF-IDF analysis and use consistent, effective tools to quantify how important various terms are in a document that is part of a collection. Since we aim to produce a global relevance indicator for monograms, we exclude words, which are not globally important base on the TF-IDF indicator. For example, terms that are not important for any document are excluded. We apply the tf-idf filter in order to exclude words that tend to exhibit high frequencies in a corpus and are not the usual stop words, for example verbs. The final one-word indicator for relevance produced is given by the log(frequency) in the remaining monograms after we apply the TF-IDF filter thus allocating higher weight to terms which are globally important.

Regarding trigrams and bigrams we estimate the log(frequency) based on which is allocated a relevance importance indicator. In the case of trigrams and bigrams there it is not necessary to apply first the TF-IDF filter since by definition the frequency distribution is more skewed to terms that are highly relevant to banking and macro economy. In order to allocate a relevance score to a document based on the bigrams or trigrams we sum the score of each bigram included in each speech. It is evident that the higher the score it means that the particular speech contains a significant amount of important bigrams making it more probable to be relevant for inclusion in the analysis.

5.2.2. Topic modelling

In text mining, we often have collections of documents, such as blog posts or news articles that we divide into natural groups so that we can understand them separately. Topic modelling is a method for unsupervised classification of such documents, similar to clustering on numeric data, which finds natural groups of items even when we are not sure what we are looking for. LDA is a particularly popular method for fitting a topic model. It treats each document as a mixture of topics, and each topic as a mixture of words. This allows documents to "overlap" each other in terms of content, rather than being separated into discrete groups, in a way that mirrors typical use of natural language.

Thus, to boost relevance, we train a topic LDA model to define the topics discussed in the whole corpus of speeches included in the train sample. We implemented the algorithm using the package "topicmodels" in R. Based on the implementation of the algorithm two topics are recognized (economic and monetary policy, financials and central banking) and based on the terms allocated to each. The output of the LDA fitted model is a probability of each word to belong each of the topics recognized. From a relevance perspective since both topics are relevant for our analysis, the higher the probability to belong in topic A or B the higher the relevance. Document/speeches that are less condensed in high probability terms exhibit overall low probability to be relevant for sentiment analysis.

5.2.3. Regression model

In this stage we build, a relevance model where each speech receives a score based on its content. Speeches that relate to macro economy, monetary policy, Global financial outlook receive a higher score. Irrelevant speeches (low score) are removed from the corpus. We develop the relevance model using the topic modelling and frequency of tokens (word, bigrams and trigrams) as described in the previous sections. Specifically, after scoring each speech based on the frequency of monograms, bigrams and trigrams and the topic LDA model we produce five indicators for relevance scoring. To fit the model we qualitatively create sub group of 200 speeches which we characterize with a binary index as relevant or not. This is the dependent variable for combining all relevance indicators by employing logistic regression.

Before we proceed with the dictionary construction analysis, we apply the relevance model and based on a threshold of 50% probability we remove speeches from the in-sample database. This led to an exclusion of approximate 800 speeches from further sentiment analysis as irrelevant.

5.3. Step 3: creation of dictionaries to perform individual sentiment analysis

We build a series of tailor-made sentiment dictionaries combining tokens with the respective 3 months-ahead performance of S&P and VIX. Each document is mapped to the three-month percentage ahead of VIX and S&P based on the date is was delivered. Based on the trigram and word tokenization we keep the most frequent. The monograms that are included in the dictionaries are the ones produced in step 1 after we apply a threshold of 10 in frequency. For trigrams the threshold applied for frequency is set to 30.

In order to score each token in the dictionaries, first we map each speech to the 3 month ahead S&P 500 and VIX percentage change. Then we trace the speeches where the specific token is included and estimate its score based on the average performance of the S&P returns and VIX movements for this sub population of speeches.

For example for "upcoming financial crisis" trigram appearing in speeches in 12 dates (t1 to t12) the average score based on S&P returns equals average (S&Pt1+3m, …., S&Pt12+3m).

1. Besides simple average, we employ several aggregating methods described by name in Appendix Table 2. We then use

these methods to estimate different variants of dictionaries and their respective scores. This process increases the features available for the multivariate classification model developed within the proposed framework. The methods applied take into account the average return of S&P 500 of each term as well as the frequency of each term. For example, terms that are allocated with an average change in S&P 500 close to zero are excluded for some features. Furthermore, we define new features by applying extra filter in the frequency of the terms. Specifically below we outline the methods implemented in our study: Simple average without any threshold applied
   - o 2 (S&P or VIX) * 2 (Monograms or Trigrams) in total 4 dictionaries.
2. We discretize the change of the index thus if S&P fell (rose) we allocate a −1 (+1) to the token for that speech, so the averaging is performed on the binary defined variables.
   - o 2 (S&P or VIX) * 2 (Monograms or Trigrams) in total 4 dictionaries.
3. Simple average but we apply two different thresholds thus if the change of the index is not significant we ignore it in the averaging
   - o 2 (S&P or VIX) * 2 (Monograms or Trigrams)* 2 Thresholds in total 8 dictionaries.
4. Simple average discretized, as in 2) but we apply a threshold as in 3)
   - o 2 (S&P or VIX) * 2 (Monograms or Trigrams)* 2 Thresholds in total 8 dictionaries.
5. Sum or division of the score for VIX and S&P dictionaries creating dictionaries that allocate high positive or negative score when both VIX and S&P exhibit outlier behaviour.
   - o In total 2 dictionaries.
6. Simple average with threshold in monogram frequency for S&P and VIX (2 vocabularies)
   - o In total 2 dictionaries.
7. Simple average discretized (−1,1) with threshold in monogram frequency for S&P and VIX (2 vocabularies)
   - o In total 2 dictionaries.

In addition, we use also four predefined dictionaries broadly used for sentiment analysis in literature. The general-purpose lexicons are.

- AFINN from Finn Arup Nielsen,
- Bing from Bing Liu and collaborators
- NRC from Saif Mohammad and Peter Turney
- Loughran from Loughran and McDonald

The first 3three lexicons include on monograms, i.e., single words. These lexicons contain many English words and the words are assigned scores for positive/negative sentiment and possibly emotions like joy, anger, sadness, and so forth. The NRC lexicon categorizes words in a binary fashion ("yes"/"no") into categories of positive, negative, anger, anticipation, disgust, fear, joy, sadness, surprise, and trust. The Bing lexicon categorizes words in a binary fashion into positive and negative categories. The afinn lexicon assigns words with a score that runs between −5 and 5, with negative scores indicating negative sentiment and positive scores indicating positive sentiment.

We also employ the Loughran and McDonald dictionary of financial sentiment terms (Loughran and McDonald 2011). This dictionary was developed based on analyses of financial reports, and intentionally avoids words like "share" and "fool", as well as subtler terms like "liability" and "risk" that may not have a negative meaning in a financial context. The Loughran data divides words into six sentiments: "positive", "negative", "litigious", "uncertain", "constraining", and "superfluous". Using the above dictionaries, in each speech is allocated a score based on the sum of matching terms, adding or counting all positive and negative terms included.

The list of trigrams dictionaries constructed is enriched based on a qualitative dictionary defined using the trigrams tokenization from the previous steps. In particular using expert judgment we filtered 1000 trigrams where we allocated +1/-1 dependent of whether the phrase is considered either optimistic/pessimistic in a financial/economic meaning. This way we have developed a custom made own dictionary (marked as trib_sent) based on a qualitative assessment of the trigrams.

To construct the adaptive nature of this dictionary we combine the use of two trained XGBoost models. Thus we produce an efficient way to enrich the dictionary without the need for further application of expert judgement when new speeches are available to the corpus. To achieve this, we train an XGBoost to apply a probability of importance/relevance in each new trigram (relevance XGBoost) and a secondary XGBoost that assigns a positive or negative meaning to the filter trigrams (sentiment XGBoost). The two models are applied in a serial structure to enrich the qualitative defined dictionary. The dataset used for training the XGBoost is comprised of the 1000 labelled trigrams that where defined with expert judgment in the previous step along with the 2000 thousand trigrams that were excluded. Note that the trigrams universe with frequency >30 was approximately 3000 produced by the tokenization process. The dependent variable for the first XGBoost is takes the value of 1 if included in the dictionary or 0 otherwise. The second XGBoost is trained on the trigrams selected (1000) based on the positive or negative score allocated to them. To build the XGBoost model a series of candidate variables/features were constructed:

- The frequency of the trigram
- The frequency of each word in the trigram
- The topic LDA score of each score in the trigram
- The default dictionaries score of each word in the trigram based on the default sentiment dictionaries (AFFINN, NRC, BING, and LOUGHRAN). For example in the trigram "global financial crisis" if we score the individual words based on LOUGHRAN dictionary we get the following three variables (NA, NA, Negative) since the terms "global" and "financial" are not included in the dictionaries since their meaning is irrelevant to sentiment.
- POS tagging each word separately in the trigram: POS tagging is the process of marking up a word in a corpus to a corresponding part of a speech tag, based on its context and definition. This task is not straightforward, as a particular word may have a different part of speech based on the context in which the word is used. To perform this POS tagging process we apply the UDpipe package in R.

After training the relevance XGBoost we apply it on the trigrams that where initially excluded from our analysis i.e the trigrams with frequency less than 30 times appearance in all speeches. By applying a threshold of 80%, we select additional trigrams relevant for our dictionary with 100% accuracy. In a latter step, we apply the sentiment XGBoost to apply −1/1 to the new generated trigrams. Our experimental results offer 98% accuracy. This way we succeed to minimize the use of expert judgment for the definition of the

trigram dictionaries in our sentiment analysis. Furthermore, we offer an automatic way to adapt the group of trigrams as new speeches are introduced to the corpus.

The 35 dictionaries scores based on the abovementioned techniques where further adjusted in three different ways mentioned below so we finally obtain 105 final vocabularies in total.

- Score is divided by number of pages in document (indexed as _p)
- Double weight allocated to score for the major Central Banks like FED (indexed as _cb)
- Score adjusted for both number of pages and size of central bank (indexed as cbp)

Using each dictionary and its corresponding token-score mapping we succeed to create numerical features from text for multivariate analysis. This is achieved by aggregating the score of all tokens located in a specific speech for each dictionary.

5.4. Step 4: scoring of speeches and combination of all individual indicators in a global sentiment index

In Step 5.3, we defined the group of dictionaries to perform individual sentiment analysis and scored speeches with the dictionaries created. This led to the allocation 105 scores for each document in the corpus. Using this dataset, we apply a set of Machine Learning algorithms in order combine all individual indicators in a Global sentiment index.

Before training the algorithm each dictionary – speech score is normalized in order to take values in (0,1). Then we aggregate each sentiment using a simple moving average monthly window taking into account all speeches given in the specific period. Based on the monthly average sentiment we estimate the percentage change of each sentiment in a one-month period. Then we define a binary variable based on expert judgment to designate financial market turbulence. This is set to one if S&P falls by more than 8% in the three-month period ahead or zero otherwise corresponding to a 10% percentile of the stock index returns in the sample (Fig. 3). To train the models we split the in sample into training and validation sample and apply parameter tuning to optimize the validation sample AUROC applying a 70%/30% rule respectively. Then we use the respective test sample (2015–2019) to estimate the performance of the models in a new dataset.

The machine learning models employed include an Extreme Gradient Boosting Algorithm (XGBoost), a Deep Neural Network (DNN), a Support Vector Machine (SVM) and a Random Forest (RF).

For the development of XGBoost and Deep Neural Networks we followed best practice regularization techniques to avoid overfitting. In the structure setup for XGBoost we apply appropriate values for the regularization parameters (lamda and alpha) as well as we limit the depth of each tree to 5. For the deep neural networks we apply the dropout regularization technique which randomly drops neurons during each training iteration as well as early stopping.

As a second step, the best setup for each model's hyperparameters is selected based on the validation sample. So even if the models perform perfectly in the train sample, in the validation sample, we observe a decrease in the performance from a discrimination accuracy perspective.

The values reported although high its driven mainly by the relatively medium size dataset along with the fact that these machine learning techniques are specified by a significant number hyperparameters.

After selecting the best candidate from each statistical technique, we estimate their generalization capacity in the test sample. The empirical results are summarized in the model validation chapter (Chapter 6).

In order to gain a further view on the importance of each vocabulary in predicting the outcome we employ a variable importance algorithm as in Friedman (2017). The results shown in Fig. 4 depict the relative importance of each vocabulary signal in overall crisis prediction. This is performed by randomizing each vocabulary signal and measure the relevant drop in the model forecasting accuracy. Signals for which the process leads in high loss off accuracy, rank higher in the importance chart.

We notice that the AFINN vocabulary weighted by the number of pages and the economic importance of the country where the Central Bank is situated provides the higher explanatory power along with the S&P based monogram averages (in levels and integers) which take the second and fourth place. An important finding is that the self-updating qualitative dictionary we created ranks with a very high importance (third place) implying that the forecasting accuracy of our framework will increase as the number of speeches belonging to the corpus will increase by time. As a

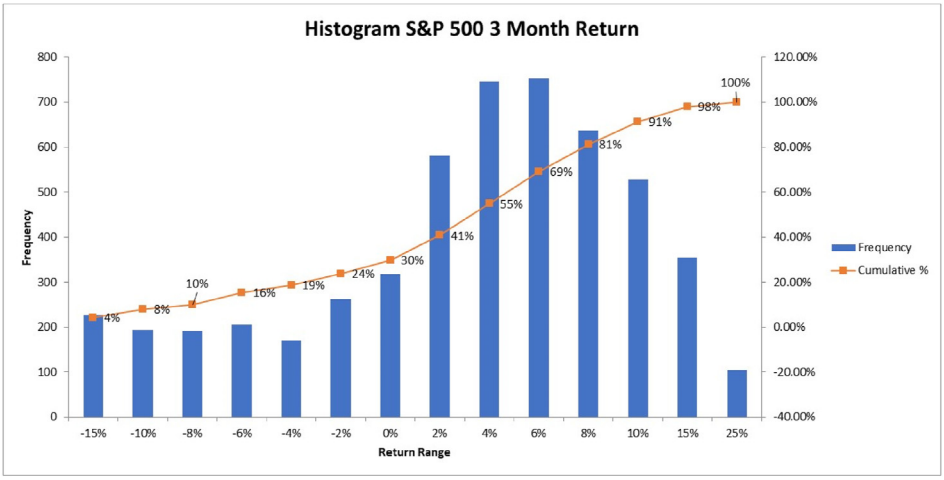

Fig. 3. Cumulative distribution of S&5 500 3-month return in the train sample.

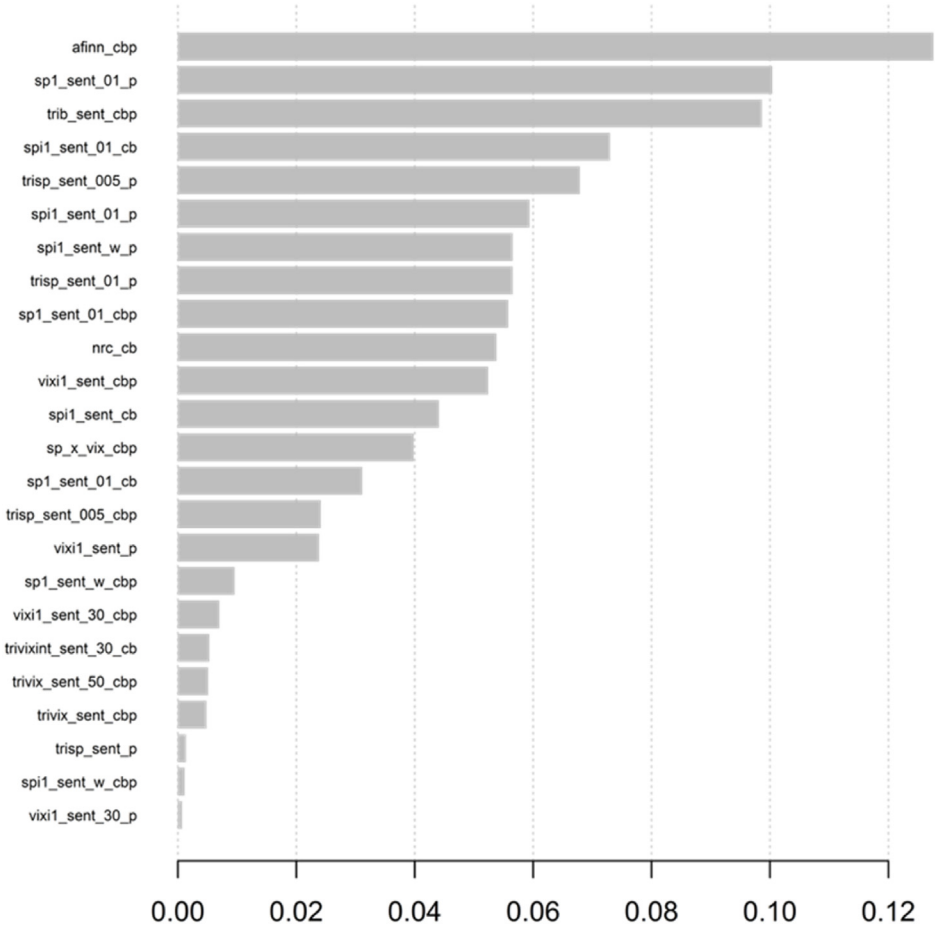

Fig. 4. Variable Importance plot constructed by randomizing each vocabulary signal and measure the relevant drop in the model forecasting accuracy. Signals for which the process leads in high loss off accuracy, rank higher in the importance chart.

general remark S&P based dictionaries on monograms provide higher forecasting capacity when combines vs VIX based dictionaries on trigrams.

6. Model validation

In the final step, we apply all the models developed and sentiment dictionaries in the corpus of speeches that are classified in the test sample to evaluate the performance of the global sentiment index.

Our experimental results provide evidence that filtering central banker's speeches to calculate a global economic sentiment index may function as an early warning system for financial market turmoil. With respect to the discriminatory power of forecasting severe financial market events, the best performing model is the XGBoost as the AUROC Coefficient is 94% for the Train sample, 93% for the validation sample and 78% for the test sample. Based on the confusion Matrix we obtain a relative high hit rate in forecasting market turmoil in a 3-month horizon but this involves a trade-off between Type I and Type II error when setting up an Early Warning system.

Based on Table 3, it is evident that the classification model based on the XGBoost model is more accurate in terms of predicting severe events in the S&P 500 index, compared to the other candidate models. DNNs are performing slightly worse than XGBoost in the test sample driven by probably overfitting in the train sample due to the high number of parameters used in their setup. Additionally the dataset size of the current study is rather restrictive for training DNNs due to their complexity. SVMs and RF are underperforming XGBoost based on our results probably to increased parameterization flexibility of the latter in capturing nonlinear patterns in the dataset. To further assess the performance of the models developed in this study we outline in Table 4 the AUROC and the KS statistic across models and across samples. Examining both classifications metrics, it is evident that XGBoost algorithm capture better nonlinearities in the sentiment scores of all dictionaries.

Further, we present in Fig. 5 the ROC curve corresponding to the best performing model (XGBoost) on both train and test basis. This curve is created by plotting the true positive rate against the false positive rate at various threshold settings. As such, they illustrate the obtained trade-offs between sensitivity and specificity, as any increase in sensitivity will be accompanied by a decrease in specificity. The closer the curve follows the left-hand border and then the top border of the ROC space, the more accurate the modelling approach. The corresponding ROC curve of extreme gradient boosting (XGBoost) is high above the 45-degree line even in the test sample supporting the high degree of efficacy and generalization capacity of the proposed employed machine learning system.

Finally, to further explore the stability of the proposed framework using XGBoost we re-calibrate the all models and dictionaries with different definition for severe events by shifting the percentile distribution. Based on the experimental results (see Table 5) overall

Table 3
Validation results by statistical model and sample.

XGBoost				
Train Period	Predicted			
Actual	0	1	Signal	Rate
0	770	132	False Alarm	15%
1	14	110	Hit Rate	89%
Test Period	Predicted			
TRUE	0	1	Signal	Rate
0	516	277	False Alarm	35%
1	4	41	Hit Rate	91%
Deep Neural Network				
Train Period	Predicted			
TRUE	0	1	Signal	Rate
0	800	102	False Alarm	11%
1	16	108	Hit Rate	87%
Test Period	Predicted			
TRUE	0	1	Signal	Rate
0	470	323	False Alarm	41%
1	7	38	Hit Rate	84%
Support Vector Machine				
Train Period	Predicted			
TRUE	0	1	Signal	Rate
0	700	202	False Alarm	22%
1	27	97	Hit Rate	78%
Test Period	Predicted			
TRUE	0	1	Signal	Rate
0	432	361	False Alarm	46%
1	13	32	Hit Rate	71%
Random Forest				
Train Period	Predicted			
TRUE	0	1	Signal	Rate
0	740	162	False Alarm	18%
1	22	102	Hit Rate	82%
Test Period	Predicted			
TRUE	0	1	Signal	Rate
0	451	342	False Alarm	43%
1	10	35	Hit Rate	78%

Table 4
AUROC and KS statistics by model for train and test sample.

Train Period	XGBoost	Deep Neural Network	Support Vector Machine	Random Forest
AUROC	94%	96%	85%	89%
KS	70%	71%	60%	63%
Test Period	XGBoost	Deep Neural Network	Support Vector Machine	Random Forest
AUROC	78%	71%	63%	67%
KS	57%	52%	45%	47%

the performance of the XGBoost -NLP approach is rather stable with slight decrease in classification accuracy when severity threshold of S&P 500 increases as the classification is performed on a more imbalanced dataset. The fact the AUROC increases when we reduce the crisis severity threshold to 6% is due to the reduction of imbalancing between crisis and not crisis days in the sample. In general, it is more difficult to classify correctly the categories in imbalanced datasets due to the preponderance of one category over the other are more so by reducing imbalancing we obtain more accurate results. In any case, we retain the 8% threshold for classifying a financial crisis event as we deem it more appropriate from a severity perspective.

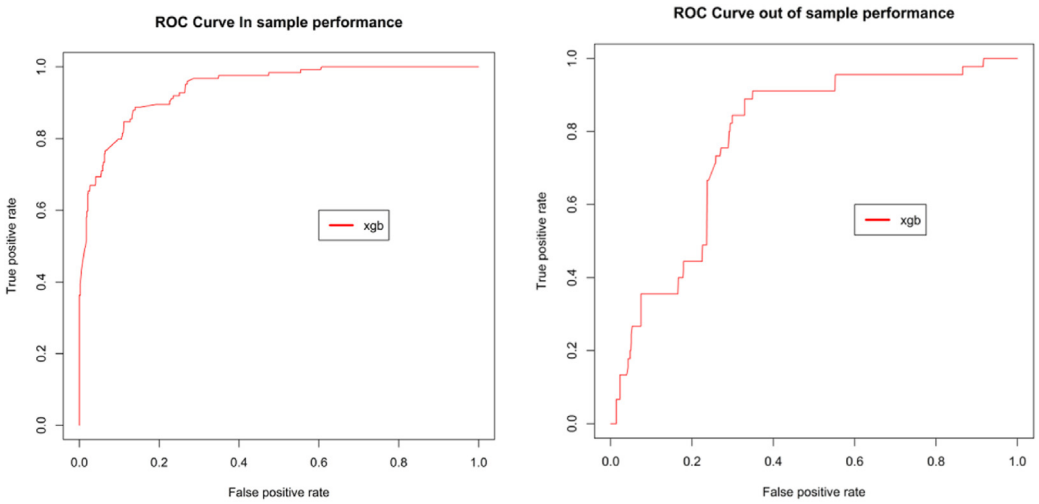

Fig. 5. ROC curve for In and out of sample discriminatory power.

Table 5
XGBoost classification accuracy sensitivity analysis compared across different severity thresholds.

Threshold for severe events in S&P 500	8%	10%	6%
XGBoost AUROC (Test Period)	78%	75%	82%

7. Conclusions and future work

This paper has addressed the question of whether the central bank speeches have a medium -term predictive ability on the occurrence of a financial turmoil event. We have investigated the question by quantifying the information contained in the Central Bank speeches using a series of dictionaries. The purpose of central bank speeches is to share information of the current central bank operations and their view about the outlook of the economy. One aim of providing such information is to enable outside observers to draw informed conclusions about future policy. Central bank speeches based on our analysis indeed provide a valuable source of information and can provide some insight about the medium-term evolution of the global economy if processed appropriately. To investigate this we perform an in-depth NLP analysis combined with machine learning algorithms, which include tokenization, dictionaries set up and training of relevance and predictive statistical models. Enhancing the flexibility of our proposed framework, we include algorithms to adjust the terms in our dictionaries in an automatic way. This is an important innovation of our work since the corpus of speeches examined cannot cover all possible trigrams. Furthermore, we believe that the forecasting accuracy of our framework will increase as the number of speeches belonging to the corpus will increase by time. We then perform thorough validation of the developed framework to test its generalization capacity. Based on our experimental results, the proposed sentiment index developed may function as a global surveillance tool and can provide an early warning framework for possible upcoming financial market turbulence in the medium term.

To investigate the robustness of using XGBoost for building the global sentiment index we compare its forecasting accuracy against three state of the art ML techniques broadly used in academia. Specifically, using the same setup a Deep Neural Network, an SVM and a random forest statistical model is implemented. Based on the empirical results the superiority of XGBoost under the specific classification problem is evident. Results also provide positive evidence for DNNs, which should be further explored in a larger corpus in the future. Examining AUROC and KS statistical metrics across all models in the test sample strong evidence exists for the outperformance of XGBoost and its generalization capacity on data. The experimental results of our study renders our approach attractive and promising to researchers and practitioners in central banking for monitoring the stability of the global financial system.

Moreover, as discussed in the introduction, it is of paramount importance the broad coverage of the speeches included in order aggregate the views of all central banks across the world. However, the results do suggest that weighting of speeches by distinguishing the size of central banks increase the forecasting accuracy of the sentiment index. Depending on how important the central bank is regarding the monetary base of the global economy, the information contained in the speeches of its member may be regarded an important variable for predicting the future evolution of the economic climate.

Our study has made three distinct contributions. First, we have made a significant methodological contribution by providing a holistic framework to analyse a corpus of speeches and aggregate all the relevant information in a sentiment index for global monitoring of the financial stability. Secondly, we have shown that NLP and ML can succeed in summarizing the themes that are included in unstructured free text speeches performed by central banks members globally. Furthermore, the proposed technical framework is highly adaptive in the introduction of additional speeches collected in real time. Third, the results of our research have indicated that using ML models can provide increased forecasting accuracy in future financial market disruptions.

In the future, to gain more insight into the effect of speeches, it may be necessary to expand the corpus in other documents produced by central banks, like the financial stability reports. This could increase the generalization with regard to predictive ability of the sentiment index to upcoming crisis events. In addition, we can investigate the benefit of the sentiment index in combination with

other conventional financial early warning indicators. Such research may show that the by including text analysis in the macro economic forecasting process may boost our awareness for future instability events. Finally, it is also important to extend our research on speeches and reports that belong to other financial participants and policy experts like insurance authorities, EBA, etc.

Appendix

Table 1
Distribution of speeches by Central Bank

Central Bank	Speeches in the sample
Bank for International Settlements	1
Bank Indonesia	49
Bank of Albania	198
Bank of Algeria	4
Bank of Argentina	17
Bank of Aruba	1
Bank of Austria	33
Bank of Bahamas	9
Bank of Bahrain	36
Bank of Barbados	70
Bank of Belgium	13
Bank of Belize	1
Bank of Bolivia	1
Bank of Bosnia and Herzegovina	4
Bank of Botswana	33
Bank of Brazil	8
Bank of Cambodia	1
Bank of Canada	296
Bank of Chile	92
Bank of China	31
Bank of Colombia	1
Bank of Curacao and Saint Maarten	13
Bank of Cyprus	3
Bank of Denmark	57
Bank of England	429
Bank of Estonia	13
Bank of Finland	107
Bank of France	208
Bank of Ghana	31
Bank of Greece	106
Bank of Guatemala	1
Bank of Guyana	2
Bank of Hungary	9
Bank of Iceland	54
Bank of India	9
Bank of Ireland	201
Bank of Israel	62
Bank of Italy	210
Bank of Jamaica	15
Bank of Japan	417
Bank of Jordan	1
Bank of Kenya	145
Bank of Korea	59
Bank of Kosovo	6
Bank of Kuwait	2
Bank of Latvia	3
Bank of Lithuania	12
Bank of Luxembourg	26
Bank of Macedonia	65
Bank of Malaysia	329
Bank of Malta	27
Bank of Mauritius	101
Bank of Mexico	77
Bank of Mozambique	1
Bank of Namibia	20
Bank of Nepal	14
Bank of Netherlands	84
Bank of Nigeria	17
Bank of Norway	130
Bank of Pakistan	67
Bank of Papua New Guinea	37
Bank of Philippines	280
Bank of Portugal	50
Bank of Romania	51
Bank of Russia	16
Bank of Samoa	5
Bank of Serbia	58
Bank of Seychelles	8
Bank of Sierra Leone	4
Bank of Slovakia	3
Bank of Slovenia	4
Bank of Solomon Islands	13
Bank of Spain	166
Bank of Sri Lanka	30
Bank of Tanzania	1
Bank of Thailand	141
Bank of Trinidad and Tobago	81
Bank of Turkey	62
Bank of Uganda	128
Bank of Ukraine	7
Bank of United Arab Emirates	5
Bank of Uruguay	1
Bank of Zambia	104
Bulgarian National Bank	22
CPMI	3
Croatian National Bank	6
Czech National Bank	32
Deutsche Bundesbank	468
Eastern Caribbean Central Bank	6
European Central Bank	1.586
Federal Reserve Bank of Atlanta	2
Federal Reserve Bank of Boston	5
Federal Reserve Bank of Chicago	8
Federal Reserve Bank of Dallas	19
Federal Reserve Bank of Kansas City	17
Federal Reserve Bank of Minneapolis	23
Federal Reserve Bank of New York	277
Federal Reserve Bank of Philadelphia	32
Federal Reserve Bank of Richmond	2
Federal Reserve Bank of San Francisco	1
Federal Reserve System	702
Hong Kong Monetary Authority	105
Maldives Monetary Authority	4
Monetary Authority of Macao	22
Monetary Authority of Singapore	168
Reserve Bank of Australia	323
Reserve Bank of Fiji	99
Reserve Bank of India	584
Reserve Bank of Malawi	21
Reserve Bank of New Zealand	85
Reserve Bank of Vanuatu	1
Saudi Arabian Monetary Agency	17
South African Reserve Bank	219
Sveriges Riskbank	191
Swiss National Bank	201
Grand Total	**10.538**

Table 2
Dictionary methods of aggregation - scoring

ndex	Agregation Methodology	mono/trigram	Index
sp1_sent	Simple average	monogram	sp
vix1_sent	Simple average	monogram	vix
spi1_sent	Simple average discretized (−1,1)	monogram	sp
vixi1_sent	Simple average discretized (−1,1)	monogram	vix
sp_x_vix	sum of sp -vix	monogram	sp-vix
sp_m_vix	product of sp -vix	monogram	sp-vix
sp1_sent_w	Simple average threshold in monogram frequency (over 30 & less than 5000)	monogram	sp
vix1_sent_w	Simple average threshold in monogram frequency (over 30 & less than 5000)	monogram	vix
spi1_sent_w	Simple average discretized (−1,1) threshold in monogram frequency (over 30 & less than 5000)	monogram	sp
vixi1_sent_w	Simple average discretized (−1,1) threshold in monogram frequency (over 30 & less than 5000)	monogram	vix
trisp_sent	Simple average	trigram	sp
trispint_sent	Simple average discretized (−1,1)	trigram	sp
trivix_sent	Simple average	trigram	vix
trivixint_sent	Simple average discretized (−1,1)	trigram	vix
sp1_sent_01	Simple average with additional threshold 10%	monogram	sp
vix1_sent_30	Simple average with additional threshold 30%	monogram	vix
spi1_sent_01	Simple average discretized (−1,1) with additional threshold 10%	monogram	sp
vixi1_sent_30	Simple average discretized (−1,1) with additional threshold 30%	monogram	vix
trisp_sent_01	Simple average with additional threshold 10%	trigram	sp
trispint_sent_01	Simple average discretized (−1,1) with additional threshold 10%	trigram	sp
trivix_sent_30	Simple average with additional threshold 30%	trigram	vix
trivixint_sent_30	Simple average discretized (−1,1) with additional threshold 30%	trigram	vix
sp1_sent_005	Simple average with additional threshold 5%	monogram	sp
vix1_sent_50	Simple average with additional threshold 50%	monogram	vix
spi1_sent_005	Simple average discretized (−1,1) with additional threshold 5%	monogram	sp
vixi1_sent_50	Simple average discretized (−1,1) with additional threshold 50%	monogram	vix
trisp_sent_005	Simple average with additional threshold 5%	trigram	sp
trispint_sent_005	Simple average discretized (−1,1) with additional threshold 5%	trigram	sp
trivix_sent_50	Simple average with additional threshold 50%	trigram	vix
trivixint_sent_50	Simple average discretized (−1,1) with additional threshold 50%	trigram	vix
afinn	general-purpose lexicon	–	–
nrc	general-purpose lexicon	–	–
bing	general-purpose lexicon	–	–
loughran	general-purpose lexicon	–	–
trib_sent	XGBoost based updating dictionary	–	–

References

Apel, Mikael, Grimaldi, Blix, Marianna, Hull, Isaiah, 2019. How Much information do monetary policy committees disclose? Evidence from the FOMC's minutes and transcripts. Working Paper Series 381, Sveriges Riksbank (Central Bank of Sweden).

Bennani, Hamza, 2019. Does People's Bank of China communication matter? Evidence from stock market reaction. C. In: Emerging Markets Review, 40. Elsevier, 1-1.

Breiman, L., 2001. Random forests. Mach. Learn. 45 (1), 5–32.

Breiman, L., Friedman, J., Stone, C., Ohlsen, R., 1984. Classification and Regression Trees. CRC press.

Bruno, G., 2016. Text mining and sentiment extraction in central bank documents. In: IEEE International Conference on Big Data (Big Data). IEEE.

Chatchawan, Sapphasak, 2019. Words Matter: Effects of Semantic Similarity of Monetary Policy Committee's Decision on Financial Market Volatility. No. 121. Puey Ungphakorn Institute for Economic Research.

Chen, Tianqi, Guestrin, Carlos, 2016. Xgboost: a scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining.

Nielsen, Finn Årup, 2011 May. A new: Evaluation of a word list for sentiment analysis in microblogs. In: Proceedings of the ESWC2011 Workshop on 'Making Sense of Microposts': Big Things Come in Small Packages 718 in CEUR Workshop Proceedings, pp. 93–98.

Friedman, Jerome H., 2001. Greedy function approximation: a gradient boosting machine. Ann. Stat. 29 (5), 1189–1232.

Friedman, Jerome H., 2017. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. springer open.

Goodfellow, Ian, Bengio, Yoshua, Courville, Aaron, 2016. Deep Learning. MIT press.

Hubert, Paul, Fabien, Labondance, February 17, 2017. Central Bank Sentiment and Policy Expectations. Bank of England Working. Paper No. 648.

Kahveci, Eyup, Odabaş, Aysun, 2016. Central banks' communication strategy and content analysis of monetary policy statements: the case of Fed, ECB and CBRT. In: Procedia-Social and Behavioral Sciences, 235, pp. 618–629.

Maqsood, H., Mehmood, I., Maqsood, M., Yassood, Y., February 2020. A local and global event sentiment based efficient stock exchange forecasting using deep learning. Int. J. Inf. Manag. 50, 432–451.

Masawi, Becksndale, Bhattacharya, Sukanto, Terry, Boulter, 2018. Does the information content of central bank speeches impact on the level of exchange rate? A comparative study of Canadian and Australian Central Bank communications. Rev. Pac. Basin Financ. Mark. Policies 21, 1850005, 01.

Mathur, Aakriti, Rajeswari, Sengupta, May 7, 2019. Analysing Monetary Policy Statements of the Reserve Bank of India.

Shapiro, Adam, Hale, Wilson, Daniel, 2019. Taking the fed at its word: a new approach to estimating central bank objectives using text analysis. In: Federal Reserve Bank of San Francisco Working Paper, 2, 2019.

Su, Shiwei, Hassan Ahmad, Ahmad, Wood, Justine, 2019. How effective is central bank communication in emerging economies? An empirical analysis of the Chinese money markets responses to the people's bank of China's policy communications. Rev. Quant. Finance Account. 1–25.

Tim Loughran, McDonald, Bill, 2011. When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. J. Finance 66 (1), 35–65.

Vapnik, Naumovich, Vladimir, Vapnik, Vlamimir, 1998. Statistical Learning Theory, vol. 1. Wiley, New York.