CentralBankRoBERTa

A Fine-Tuned Large Language Model for Central Bank Communications

Moritz Pfeifer[1] and Vincent P. Marohl[2]

[1]Institute for Economic Policy, University of Leipzig, 04109 Leipzig, Germany
pfeifer@wifa.uni-leipzig.de
[2]Department of Mathematics, Columbia University, New York NY 10027, USA
vincent.marohl@columbia.edu

Abstract. *Central bank communications are an important tool for guiding the economy and fulfilling monetary policy goals. Natural language processing (NLP) algorithms have been used to analyze central bank communications, but they often ignore context. Recent research has introduced deep-learning-based NLP algorithms, also known as large language models (LLMs), which take context into account. This study applies LLMs to central bank communications and constructs* CentralBankRoBERTa*, a state-of-the-art economic agent classifier that distinguishes five basic macroeconomic agents and binary sentiment classifier that identifies the emotional content of sentences in central bank communications. We release our data, models, and code.*

Keywords: Central Bank Communication, Sentiment Analysis, Multiclass Classification, Large Language Model, Monetary Policy.

1 Introduction

Today, central bank communications are considered an important monetary policy tool. Through communications, central banks have been said to guide the economy. Communications are supposed to help the central bank better fulfill monetary policy goals such as price stability, stable interest rates or employment. A burgeoning literature has explored the use of natural language processing (NLP) to analyze central bank communications. Most, if not all, of these studies rely on NLP algorithms that assume a bag-of-words structure. That is, the algorithms ignore context and instead analyze texts as a collection of individual words treated independently without regard for grammar or word order. More recently, research in computational linguistics has introduced deep-learning-based NLP algorithms often referred to as large language models (LLMs) due to their large number of parameters (up to billions). These algorithms learn the syntactic and semantic relationships between words from a large body of texts. This allows them to take context into account.

The aim of this study is to apply LLMs to central bank communications. Using a sample of pre-labeled sentences from the Federal Reserve Bank (Fed), the Bank of International Settlements (BIS) and the European Central Bank (ECB), we construct the first economic agent classifier of central bank communications. This classifier distinguishes five basic macroeconomic agents in central bank communications: households, firms, the financial sector, governments, and the central banks themselves. It is our view that such a classifier is fundamental for analyzing central bank communications. Central bank communications do not take place in a vacuum. Monetary policy may address different, and sometimes even opposing agents within the economy, so that signals that are relevant to one agent may be adversely relevant to another, or not relevant at all. Research that is interested in better understanding how central bank communications address different actors within the economy; whether communications exhibit a bias of some actors over others; and which actors in the economy are paying the closest attention to what a central bank says may find such a classifier particularly useful.

Next, we may want to know whether a given communication emits a positive or negative signal about, and arguably to, the economic actor that a central bank is talking about. For this purpose, we have trained a binary sentiment classifier. The classifier is able to distinguish sentences with regards to their emotional content. A word of caution on the interpretation of sentiment. In the literature, the emotional content of central bank communications is often equated with monetary policy stances. Positive emotions are then supposed to represent "doves", i.e. central bankers in favor of accommodative monetary policies, and negative emotions represent "hawks", i.e. central bankers in favor of restrictive monetary policies. While it may be generally useful to detect hawkish and dovish stances in central bank communications, we believe that it is misleading to attribute binary emotions to such stances. From a theoretical point of view, there is no reason to believe that a "hawkish" central banker will only talk "negatively" in all contexts. Indeed, if soft speech is like oil on bruised skin, hawkish central bankers would surely be well advised to speak positively on occasion without ever having to compromise their stance. Our sentiment classifier may be useful for research interested in monitoring how central banks feel about the constituents of the economy and whether these feelings manage to steer the economy in specific directions or not.

The rest of this paper is structured as follows. In the first part we review some of the recent literature of central bank communication with a focus on the most common practices of natural language processing. In the second part, we give an overview of our data and labeling method. The third part presents the experiments and results of our classification model and its performance against other LLMs such as BERT, FinBERT and XLNet as well as more traditional machine learning algorithms such as Naïve Bayes, Support Vector Machines or Random Forest. We conclude our paper with a few suggestions for the type of analysis our model could be used for.

2 Measuring Central Bank Communication

Empirical analyses of central bank communication have relied on natural language processing (NLP) techniques to quantify textual information. Most studies employ NLP for downstream econometric purposes, which may be the reason why less time has been spent studying the accuracy of the linguistic models employed. In our view, NLP models that have been used so far in measuring central bank communications have failed on two fronts. First, they are unable to take into consideration the contextual information of central bank communications and, second, they do not view communication as an interaction between a speaker and an audience. The first problem would be less nagging if central bank communications were linguistically simple. Alas, they are not. The Flesch-Kincaid score, a test giving information on the easiness of understanding a text, of an average Fed speech requires a college degree (Siklos et al. 2018). As former Fed president Alan Greenspan quipped, "If you have understood me, I must have misspoken." The second problem disregards the primary purpose of language, the transmission of information from one person to another. Empirical studies of central bank communication have been remarkably cavalier about defining the public a central bank is supposed to address. This is all the more surprising because many of the most widespread econometric models distinguish between different types of agents in the economy. It thus appears appropriate to proceed similarly when classifying the public central banks are communicating with.

Sentiment analysis has become a popular method to evaluate the effects of information contents on market actors. To our knowledge, the sole approach in the literature has been to employ dictionary methods, in which a set of n-grams are predefined and a document's sentiment score is calculated based on the relative occurrence of positive (e.g. "dovish") and negative (e.g. "hawkish") n-grams. This allows for interpretability and consistency. Picault & Renault (2017) (based on ECB press conferences), Bennani & Neuenkirch (2017) (ECB Governing Council speeches), create a field-specific hawkish/dovish sentiment dictionary to subsequently create time-series sentiment data on ECB speeches; other examples of dictionary techniques include Schmeling & Wagner (2015), Correa et al. (2017), Goldfarb et al. (2005), Stekler & Symington (2016) and Bennani (2020), the latter using an optimistic/pessimistic dictionary variant. Moniz & de Jong (2014) use Naïve Bayes, a supervised machine learning method, to group MPC minute sentences from the Bank of England by topic and then infer sentiment via a dictionary method. Popular sentiment dictionaries that use the hawkish/dovish distinction include Loughran & Mcdonald (2011) and Apel & Blix Grimaldi (2012). Some studies such as Gáti & Handlan (2022) focus on longer n-grams to capture more context. In all cases, there are no scores to capture the accuracy of n-gram classification methods other than manual verification of results. It is thus difficult to estimate if the produced sentiment scores truly reflect what the authors intended to constitute a positive/negative communication event.

A challenge of using bag-of-word techniques is to account for the linguistic idiosyncrasies of central bank communications. These have a notorious character, which central bankers such as Alan Greenspan used to wear like a crown and which have been laid aside only recently (Blinder et al., 2022; Ferrara & Angino, 2022; Issing,

2019). Central bank communications highly depend on context, yet n-grams cannot capture this: the bigram "raise rates" flags both the sentences "we will not raise rates" and "we will raise rates", which are opposite in meaning. This is further aggravated as dictionaries typically exclude qualitative adjectives (e.g., Loughran & Mcdonald, 2011), thus leaving the inference of sentiment to the interpretation of nouns. Similarly, topic models like LDA may group communication about both negative and positive inflation shocks under the same topic, as these typically include the same terms (e.g. "CPI", "interest rates", "prices"). More sophisticated methods (e.g. using longer n-grams) to address such challenges have the adverse effect of substantially reducing data points, leading to less robust results when analyzing infrequent communications.

Another strand of literature is focused on topic models for analyzing central bank communications. Boukus & Rosenberg (2006) use Latent Semantic Analysis (LSA) to group FOMC statements by themes. Meade & Acosta (2015) and Ehrmann & Talmi (2016) use unweighted and TF-IDF-weighted document-term matrices to determine cosine-similarity between speeches. Following this topic model approach, Hansen & McMahon (2016) and Feldkircher et al. (2023) pursue an LDA-based (Blei et al., 2003) approach to group speeches by topics or ideological position, respectively. One weakness of topic-model approaches is that they do not directly provide an interpretable sentiment, but merely show connections within document term matrices. Further, data-driven methods do not allow for comparability between datasets, which makes it difficult to generalize empirical findings with regards to the contents of central bank communication.

A solution may be found in recent finance studies that show that machine learning algorithms outperform bag-of-word methods in sentiment analysis, including domain-specific dictionaries (Frankel et al., 2022; Purda & Skillicorn, 2011). Further, Huang et al., (2022) demonstrate that deep-learning models outperform non-deep-learning models in finance sentiment analysis. Such deep learning models, also referred to as large language models (LLMs), are typically pre-trained on a large text corpus (Devlin et al., 2018; Peters et al., 2018; Radford et al., 2018) and then fine-tuned on a task-specific dataset to extrapolate sentiment for out-of-sample data. Unlike dictionary approaches, LLMs detect more general patterns and can thus determine sentiment without a predefined n-gram present. LLMs have the additional advantage that they consider the context of a word in a given sentence, e.g., negation is detected.

On the conceptual side, we observe that a large theoretical literature has underlined heterogeneity of audiences in central bank communication (Blinder et al., 2022; Coibion et al., 2020; Vayid, 2013). For example, Binder (2017) and Coibion & Gorodnichenko (2015) consider inflation expectations to be formed differently in households, firms and financial markets, which according to them would motivate more audience-specific communications from central banks. Reid & Siklos (2020) bemoan the lack of audience-discriminating data in central bank communications. While awareness for audience-specific analysis has thus received substantial theoretical support, it is yet to materialize in empirical analyses. Blinder et al. (2008) observe that "virtually all the research to date has focused on central bank communication with the financial markets" (p. 941). Pfeifer et al. (2022) conduct an audience-based sentiment

analysis and demonstrate that Fed communications distinguish between economic agents. This means the *cui bono* question remains unanswered: how valid is the classification of an event as positive or negative if heterogeneous audiences have opposing interests? The sentence "we will lower interest rates" may constitute a positive event for households who enjoy lower borrowing costs and a negative event for banks whose lending profitability is reduced.

Two problems exist with the current NLP models: the inability to consider contextual information (context sensitivity) and the lack of attention to communication as an interaction between the speaker and the audience (audience specificity). In the next section, we will construct two deep learning-based classification models with the hope that they will remedy these two shortcomings of measuring central communications.

3 Training Data and Labeling Methodology

Our corpus comprises the three largest English-speaking databases of central bank speeches, the BIS (1998-2022), which includes speeches of 83 central banks from Albania to Zambia; the Fed (1948-2022), including speeches of the board of governors and of the 12 district banks; and of the ECB (1999-2022), including speeches of the board members and of the 20 national member banks. The BIS dataset consists of 10,211 speeches, the Fed dataset of 6,765 speeches and the ECB of 2,405 speeches. Unlike many NLP algorithms, deep-learning algorithms such as BERT do not require pre-processed inputs (e.g., removing stop words and punctuation, stemming, and lemmatizing). Instead, entire raw sentences are taken as inputs. We have thus kept pre-processing at a minimum by only separating each speech into individual sentences that have at least 20 characters.

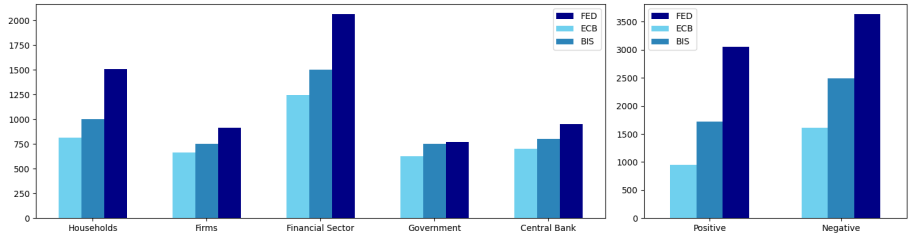
In order to train our economic agent classifier, we labeled 6,205 randomized sentences from the Fed database as speaking either about households, firms, the financial sector, the government, or the central bank itself. The question of accurately labeling sentences according to these relatively simple criteria can illustrate some of the ambiguity issues raised above. For instance, when a central banker speaks about mortgages, the information may be relevant to several actors at once. It could be relevant to households looking for home loans, to businesses looking for commercial loans, to the financial sector administering such loans or to the public sector, in the case of the government-backed mortgages. Lastly, in a context of supervisory oversight, it could pertain to the activities of the central bank itself. In each case, context is key to determining the semantic content of the sentence. We have therefore only labeled sentences where it is unambiguously clear that one and only one of our economic agents is being spoken about. Sentences about more than one economic agent have not received multiple labels.

The training dataset for our sentiment classifier consists of 6,683 pre-labeled sentences from the Fed database, which are either labeled as being positive or negative. We understand positive and negative sentiment as any sentence that talks positively or negatively about one and only one of our economic agents. We have thus only labeled

the sentiment for sentences that had previously been attributed a label of one of the economic agents. Positive and negative sentiment labels include descriptive sentences, e.g. sentences that simply portray the current, past or future economic situation of an economic agent, as well as prescriptive labels, e.g. sentences that make claims about how the economic situation should or should not be. It would be tempting to suggest that our classifiers are not only able to assess when a central bank is talking negatively or positively *about* the economic situation and prospect of a given economic agent, but also *to* the agent. While we would be cautious in interpreting our economic agent classifier in this way, the sentiment classifier may be considered to fulfill this task. Sentiment analysis is target-oriented, so that it is reasonable but not strictly necessary to assume that expressions of opinions and attitudes about an economic agent are also intended to be directed towards that agent. Forward guidance, the idea that communicating about the future course of monetary policy, can influence economic conditions today, is based on a similar assumption. It considers that whatever central banks are speaking about is a mode of address.

We have created two additional datasets of pseudo-labels from the BIS and ECB datasets for each classifier. Pseudo-labelling is the process of adding confident predicted test data to training data (Riloff 1996; Lee 2013). To further fine-tune our economic agent classifier, only additional pseudo-labels for economic agents with a high confidence threshold were used. A lower threshold of 77% for the BIS and 79% for the ECB was applied. As a result, 4,804 additional labels were created for the BIS and 4,051 for the ECB. This leaves us with a total of 15,060 labels for classifying economic agents in central bank communications. The training dataset for the sentiment classifier contains an additional 2,563 pseudo-labels generated from the ECB dataset and an additional 4,212 pseudo-labels generated from the BIS dataset, both with a lower threshold of 87%. All together, we have 13,458 labels for our sentiment classifier, of which 7,730 are negative and 5,728 are positive. The proportion of each label reflects the proportion of the randomized sentences of the Fed dataset. An overview of the training data for both classifiers is given in Fig. 1.

Figure 1. Training Data



4 Experiments and Results

Both our economic agents and sentiment classifier are based on the RoBERTa model (Liu et al., 2019), which is pre-trained on a large unlabeled text corpus and fine-tuned

for our respective downstream tasks. The RoBERTa model is a re-implementation of BERT (Devlin et al., 2018) that outperforms it in common benchmarks (GLUE, SQuAD). It is a more streamlined version by removing BERT's next sentence pre-training and operating with larger mini-batch size, learning rates, and a training dataset of 160GB compared to BERT's 16GB (Liu et al., 2019).

We use the PyTorch implementation of RoBERTa and of our baseline LLM comparison models (BERT, FinBERT, XLNet) via the HuggingFace transformer open-source library. We use the base versions of all models because performance increases of the large versions were observed to be marginal and could therefore not justify the additional computing costs. All training and testing were carried out on a A100 SXM4 GPU with 40 GB memory within the Google Colab environment.

4.1 Economic Agents Classifier

We split the economic agent dataset of 15,060 sentences into 80% (12,048) training and 20% (3,012) validation data. For this classifier, we focus on RoBERTa with BERT and XLNet (Yang et al., 2019) as baseline comparisons. These LLMs are trained to assign each sentence from our dataset to one of five economic groups: Households, Firms, Financial Sector, Government and Central Bank. We fine-tune these models by testing parameters for batch size, learning rate (LR) and epochs. Manual testing has yielded the hyperparameters in Fig. 2 to be optimal. Warmup steps (n=500), gradient accumulation steps (n=2) and decay weight (0.01) parameters are set equal for all models. To increase the batch size and still be able to fit higher batch sizes into memory, we add gradient accumulation steps (GAS) to our model. Dropout rates and parameter weights are left at the default settings. The results of our training are presented in Fig. 3. We find that the RoBERTa model scores higher than both the BERT and XLNet models for our training task.

Figure 2: Hyperparameters Agent Classifier

Model	GAS	Batches	LR	Epochs
BERT	2	32	4e-6	5
XLNet	2	32	4e-6	5
RoBERTa	2	64	7e-6	5

Figure 3: Result Agent Classifier

Model	Precision	Recall	F1
BERT	0.92	0.92	0.92
XLNet	**0.93**	0.92	0.92
RoBERTa	**0.93**	**0.93**	**0.93**

4.2 Sentiment Classifier

We split the sentiment dataset of 13,458 sentences into 80% (10,766) training and 20% (2,692) validation data. For this classifier, we use a RoBERTa model. For our baseline comparison, we consider the LLMs BERT, XLNet and FinBERT (Huang et al., 2022). We also consider other machine learning approaches: Support Vector Machine (SVM), Random Forest, and a two-step TF-IDF and Naïve Bayes (NB) algorithm. SVM separates categories in data by constructing hyperplanes (Mullen & Col-

lier, 2004). Random Forest Algorithms combine decision trees for data discrimination (Frankel et al., 2022). TF-IDF weights words based on frequency, and NB assigns a most likely label to a sentence using a simple bag-of-words approach (Li, 2010). These additional machine learning models were included in the baseline because they represent popular methods in the sentiment analysis literature and thus offer an additional point of comparison of our model with more traditional machine learning methods.

The models are trained to assign each sentence from our dataset a binary positive/negative label. As before, we fine-tune the LLM models by testing parameters for optimal batch size, learning rate and epochs, the results of which are reported in Fig. 4. Warmup steps (n=500), gradient accumulation steps (n=2) and decay weight (0.01) are set equal, and dropout rates and parameter weights are on default settings. When applicable, we test for optimal parameters for the machine learning models via grid search. The results of our training are presented in Fig. 5. Our findings show that all deep-learning models outperform all non-deep-learning models for our task. Among the LLMs, the RoBERTa model outperforms the BERT, FinBERT and XLNet fine-tuned models.

Figure 4: Hyperparameters Sent Classifier

Model	GAS	Batches	LR	Epochs
LSTM	0	64	4e-4	3
BERT	2	64	2e-6	5
FinBERT	2	32	1e-6	5
XLNet	2	32	9e-7	5
RoBERTa	2	64	7e-6	5

Figure 5: Result Sent Classifier

Model	Precision	Recall	F1
TF-IDF	0.82	0.76	0.74
RF	0.80	0.80	0.79
LSTM	0.81	0.81	0.81
SVM	0.83	0.83	0.83
BERT	0.85	0.85	0.85
FinBERT	0.85	0.85	0.85
XLNet	0.87	0.87	0.87
RoBERTa	**0.88**	**0.88**	**0.88**

5 Conclusion

Communications relies almost exclusively on bag-of-words methods that ignore contextual information. There has also been virtually no empirical literature that distinguishes central bank communications according to what economic agents are being addressed. We introduce CentralBankRoBERTa, a state-of-the-art LLM based on the pre-trained RoBERTa model. We fine-tune the model for two tasks: (1) classification of economic agents and (2) sentiment in central bank communication. We train our model on two novel sets of pre-labeled sentences from FED, ECB and BIS communications. We benchmark CentralBankRoBERTa on other LLMs (BERT, FinBERT,

XLNet) and machine learning techniques (NB, SVM, RF) and find it outperforms them in both tasks.

CentralBankRoBERTa provides a novel way to account for context in central bank communication. Our economic agent classifier may be useful for research interested in understanding the ways in which central bank communications address different actors within the economy. This can refine previous analyses by discriminating the effects of communication on market events by separating the addressed economic agent. It also allows researchers to gain insights into whether certain actors are being favored or disadvantaged by the central bank's communications and, by extension, monetary policy. By opening up new ways to identify the most attentive audiences, researchers can gain insights into which groups are most likely to be impacted by monetary policies. The sentiment classifier can be a valuable tool for researchers interested in monitoring how central banks feel about the various constituents of the economy and whether these feelings influence economic outcomes. Further, CentralBankRoBERTa's two tasks synergize, and researchers may for the first time analyze group-specific sentiment, which can spawn novel economic analyses.

References

1. Apel, M., & Blix Grimaldi, M. (2012). The Information Content of Central Bank Minutes (Issue 261). Sveriges Riksbank (Central Bank of Sweden).
2. Bennani, H., & Neuenkirch, M. (2017). The (home) bias of European central bankers: New evidence based on speeches. *Applied Economics*, *49*(11), 1114 -1131. https://doi.org/10.1080/00036846.2016.1210782
3. Binder, C. (2017). Fed speak on main street: Central bank communication and household expectations. *Journal of Macroeconomics*, *52*, 238–251. https://doi.org/10.1016/j.jmacro.2017.05.003
4. Blinder, A., Ehrmann, M., de Haan, J., & Jansen, D.-J. (2022). *Central Bank Communication with the General Public: Promise or False Hope?* (No. w30277; p. w30277). National Bureau of Economic Research. https://doi.org/10.3386/w30277
5. Blinder, A. S., Ehrmann, M., Fratzscher, M., De Haan, J., & Jansen, D.-J. (2008). *Central Bank Communication and Monetary Policy*. ECB.
6. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.
7. Boukus, E., & Rosenberg, J. V. (2006). The Information Content of FOMC Minutes. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.922312
8. Coibion, O., & Gorodnichenko, Y. (2015). Is the Phillips Curve Alive and Well after All? Inflation Expectations and the Missing Disinflation. *American Economic Journal: Macroeconomics*, *7*(1), 197–232. https://doi.org/10.1257/mac.20130306
9. Coibion, O., Gorodnichenko, Y., Kumar, S., & Pedemonte, M. (2020). Inflation expectations as a policy tool? *Journal of International Economics*, *124*, 103297. https://doi.org/10.1016/j.jinteco.2020.103297
10. Correa, R., Garud, K., Londono, J. M., & Mislang, N. (2017). Sentiment in Central Bank's Financial Stability Reports. International Finance Discussion Paper, 2017(1203), 1–46. https://doi.org/10.17016/ifdp.2017.1203

11. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://doi.org/10.48550/ARXIV.1810.04805
12. Ehrmann, M., & Talmi, J. (2016). Starting from a Blank Page? Semantic Similarity in Central Bank Communication and Market Volatility (Issues 16–37). Bank of Canada.
13. Feldkircher, M., Hofmarcher, P., & Siklos, P. L. (2023). Cacophony in Central Banking? Evidence from Euro Area Speeches on Monetary Policy. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.4196226
14. Ferrara, F. M., & Angino, S. (2022). Does clarity make central banks more engaging? Lessons from ECB communications. *European Journal of Political Economy*, *74*, 102146. https://doi.org/10.1016/j.ejpoleco.2021.102146
15. Frankel, R., Jennings, J., & Lee, J. (2022). Disclosure Sentiment: Machine Learning vs. Dictionary Methods. *Management Science*, *68*(7), 5514–5532. https://doi.org/10.1287/mnsc.2021.4156
16. Gáti, L., & Handlan, A. (2022). Monetary Communication Rules. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.4260246
17. Goldfarb, R. S., Stekler, H. O., & David, J. (2005). Methodological issues in forecasting: Insights from the egregious business forecast errors of late 1930. Journal of Economic Methodology, 12(4), 517–542. https://doi.org/10.1080/13501780500343524
18. Hansen, S., & McMahon, M. (2016). Shocking language: Understanding the macroeconomic effects of central bank communication. 38th Annual NBER International Seminar on Macroeconomics, 99, S114–S133. https://doi.org/10.1016/j.jinteco.2015.12.008
19. Huang, A. H., Wang, H., & Yang, Y. (2022). FinBERT: A Large Language Model for Extracting Information from Financial Text*. *Contemporary Accounting Research*, *n/a*(n/a). https://doi.org/10.1111/1911-3846.12832
20. Issing, O. (2019). The long journey of central bank communication. MIT Press.
21. Lee, D.-H. (2013). Pseudo-Label: The Simple and Efficient Semi-Supervised Learning Method for Deep Neural Networks. *ICML 2013 Workshop : Challenges in Representation Learning (WREPL)*.
22. Li, F. (2010). The Information Content of Forward-Looking Statements in Corporate Filings-A Naïve Bayesian Machine Learning Approach: The information content of corporate filings. *Journal of Accounting Research*, *48*(5), 1049–1102. https://doi.org/10.1111/j.1475-679X.2010.00382.x
23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. https://doi.org/10.48550/ARXIV.1907.11692
24. Loughran, T., & Mcdonald, B. (2011). When Is a Liability Not a Liability? Textual Analysis, Dictionaries, and 10-Ks. *The Journal of Finance*, *66*(1), 35–65. https://doi.org/10.1111/j.1540-6261.2010.01625.x
25. Moniz, A., & de Jong, F. (2014). Predicting the Impact of Central Bank Communications on Financial Market Investors' Interest Rate Expectations. In V. Presutti, E. Blomqvist, R. Troncy, H. Sack, I. Papadakis, & A. Tordai (Eds.), *The Semantic Web: ESWC 2014 Satellite Events* (Vol. 8798, pp. 144–155). Springer International Publishing. https://doi.org/10.1007/978-3-319-11955-7_12
26. Meade, E. E., & Acosta, M. (2015). Hanging on Every Word: Semantic Analysis of the FOMC's Postmeeting Statement. FEDS Notes, 2015(1580), Article 1580. https://doi.org/10.17016/2380-7172.1580
27. Mullen, T., & Collier, N. (2004). *Sentiment Analysis using Support Vector Machines with Diverse Information Sources*. 412–418.

28. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. https://doi.org/10.48550/ARXIV.1802.05365
29. Pfeifer, M., El Guindi, M., & Salazar-Caicedo, G. (2022). *Secrets of the Temple or Noise of the Agora?* Paper presented at the 25th Central Bank Macroeconomic Modeling Workshop, November 9, 2022.
30. Picault, M., & Renault, T. (2017). Words are Not All Created Equal: A New Measure of ECB
31. Communication. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2980777
32. Purda, L. D., & Skillicorn, D. (2011). Identifying Fraud from the Language of Financial Reports. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1670832
33. Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., & others. (2018). *Improving language understanding by generative pre-training*.
34. Reid, M., & Siklos, P. (2020). *Building Credibility and Influencing Expectations The Evolution of Central Bank Communication* (Issue 10144). South African Reserve Bank.
35. Riloff, E. (1996). Automatically generating extraction patterns from untagged text. *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2*, 1044–1049.
36. Schmeling, M., & Wagner, C. (2015). Does Central Bank Tone Move Asset Prices? SSRN Electronic Journal. https://doi.org/10.2139/ssrn.2629978
37. Siklos, P., Amand, S. S., & Wajda, J. (2018). The evolving scope and content of central bank speeches.
38. Stekler, H., & Symington, H. (2016). Evaluating qualitative forecasts: The FOMC minutes, 2006–2010.*International Journal of Forecasting*, *32*(2), 559–570. https://doi.org/10.1016/j.ijforecast.2015
39. Vayid, I. (2013). *Central bank communications before, during and after the crisis: From open-market operations to open-mouth policy* (Working Paper No. 2013–41). Bank of Canada Working Paper.
40. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R., & Le, Q. V. (2019). *XLNet: Generalized Autoregressive Pretraining for Language Understanding*. https://doi.org/10.48550/ARXIV.1906.08237