Automated Detection of Emotion in Central Bank Communication: A Warning

Nicole Baerg and Carola Binder*

Abstract

Central banks have increased their official communications. Previous literature measures complexity, clarity, tone and sentiment. Less explored is the use of fact versus emotion in central bank communication. We test a new method for classifying factual versus emotional language, applying a pretrained transfer learning model, fine-tuned with manually coded, task-specific, and domain-specific datasets. We find that the large language models outperforms traditional models on some occasions, however, the results depend on a number of choices. We therefore caution researchers from depending solely on such models even for tasks that appear similar. Our findings suggest that central bank communications are not only technically difficult but also subjectively difficult to understand.[1]

JEL codes: E58, E59, C10, C55

1 Introduction

A large and growing literature examines central bank communication and its effects on public officials, markets, and the mass public. Interest in central bank communication has increased as central banks have entered into what Haldane et al. (2021) refer to as the "second wave" of central bank communication, in which central bankers are asking themselves, "How should we communicate this in a way that engages a broader cross-section of society?" (p. 279). In this quest for engagement, central bankers may have increased their reliance on emotional appeals. Classifying the emotional content of central bank speeches is both an interesting methodological challenge and potentially important for making sense of the evolution and impacts of central bank communication. This paper

*Baerg: University of Essex, nicole.baerg@essex.ac.uk, Binder: Haverford College, cbinder1@haverford.edu

[1]The authors would like to thank James Brookes, Michael McMahon, Chad Hazlett, and participants of the NIER Workshop on 'Advances in Central Banking' for helpful comments and suggestions. We would also like to thank our human coders and research assistants Devansh Goyal and Samuel Ross for carefully coding our sample of sentences and David Yen-Chieh Liao for help with collecting and organising our corpus.

makes first steps toward that goal, while also raising notes of caution about the difficulty of this and related tasks.

A subsection of the central bank communication literature has considered sentiment rather than emotion. Sentiment analysis is a means of assessing if language in a given text is positive, negative, or neutral. Previous studies have used sentiment analysis to uncover the monetary policy stance conveyed by central bank communications (Ehrmann and Fratzscher, 2007; Ehrmann and Wabitsch, 2022; Hayo and Neuenkirch, 2010; Hubert and Fabien, 2017); to build metrics of market sentiment (for a review in finance see Kearney and Liu (2014)); and to estimate a central bank's policy position (Shapiro and Wilson, 2021). Sentiment analysis has also been used to understand the movements of key economic variables (Shapiro et al., 2020), which are important inputs in monetary policymaking. Most research to date has used dictionary based approaches or traditional bag-of-words machine learning models. New research also considers sentiment in the context of deep-learning models, specifically Large Language Models (LLMs). In work most related to ours, Pfeifer and Marohl (2023) build a sentiment classifier of central bank communications using manually labeled central bank speech data. Unlike our model, Pfeifer and Marohl (2023) use training data coded for sentiment not emotion, though they use sentiment and emotion interchangeably.

Sentiment and emotion are related but not exactly the same.[2] Emotional language is defined as language that is trying to invoke feelings in the receiver. As cited in Cochrane et al. (2022) emotional language "causes brain activity [in the listener] associated with the retrieval of memories about those emotions, which helps people to more quickly resolve ambiguous affective states." Sentiment, by contrast, refers to tone or *polarity*, assessing the positive or negative tone in an expression. Emotion and sentiment is also different with respect to time with emotions often experienced within a relatively short time period ("I am angry"), whereas sentiments are felt much longer ("I enjoyed the talk and found the speaker convincing"). In addition, sentiments are often expressed in relation to an object ("I felt good about the interview"), whereas emotions are not necessarily object-anchored ("I feel sad") (Munezero et al., 2014). In this paper, we examine emotion versus fact based statements in central bank communications and as distinct from sentiment.

Researchers in economics and political science have recently started measuring political and

[2]See (Liu, 2020) for an approach from computational linguistics.

economic texts for emotional language using approaches from computational social science. For example, Cochrane et al. (2022) look specifically at the use of emotional language in parliamentary speeches in Canada. They find that while video recordings of parliamentary speeches exhibit more emotion than transcripts, transcripts still transmit emotion. Similarly, Gennaro and Ash (2022) measure emotion in US Congressional speeches between 1858–2014. Showing how emotion can vary with covariates, these authors find that US Congressional speeches are more emotional during times of political conflict. Important for us, Gennaro and Ash (2022) also find that emotion is distinct from positive and negative sentiment (p.1038). This along with some of the conceptional work in computer science indeed suggests that that emotion is related but different from tone empirically.

In this paper, we utilise previous research on central bank sentiment and large language models but examine emotion in central bank speeches. At first glance, central bankers may rarely appeal to emotion and favour facts. This might be because central bankers have roles that are quite technocratic in nature. Despite this, we find instances where central bankers use emotional appeals in their speeches. In our sample of central bank speeches, we find that 60% of the sentences that are manually code are sentences labeled as facts whereas the remaining 40% express emotional language. Using these sentences, we then classify unseen sentences for emotional language using state of the art computational methods from natural language processing. We also run a number of experiments to examine the performance of different model variations. We introduce additional layers of pre-trained off-the-shelf labeled data, using both task specific datasets that label sentences for emotion versus fact and in-domain specific pre-training data that label sentences for sentiment specifically in a corpus of central bank communications. We also run more traditional machine learning algorithms and compare the results.

We find that the state of the art large language models are useful but offer researchers a variety of choices in their implementation and require a significant amount of tinkering. Further, we find that tinkering produces large variations in the results and performs poorly on difficult to label texts. Furthermore we find that while using existing pre-training data for central bank sentiment is helpful, within-domain language exposure is not a magic bullet. We therefore advocate a cautious approach to researchers that are considering using off the shelf LLM for the study of central bank communication. We specifically highlight how customisation, layering, and coder agreement all

matter for model performance.

As a result of our experiments, we suggest that researchers studying central bank communication broaden their interpretation of textual complexity and textual difficulty. To date, most research has measured textual complexity using simple readability metrics (Bholat et al., 2019).[3] Yet, as we show in this paper, central bank communications are subjectively complex in terms of affect, feelings and emotion. In other words, labeling central bank communications for emotion is itself difficult because central bankers use both stories and numbers to convey meaning sometimes interchangeably. This suggests that central bank communication is not only hard to read but also subjectively difficult to discern. Our findings are also consistent with recent research that argues that central bank communications is often cognitively complex (McMahon and Naylor, 2023). Finally, our findings also contribute to recent literature which shows that transformer models often struggle with economic texts and that relatively simple word count models do surprisingly well across all sorts of tasks within economics and finance (Ahrens et al., 2024).

2 Training Data and Labeling Methodology

The Bank of International Settlements (BIS) provides the text of nearly all speeches by central bank officials, staring in 1997. To utilise this data, we scraped the text of all available speeches including meta-data such as the speaker's position, speech-title and affiliation, and the date the speech was delivered. We therefore have a corpus construction of all published central bank speeches from January 1, 1997 and April 1, 2021, or 16,784 speeches.

Because we want to train and test some models, we take some of this data for manually coding, training, and testing. In this paper, we restrict the sample to central bank speeches from central bankers where English is the bank's majority language (e.g. USA, Canada, Ireland, New Zealand, UK and Australia). To ensure that we get a sample that contains emotive language, we use purposeful selection of speeches for our training data. To identify possible speeches for our training data that include emotive language, we applied a simple dictionary of people-centered language including the words: "the people", "the public", "consumers", "citizens", "voters", "taxpayers", and "population"

[3]Similarly in studies of legislative text, researchers also depend on such metrics (Benoit et al., 2019; Spirling, 2016; McDonnell and Ondelli, 2022)

as well as their derivatives. We then took a sample of speeches that scores relatively high on these terms to be used for training. The rationale for using purposeful rather than random selection is that we wanted to ensure sentences in our training and test sets contained the subjective language that we are interested in.

Once we identified speeches that contained emotive or subjective language, we then split these speeches into sentences. This gave us a total of 750 sentences that we used for human labeling/annotation. To label the data, we use four human annotators and ask them to label our central bank sentences into a binary classification, a sentence can be either a "FACT" or "FEEL" sentence. Our four coders included both authors of this paper (female, tenured academics whose native language is North American English) and two male undergraduate students in economics whose native language is American English. The instructions for manual coding given to the annotators was identical to those used in previous research that asked annotators to label FEEL and FACT sentences from an internet forum (more on this below). The annotators were instructed with the following question: *Is the speaker attempting to make a fact based argument or appealing to feelings and emotions?*. In total, our four human annotators coded 750 sentences, and these 750 sentences were then earmarked for training.

As we had four coders, we also assessed the human coding for inter-coder agreement across the human annotators. Of those 750 sentences, only 486 (65%) of the sentences had complete coder agreement for a particular sentence (sentences were either coded by pairs or triplets of annotators). In order to be included, there needed to be at least a pair of annotators and there needed to be agreement across all annotators. From these 486 sentences, we then split the sentences into a training and testing set. The training test split that we used was a random sample of 80/20 sentences. For the remaining 206 "difficult' sentences where annotator agreement was not unanimous amongst the coders, we use these sentences as unseen (out of sample) validation for all models. Table 1 shows some examples of our coded sentences:

In summary, we collected a corpus of approximately 16,500 central bank speeches made by central bankers around the world between 1997 and 2021, archived on the Bank for International Settlements (BIS) website. From this corpus, we extract speeches from central bankers in majority English speaking countries (USA, Canada, Ireland, New Zealand, UK and Australia). We then use a

Table 1: Human Annotation of Central Bank Statements for Fact and Feeling

Agreement	Label	Sentence
Yes	Fact	Inflation was high in the 1980s and lower more recently – even with lower unemployment – because that's what people expected.
Yes	Fact	Similarly, there has been a downward shift in estimates of potential growth in Asia, especially in China and South Korea.
Yes	Feel	We fear that if we make such statistics available, all of our hard efforts to communicate the outlook as a whole get washed away in an extreme focus on point estimates.
Yes	Feel	We do have a stake in supporting strong and sustainable growth, and that is why we play an important advisory role and help shed light on some of the trade-offs at play.
No	Unclear	It ignores history, which clearly shows that those societies that have done the most to improve the economic well-being of their citizens are those where the public sector has provided the right social climate for the dynamism and creativity of individuals and businesses to thrive.
No	Unclear	But there is always room for improvement.

simple dictionary method to try to find central bank speeches that contained emotional language and purposefully selected a sample of speeches with this language to build our human-labeled training, testing, and validation sets. We had either two or three annotators code approximately 750 sentences into a binary classification: FACT versus FEEL sentences. We use only those sentences where all coders provide unanimous annotation for our training and testing set. We hold out those sentences that the annotators find especially difficult to agree on and use these sentences for out of sample, validation data. In our analysis, all our models are assessed on this unseen and relatively difficult to classify data.

2.1 Experiments and Results

We construct a binary classification tool that can be use to classify whether a sentence in a central bank speech is using either FACT or FEEL language. To build this classification tool, we explore a number of models including state of the art large language models as well as traditional supervised learning models based on bag-of-word representations such as Naive Bayes and Logistic Regression.

In this section we discuss the models and present our results.

The most sophisticated language model that we use is a BERT model. BERT stands for "Bidirectional Encoder Representations from Transformers" (BERT). The idea behind BERT is similar to word embeddings (for a review of word embeddings see Rodriguez and Spirling (2022)) in that context around language is helpful for understanding meaning. BERT was developed by Google and was originally trained to learn about the (English) language (context) by training on English language books (Book Corpus) as well as Wikipedia pages (Devlin et al., 2018). The type of transfer learning model we use is distilBERT, which is a smaller but powerful derivative of BERT.

Modelling language via (distil)BERT means that a model has an pre-understanding of textual data from books and Wikipedia. The underlying language data is not human labeled rather the model is exposed to vasts amounts of texts and the model uses these textual examples to learn about word or token associations (context) given what it observes in the training data. BERT was originally pre-trained to do two main things, language modelling and sentence prediction. In the case of language modelling, for a given text, a proportion of words or tokens is hidden from the BERT model. The model is then trained to predict the missing tokens from context. As a result of the training process, BERT learns contextual embeddings for words from an enormous and generic corpus.

What is beneficial to us as social science researchers is that this pre-training, which is computationally expensive, comes already available for customization. This is why it is called transfer learning. Transfer learning is the improvement of learning for a given new task as a results of the transfer of knowledge from a another task that has already been learned. Running models based on BERT, the researcher starts with a base of knowledge about language (a language model). The researcher can then fine-tune the generic language model from BERT with more customized or smaller datasets to optimize its performance for the specific user defined task. For pre-trained language models like BERT, domain adaptation through the use of pre-training improves their use for downstream, in-domain tasks (Röttger and Pierrehumbert, 2021). In our case we use a BERT based model distilBERT (Sanh et al., 2019) as our underlying language model and then supplement this language with within domain language of central bank communications using our hand-coded sentences.

We also run a second set of experiments where we layer the large language model with two sets of off the shelf, manually coded data. The first off the shelf manually coded data is task specific but out of domain. This dataset comes from the Internet Agreement Corpus (IAC) and made available by researchers online (Walker et al., 2012). The underlying textual data was scraped from *4forums.com*, a website for political debate and discourse. Importantly for us, the corpus was manually annotated for emotion at the sentence level. Also interesting is that coders found coding for emotion relatively difficult (Krippendorff's $\alpha = 0.32$). The annotations are generated using Mechanical Turk workers who label participant emotional stance in a number of question/response pairs on political topics. Workers were asked to rank the question-response pair in terms of the level of emotional content of the text using the same instructions that we used: *Is the respondent attempting to make a fact based argument or appealing to feelings and emotions?* Annotators ranked each sentence a score from -5 to 5. The researchers calculated a mean score and ultimately, each sentence is ranked as either being predominately FACT or FEEL, for a total of annotated 4070 annotated sentences. As the economy was not a topic that was discussed by the internet forum users, this dataset shares the same task (within task) but is outside of the domain of central bank communications (out of domain).

The second off the shelf dataset comes from Pfeifer and Marohl (2023). These authors construct a dataset of 6683 manually labeled sentiment scores for central bank communications. The training dataset is only from central bank speeches given by the U.S. central bank. Unfortunately, the researchers do not go into much detail about the labeling process and it is unclear the number of individuals that annotated the data nor do they specifically mention the difficulty (or not) of the labeling task. The researchers only label for positive and negative sentences. They say they also remove labels about the central bank or those that are vague. In their sample, the researchers find that negative sentiment is expressed more often than positive sentiment. Their labeled data can be found on GitHub.[4] Different from the above, this dataset is in the domain of central bank communications (in domain) but considers a different task (outside task).

In addition to DistilBERT and its derivative models, we also run more traditional supervised machine learning models including Naive Bayes and Logistic Regression. Unlike the above language model, the Naive Bayes model does not consider words and their context but is based on the "bag of

[4]CentralBankRoBERTa

words" assumption, or the term frequency tokens. The Naive Bayes model computes probabilities of tokens for each class. Class predictions are then made from summing the probabilities for tokens in each sentence and assigning the predicted class to whichever class probability is largest. Similarly, we also use a logistic regression model. The logistic regression classifier uses a weighted combination of tokens and passes these weights through a sigmoid function. The sigmoid function then transforms the input to a number between 0 and 1, which we can interpret as class probabilities.

In order to compare class predictions across the models, we convert each FACT and FEEL predictions into predicted probabilities. We also calculate commonly reported model metrics and present those as well. Finally as mentioned above, our annotators found the task relatively difficult (Krippendorff's $\alpha = 0.52$ for the statements where coders agree but α -0.426 for the disagreeing statements). We therefore present the results when we use different coders as the gold standard for the unseen data.

Table 2 gives performance metrics across the different models reporting accuracy, precision, recall, and F1-score when the gold standard is generated by annotator one. Accuracy is the proportion of true predictions made by the models. We can see that both the generic LLM and the within domain models are more accurate than the other models. The F1 score is the harmonic mean of precision and recall and takes into account not only the number of prediction errors that the model makes, but also the type of errors made. Table 2 shows that all of the models have a comparable F1 score. If we look at the component parts, we see that all of the models have a much higher recall score than precision score. Models with high recall identify the positive cases in the data, even though they may also wrongly identify some negative cases as positive cases ($true\ positives/(true\ positives + false\ negatives)$). Precision on the other hand counts the percentage of correctly identified FEEL sentences over all those that were classified as FEEL sentences ($true\ positives/(true\ positives + false\ positives)$). The table shows none of the models are very precise.

Moving to results for our second annotator in Table 3, here we find that the traditional machine learning models are performing better against the gold standard. Unlike those in Table 2, we have much higher precision across most of our models but at the cost of recall. We also have higher accuracy across all of our models, with the exception of the in-domain trained sentiment LLM.

One point of concern is that the class balance in the out of sample data is significantly different

Table 2: Performance Metrics for FACT vs FEEL classification with Coder 1 Gold Standard

	Accuracy	Precision	Recall	F1
DistilBERT	0.47	0.19	0.95	0.30
DistilBERT with Emotion	0.25	0.19	0.95	0.32
DistilBERT with Central Bank Sentiment	0.65	0.23	0.38	0.28
Naive Bayes	0.52	0.22	0.62	0.32
Logistic Regression	0.26	0.19	0.92	0.32

Table 3: Performance Metrics for FACT vs FEEL classification with Coder 2 Gold Standard

	Accuracy	Precision	Recall	F1
DistilBERT	0.55	0.91	0.91	0.70
DistilBERT with Emotion	0.85	0.92	0.91	0.91
DistilBERT with Central Bank Sentiment	0.33	0.90	0.29	0.44
Naive Bayes	0.55	0.91	0.56	0.70
Logistic Regression	0.84	0.93	0.89	0.91

than the training and testing datasets. In the training and testing data, we had a 40%, 60% spilt of FEEL to FACT. In the validation sentences we have the opposite and more extreme imbalance such that 76% of validation are FEEL and 24% are FACT. Because we want to apply this model to unseen data, we do not want a model that is restricted to doing well only on similarly distributed data. There is some evidence from computer science that the transformer models do better than more traditional models for this particular problem (Hendrycks et al., 2020), however, the wide variation in model performance when compared against different coding standards suggests that the task may be too hard at least given the instructions the annotators were provided with. We have of course given the computer a difficult task in the first place by using as held out data only those sentences where there was no human annotator agreement in the first place.

So as to try to evaluate qualitatively where the classifiers are going wrong (or not), we took a sub-sample of sentences that our two undergraduate annotators has disagreed on and asked them to reconsider the sentences and produce a consensus label. In addition to the consensus label, they were also asked to rank their level of confidence in their consensus label indicating either low, medium, or highly confident. We then compare sentence level predictions across all of the models when the

annotators have "high certainty" and "low certainty".

As above, we see a lot of variation in the results. For the models where the coders reach a consensus with "high certainty" results in Table 4 show that none of the models do particularly well. The number of "high certainty" sentences are about 40 percent of the data. The LLM model pre-trained on sentiment sentences performs slightly better in terms of accuracy than other models, however, if we consider the F1 score (the harmonic mean of precision and recall) the Naive Bayes model does just as well. All of the models do poorly on precision. All of the models over-predict the FACT label, which is the dominant label in the training set.

Table 4: Performance Metrics for FACT vs FEEL classification with Consensus "High" Certainty

	Accuracy	Precision	Recall	F1
DistilBERT	0.49	0.10	0.83	0.19
DistilBERT with Emotion	0.17	0.07	0.83	0.12
DistilBERT with Central Bank Sentiment	0.70	0.12	0.50	0.19
Naive Bayes	0.53	0.11	0.83	0.20
Logistic Regression	0.20	0.08	1.00	0.15

If we shift to the "low certainty" sentences, we see differences again as shown in Table 5. For this sub-sample of sentences, the DistilBERT model with no additional training data and the Naive Bayes model perform the best. There is no clear advantage of using either within domain or within task additional language for these sentences. We find it particularly interesting that the models seem to perform better on the uncertain labels than the certain labels. This might suggest that the annotators led each other afield when generating a consensus label or that the models are significantly influenced by the class balance in the training data.

Table 5: Performance Metrics for FACT vs FEEL classification with Consensus "Low Certainty"

	Accuracy	Precision	Recall	F1
DistilBERT	0.57	0.50	0.75	0.60
DistilBERT with Emotion	0.43	0.43	0.96	0.59
DistilBERT with Central Bank Sentiment	0.52	0.41	0.25	0.31
Naive Bayes	0.55	0.49	0.57	0.52
Logistic Regression	0.40	0.41	0.89	0.56

One possible critique of the above is related to the initial low inter-reliability of the human coded data and/or the fact that maybe one coder dominated the other coder in the consensus coding. Because the human raters had low agreement on how to classify central bank communications for emotions to start with means that unsurprisingly it is difficult for the machine to do well at classification too. One possible solution is to therefore up the quality of manual annotations. In other words, if we invested more attention and instruction to our human coders then consequently the machine would have a more meaningful target and we would get a better result. However, in this paper, we specifically used datasets whose authors suggested that they be applied by other researchers and made them publicly available for that use. Arguably, we do not want to generate large amounts of manually coded, customised, and as a result, relatively expensive datasets for every research task. One significant take away from this paper therefore is that we show how even seemingly good fitting, high quality, labeled, off the shelf and in-domain data (e.g. the central bank sentiment data) under-performs expectations, as does the lesser quality, labeled, off the shelf data (IAC data).

Missing from this paper are extensive model fitting techniques such as hyperparamter tuning. One possible criticism of our paper is that, by excluding a hyperparameter tuning stage, the findings in this paper over-emphasizes performance issues related to LLM. One possible additional step we could take would be to include a third split (train/tune/test) in which to tune hyperparameters while insuring the test data remains unseen. In this paper, we specifically excluded doing this because we wanted to point out some of the non-trivial ways in which central bank communications, in their use of natural language, generates challenges beyond those "solved" by model optimisation. Table 4 in Pfeifer and Marohl, 2023 shows hyperparameter tuning including gradient accumulation, batch size, learning rate, and training epochs. As reported in their paper, even with this hyperparameter tuning stage, the BERT model performs still only slightly better (precision = 0.85) than the more traditional Naive Bayes model (precision = 0.82). Similarly, findings in Ahrens et al. (2024) shows that even with hyperparameter optimisation using ensemble based models, bag of words models perform remarkably well.

Natural language that central bankers use has a host of subjective ambiguities, cultural nuances and institutional constraints. Central banking has cultures of communication styles. Furthermore

institutional features as well as the political and economic climate all affect their communications (Baerg, 2020). Our main message therefore is that researchers considering applying state of the art tools like large language models to complex domains, need to consider the costs of using such models. Of course, if one wants to apply such an algorithm to predict massive amounts of unseen data then prediction accuracy might matter more than interpretation and such a choice may be warranted. Yet, in our view, trade-offs between using embedding type models versus bag of words models are rarely discussed.

In summary, results from our experiments do not give us confidence in using large language models to detect emotion in central bank communication. Furthermore, the general lack of transparency about the class contributions at the token level when using such models, which are easily retrieved using Naive Bayes, makes model exploration relatively impossible. These challenges coupled with challenges over determining the optimal size and number of pre-training layers and issues of coder agreement and class imbalance in pre-training and training data and selecting and implementing hyperparameter optimisation makes us skeptical about calls to abandon more traditional bag of words approaches at least in the domain of central bank communications. From our experiments, we find that the gains over standard approaches are slight, if any.

3 Conclusion

This paper presents complex language models and traditional machine learning models to help classify emotion in central bank communication. We classify central bank speeches at the sentence level. We find that transfer learning models sometimes outperform traditional machine learning models, but we also find that the results are sensitive to a number of model choices, including the number of pre-training layers, the balance of classes, the use of in-domain versus within-task pre-training data, and the selection and agreement of the labelled gold standard.

Traditional machine learning and bag of words type models are often criticised despite their relative simplicity and elegance. The argument is that bag-of-words models do not include context and therefore miss subtleties that are common in central bank communication. It is presumed that such LLM can discover such subtleties through context, though we find very little evidence that they

do. Given that traditional models allow researchers a greater ability to look "under the hood" of the models, we argue that researchers should be wary of favouring context and newer large language models at the expense of more traditional models. In fact, we find that the performance of these models on a closely related set of tasks is relatively poor, even when the training data is from the same domain.

In addition, researchers studying central banks have generally assumed that the complexity in central bank communication is lexical complexity, i.e. complexity in jargon and vocabulary. Consequently, researchers have prescribed readability metrics and have advocated simplifying central bank communication on the basis of making statements more readable. As we have shown in this paper, lexical complexity is only one type of textual complexity. We therefore offer these sober findings to researchers and policymakers interested in central bank communications.

We find that central bank texts are complex in terms of their use of affect, feelings and emotions - what we call subjective complexity. Readers of central bank communication also struggle with these nuances. If central banks are indeed interested in increasing public trust and the clarity of their communication, increasing subjective understanding may also contribute to improving the clarity of communications.

References

Ahrens, M., D. Gorguza, and M. McMahon (2024). Ecofinbench: A natural language processing benchmark for economics and finance. *PrePrint*.

Baerg, N. (2020). *Crafting consensus: Why central bankers change their speech and how speech changes the economy*. Oxford University Press, USA.

Benoit, K., K. Munger, and A. Spirling (2019). Measuring and explaining political sophistication through textual complexity. *American Journal of Political Science 63*(2), 491–508.

Bholat, D., N. Broughton, J. Ter Meer, and E. Walczak (2019). Enhancing central bank communications using simple and relatable information. *Journal of Monetary Economics 108*, 1–15.

Cochrane, C., L. Rheault, J.-F. Godbout, T. Whyte, M. W.-C. Wong, and S. Borwein (2022). The automatic analysis of emotion in political speech based on transcripts. *Political Communication 39*(1), 98–121.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Ehrmann, M. and M. Fratzscher (2007). Communication by central bank committee members: different strategies, same effectiveness? *Journal of Money, Credit and Banking 39*(2-3), 509–541.

Ehrmann, M. and A. Wabitsch (2022). Central bank communication with non-experts–a road to nowhere? *Journal of Monetary Economics 127*, 69–85.

Gennaro, G. and E. Ash (2022). Emotion and reason in political language. *The Economic Journal 132*(643), 1037–1059.

Haldane, A., A. Macaulay, and M. McMahon (2021). *Independence, Credibility, and Communication of Central Banking*, Chapter The Three E's of Central-Bank Communication with the Public, pp. 279–342. Central Bank of Chile.

Hayo, B. and M. Neuenkirch (2010). Do federal reserve communications help predict federal funds target rate decisions? *Journal of Macroeconomics 32*(4), 1014–1024.

Hendrycks, D., X. Liu, E. Wallace, A. Dziedzic, R. Krishnan, and D. Song (2020). Pretrained transformers improve out-of-distribution robustness. *arXiv preprint arXiv:2004.06100*.

Hubert, P. and L. Fabien (2017). Central bank sentiment and policy expectations.

Kearney, C. and S. Liu (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis 33*, 171–185.

Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.

McDonnell, D. and S. Ondelli (2022). The language of right-wing populist leaders: Not so simple. *Perspectives on Politics 20*(3), 828–841.

McMahon, M. and M. Naylor (2023). Getting through: communicating complex information. *Bank of England Working Paper*.

Munezero, M., C. S. Montero, E. Sutinen, and J. Pajunen (2014). Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing 5*(2), 101–111.

Pfeifer, M. and V. P. Marohl (2023). Centralbankroberta: A fine-tuned large language model for central bank communications. *The Journal of Finance and Data Science 9*, 100114.

Rodriguez, P. L. and A. Spirling (2022). Word embeddings: What works, what doesn't, and how to tell the difference for applied research. *The Journal of Politics 84*(1), 101–115.

Röttger, P. and J. B. Pierrehumbert (2021). Temporal adaptation of bert and performance on downstream document classification: Insights from social media. *arXiv preprint arXiv:2104.08116*.

Sanh, V., L. Debut, J. Chaumond, and T. Wolf (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Shapiro, A. H., M. Sudhof, and D. J. Wilson (2020). Measuring news sentiment. *Journal of econometrics*.

Shapiro, A. H. and D. Wilson (2021). Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. Federal Reserve Bank of San Francisco.

Spirling, A. (2016). Democratization and linguistic complexity: The effect of franchise extension on parliamentary discourse, 1832–1915. *The Journal of Politics 78*(1), 120–136.

Walker, M. A., J. E. F. Tree, P. Anand, R. Abbott, and J. King (2012). A corpus for research on deliberation and debate. In *LREC*, Volume 12, pp. 812–817. Istanbul, Turkey.