National Institute Economic Review (2025), 1–10
doi:10.1017/nie.2024.31

NIER CAMBRIDGE UNIVERSITY PRESS

RESEARCH ARTICLE

AUTOMATED DETECTION OF EMOTION IN CENTRAL BANK COMMUNICATION: A WARNING

Nicole Baerg[1] and Carola Binder[2]

[1]Department of Government, University of Essex, Colchester, UK and [2]School of Civic Leadership, University of Texas, Austin, TX, USA
Corresponding author: Nicole Baerg; Email: nicole.baerg@essex.ac.uk

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 data sets. We find that the large language models outperform 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 but also subjectively difficult to understand.

Keywords: Central banks; machine learning; central bank communications

JEL Codes: E52; E58; D83; C55; C82

1. Introduction

A large and growing body of 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.* (2020) 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 article makes first steps towards 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 policy-making. 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, although they use sentiment and emotion interchangeably.

Sentiment and emotion are related but are not exactly the same.[1] Emotional language is defined as a 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 are 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 article, 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 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 emotions than transcripts, transcripts still transmit emotions. Similarly, Gennaro and Ash (2022) measured emotion in the US Congressional speeches between 1858 and 2014. Showing how emotion can vary with covariates, these authors find that the U.S. Congressional speeches are more emotional during times of political conflict. It is worth noting that Gennaro and Ash (2022) also find that emotion is distinct from positive and negative sentiments (p. 1038). This, along with some of the conceptional work in computer science suggests that emotion is related to but different from tone empirically.

In this article, we utilize previous research on central bank sentiment and LLMs 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 coded are sentences labelled 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 several experiments to examine the performance of different model variations. We introduce additional layers of pre-trained off-the-shelf labelled data, using both task-specific data sets 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 LLMs 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 who 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).[2] Yet, as we show in this article, central bank communications are subjectively complex in terms of affect, feelings and

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

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

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

emotion. In other words, labelling 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 are 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, starting in 1997. To utilize this data, we scraped the text of all available speeches including metadata 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 article, 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', as well as their derivatives. We then took a sample of speeches that scored 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 were 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 labelling/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 article (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 were 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). 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 among 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 ~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 the majority English English-speaking countries (USA, Canada, Ireland, New Zealand, UK and Australia). We then used a 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-labelled training, testing and validation sets.

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

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 is 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.

We had either two or three annotators code ~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 used to classify whether a sentence in a central bank speech is using either FACT or FEEL language. To build this classification tool, we explore several models including state-of-the-art LLMs, 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 helps in understanding the 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 a pre-understanding of textual data from books and Wikipedia. The underlying language data are not human labelled; rather, the model is exposed to vast 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, is already available for customisation. This is why it is called transfer learning. Transfer learning is the improvement of learning for a given new task as a result of the transfer of knowledge from 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

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

language model from BERT with more customized or smaller data sets 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 LLM with two sets of off-the-shelf, manually coded data. The first off-the-shelf manually coded data is task-specific but out of the domain. This data set comes from the Internet Agreement Corpus (IAC) and is made available by researchers online (Walker *et al.*, 2012). The underlying textual data were 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 participants' emotional stances in several 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 with 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 4,070 annotated sentences. As the economy was not a topic that was discussed by the internet forum users, this data set shares the same task (within a task) but is outside of the domain of central bank communications (out of domain).

The second off-the-shelf data set comes from Pfeifer and Marohl (2023). These authors construct a data set of 6683 manually labeled sentiment scores for central bank communications. The training data set is only from central bank speeches given by the US central bank. Unfortunately, the researchers do not go into much detail about the labelling process and it is unclear the number of individuals that annotated the data nor do they specifically mention the difficulty (or not) of the labelling task. The researchers only label positive and negative sentences. They say they also remove labels about the central bank or those that are vague. In their sample, the researchers found that negative sentiment is expressed more often than positive sentiment. Their labelled data can be found on GitHub.[3] Different from the above, this data set is in the domain of central bank communications (in the 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 in their context but is based on the 'bag-of-words' assumption or the term frequency tokens. The Naive Bayes model computes the probabilities of tokens for each class. Class predictions are then made by summing the probabilities for tokens in each sentence and assigning the predicted class to whichever class probability is the 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.

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

[3]CentralBankRoBERTa.

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

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

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 the 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, except the in-domain trained sentiment LLM.

One point of concern is that the class balance in the out-of-sample data is significantly different than the training and testing data sets. 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 is FEEL and 24% is 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.

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 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.

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

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

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

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 is about 40% 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 overpredict the FACT label, which is the dominant label in the training set.

If we shift to the 'low certainty' sentences, we see differences again as shown in Table 5. For this subsample of sentences, the DistilBERT model with no additional training data and the Naive Bayes model performs 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 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.

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 a 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 article, we specifically used data sets 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, customized and, as a result, relatively expensive data sets for every research task. One significant takeaway from this article, therefore, is that we show how even seemingly good fitting, high-quality, labelled, off-the-shelf and

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

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

in-domain data (e.g., the central bank sentiment data) underperforms expectations, as does the lesser quality, labelled, off-the-shelf data (IAC data).

Missing from this article are extensive model-fitting techniques such as hyperparameter tuning. One possible criticism of our article is that, by excluding a hyperparameter tuning stage, the findings in this article overemphasizes 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 ensuring the test data remains unseen. In this article, 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, generate challenges beyond those 'solved' by model optimization. Table 4 in Pfeifer and Marohl, 2023 shows hyperparameter tuning including gradient accumulation, batch size, learning rate and training epochs. As reported in their article, 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) show that even with hyperparameter optimization using ensemble-based models, bag-of-words models perform remarkably well.

The 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). Therefore, our main message is that researchers considering applying state-of-the-art tools like LLMs 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, the results from our experiments do not give us confidence in using LLMs 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 optimization make us skeptical about calls to abandon a 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 article 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 several 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 criticized 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 LLMs at the expense of more traditional models. 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, that is, complexity in jargon and vocabulary. Consequently, researchers

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

have prescribed readability metrics and have advocated simplifying central bank communication based on making statements more readable. As we have shown in this article, 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.

Acknowledgements. We 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 organizing our corpus.

References

Ahrens, M., Gorguza, D. and McMahon, M. (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*, New York: Oxford University Press.

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

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

Cochrane, C., Rheault, L., Godbout, J.-F., Whyte, T., Wong, M.W.-C. and Borwein, S. (2022), 'The automatic analysis of emotion in political speech based on transcripts', *Political Communication*, **39**, 1, pp. 98–121.

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

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

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

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

Haldane, A., Macaulay, A. and McMahon, M. (2020). The 3 E's of Central bank communication with the public. CEPR Discussion Paper No. 14265.

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

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

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

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

Liu, B. (2020), *Sentiment Analysis: Mining Opinions, Sentiments, and Emotions*, New York: Cambridge University Press.

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

McMahon, M. and Naylor, M. (2023), 'Getting through: Communicating complex information.' Discussion Paper Series 18537, CEPR.

Munezero, M., Montero, C.S., Sutinen, E. and Pajunen, J. (2014), 'Are they different? Affect, feeling, emotion, sentiment, and opinion detection in text. ', *IEEE Transactions on Affective Computing*, **5**, 2, pp. 101–111.

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

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

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

Sanh, V., Debut, L., Chaumond, J. and Wolf, T. (2019), 'Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter', Preprint, arXiv:1910.01108.

https://doi.org/10.1017/nie.2024.31 Published online by Cambridge University Press

Shapiro, A.H., Sudhof, M., and Wilson, D.J. (2020), 'Measuring news sentiment', *Journal of Econometrics* **228.2** (2022): 221–243.

Shapiro, A.H. and Wilson, D.J. (2021), Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. *The Review of Economic Studies* **89.5** (2022): 2768–2805.

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

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

Cite this article: Baerg, N. and Binder, C. (2025), 'Automated detection of emotion in central bank communication: A warning', *National Institute Economic Review*, pp. 1–10. https://doi.org/10.1017/nie.2024.31