FedBERT

Camilla Hao Guo[*] **Vincent Lamine**[†] **Silvi Pandey**[‡] **Jose Regalado**[§] **Jerry Qinghui Yu**[¶]

University of California Berkeley - Haas School of Business

Abstract

This paper is meant to illustrate Natural Language Processing techniques in the context of financial data. More specifically, we compare the performance of various BERT models in understanding the language used in Federal speeches. We extend the work of [Desola et al., 2019] to Fed speech, statements, and minutes data to create FedBERT, which outperforms the classical BERT model in capturing the context of the financial jargon, specifically macroeconomics and Fed speech.

[*] hao_guo@berkeley.edu
[†] vincent_lamine@berkeley.edu
[‡] silvi_pandey@berkeley.edu
[§] jose.regalado@berkeley.edu
[¶] jerryqhyu@berkeley.edu

Preprint. Under review.

1 Introduction

The Fed (Federal Reserve) and its actions have never been as important and impactful as today. The actions taken by the Fed are of capital importance in studying the stability and dynamics of the financial markets. In recent times, speeches and written releases by the Fed have garnered an unprecedented level of scrutiny, from investors and politicians alike. In the wake of the COVID-19 crisis, never the adage "lucky who can predict the Fed" has been so true, as the support the Fed provided to various financial markets were at the origin of major market moves throughout the crisis. Not only is this true for the Fed, but generally speaking major central Banks have also gained in credibility and power over the last several decades (e.g the European Central Bank). Naturally, this argument supports our willingness to improve the ability to forecast what has become the holy grail of finance professionals, and Natural Language Processing appeared as a powerful tool to do so.

Indeed, the last decade have seen the rise of Natural Language Processing in many applications, and the financial industry is not exception to the rule. Furthermore, Google published BERT models in 2019 [Devlin et al., 2018] and paved the way for an even greater efficiency in accomplishing those tasks, significantly improving the performances of NLP predictive ability as measured by many metrics such as next sentence prediction or masked words. BERT was trained mainly on Wikipedia over several hundreds of pages, and its first performances were really astonishing on many problems. But when trying to evaluate BERT models on Financial jargon (for example - on company filings), researchers discovered the need for contextual training. This was therefore realized for FinBERT in [Desola et al., 2019]).

With this idea in mind, we wanted to propose a similar application covering Fed communications (FOMC meeting written releases and speeches), with the far-reaching goal of quickly incorporating this information into financial markets. We therefore developed a model specifically trained on Fed jargon, and tested its performances in predicting such specific outcomes. It turned out that our results show the necessity for taking into account the context in which the NLP model is used when training, as already illustrated with FinBERT on company filings.

2 Related Work

BERT is short for "Bidirectional Encoder Representations from Transformers", which is a powerful tool in the field of NLP, especially in language understanding. However, the performance of BERT is largely defined by the data used to pre-train the model. Given the sentences from Wikipedia, there is no doubt that the pre-trained BERT from Google is to analyze more general data. When it comes to articles in some specific fields such as biology and finance, it fails to understand some widely used terminologies and the specific structure of sentence. This leads to the development of those domain-specific models such as BioBERT [Lee et al., 2019] and FinBERT [Desola et al., 2019]. The latter of which is the one inspires us most and also the one we will illustrate more in the following paragraphs.

The data FinBERT focused on is SEC fillings since it contains both extremely technical knowledge and opinionated text in the Management Discussion and Analysis (MD&A) sections. As included in the article, their goal is to "include most of the domain-specific words commonly used by financial experts, by training on corpora that include contextual information not present in general language models like BERT." The authors of [Desola et al., 2019] chose to pre-train FinBERT in three different ways: The first model (FinBERT-Prime), is trained from scratch using 10K filings in the past three years, capturing specific context on financial documents without any of the references from BERT, such as pop-culture or history. The second model (FinBERT-Pre2K), is trained from scratch but using 10K filings from 1998 and 1999 to compare how financial language has changeover the last two decade. The last model (FinBERT-Combo), is trained on top of BERT-Base Uncased model to take advantage of transfer learning from the original BERT.

3 Data and Model

3.1 Data

The BERT model was trained with a general corpus consisting of Wikipedia articles and BookCorpus, so understanding the Fed nuances would require to train the model using Fed communications. These communications were mainly taking two forms : Fed speeches and written releases (minutes and statements). FOMC statements and minutes were selected because they reflect the FOMC decisions and are closely followed by market participants. Additionally, Fed board members speeches were used due to the possibility that these articles transmit hints about board members economic views and consequently about their possible vote directions.

Fortunately, all this information is available in the Fed webpage[6]. It is a well designed webpage, so it is relatively straightforward to obtain the required information using web scrapping techniques. However, it should be noted that there is a abundant JavaScript content so the Python package Selenium had to be used. Once we are able to navigate through the java controls extracting the documents we need was relatively easy. A separate process was used for the speeches prior to 2006 since they are stored in a different webpage with a different format[7]. The main difference was that rather than interacting with JavaScript controls we only needed to change the year in the URL address.

The text that was extracted from each web page using an html parser and looking for the id "*articles*" and "*leftText*" (for documents prior to 2012). The text delimited by these IDs is almost clean and we only had to eliminate some words at the end of the documents that do not add meaning to it, such as "*share*" or "*return to top*".

Once we have completed the data extraction and cleaning, we have the following data:

Table 1: Fed statements, minutes, speeches data count

	Number	Period	Words
Statements	206	1994-2020	78806
Minutes	243	1993-2020	1539020
Speeches	890	1996-2020	2781519
Total	1339	1993-2020	4399345

To put these numbers in context, this is the data used by the BERT model:

Table 2: BERT training data count

	Words (M)
Wikipedia	2500
BookCorpus	800
Total	3300

3.2 Processing Data

The data collected has numerous issues, including meaningless sentences, redundant dates and tables, and poor formatting. To remedy these issues, we perform a two-step processing. We sifted sentences revealing more information in speeches and statements. During this process, some less important sentences are eliminated. For example, "Good afternoon. It's great to be with you, and I look forward to our discussion." should be eliminated while "The Federal Reserve is committed to using its full range of tools to support the U.S. economy in this challenging time, thereby promoting its maximum employment and price stability goals." should be included. Our strategy is that we develop a Fed-keyword list and only keep sentences containing the keywords.

[6] https://www.federalreserve.gov/monetarypolicy/materials/

[7] https://www.federalreserve.gov/newsevents/speech/2005speech.htm

To better find the keywords in Fed speeches and statements, we counted the frequency of each word and selected some words with more probability to indicate financial information. That is, we mainly pick words with highest frequency except meaningless words such as prepositions and the most common verbs or adjectives. We list several selected keywords as an example. Finally, we split data and combined sentences into the form that compatible to BERT's architecture.

financial, federal, reserve, inflation, policy, monetary, interest

3.3 Model

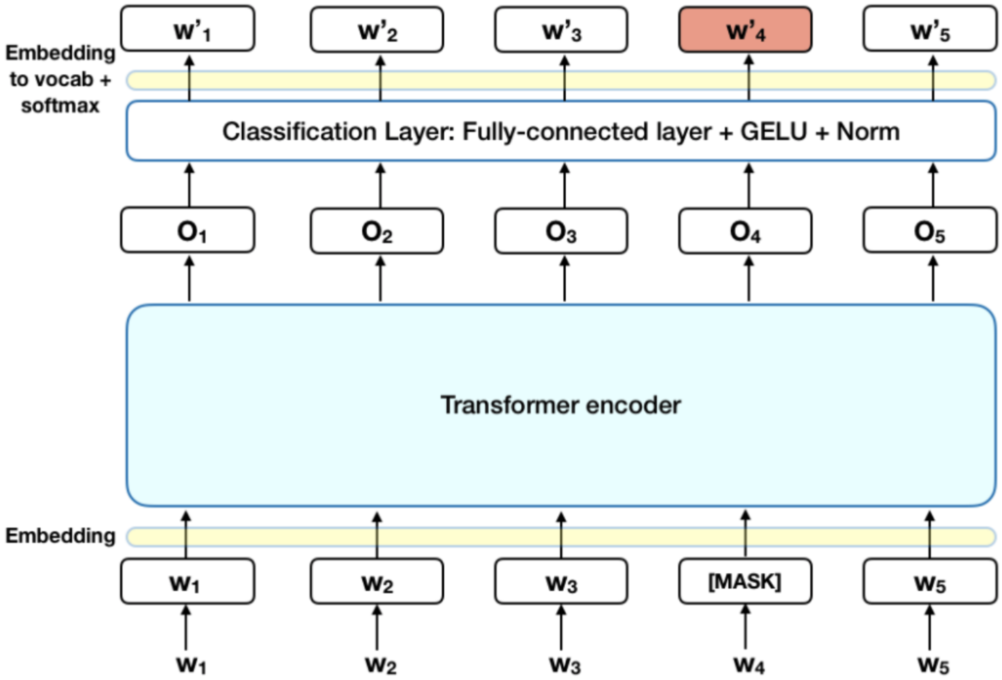
One of the main reasons for the good performance of BERT on different NLP tasks was the use of Semi-Supervised Learning. This means the model is trained for a specific task that enables it to understand the patterns of the language.

BERT is basically an encoder stack of transformer architecture. A transformer architecture is an encoder-decoder network that uses self-attention on the encoder side and attention on the decoder side. BERT has 24 layers in the Encoder stack. BERT architecture also has larger feed-forward networks with 1024 hidden units, and 16 attention heads. The total number of parameters is 340M.

The model takes [CLS] token as input first, then it is followed by a sequence of words as input. Here [CLS] is a classification token. It then passes the input to the above layers. Each layer applies self-attention, passes the result through a feed-forward network after then it hands off to the next encoder.

The model outputs a vector of hidden size (1024 for BERT). If we want to output a classifier from this model we can take the output corresponding to [CLS] token.

Here is the BERT architecture


w'1
w'2
w'3
w'4
w'5
Embedding to vocab + softmax
Classification Layer: Fully-connected layer + GELU + Norm
O1
O2
O3
O4
O5
Transformer encoder
Embedding
w1
w2
w3
[MASK]
w5
w1
w2
w3
w4
w5

Our work draws inspiration from FinBERT which was trained using 10k filings of companies. FinBERT consists of 3 sub-models. The sub-models are - FinBERT Prime, FinBERT Pre-2k and FinBERT Combo. After analyzing the results obtained by using the three different models as baselines, we concluded that we wanted to go ahead with FinBERT Prime as our baseline for FedBERT-Fin to provide a better contrast with FedBERT-BERT, which uses a Google BERT model as baseline.

We also train a model from scratch which we call FedBERT-Prime. The training parameters are the common parameters for BERT training, with a learning rate of 0.0001.

4 Results and Discussion

4.1 Training and validation loss

We train our model using Fed speech data then evaluate it using Fed statement data. We evaluate the model using a classic in language representation models — masked language modeling objectives. We mask out tokens in a random fashion and use the model to predict the masked word. The loss is defined using a softmax function on the probabilities over a defined dictionary of words. Usually BERT model is trained from scratch using a dual-objective, but unfortunately due to hardware constraints we did not have the opportunity to train next sentence prediction task. (GCP has forbidden GPU use for new users)

Figure 1 shows the loss as a function of masking probability. As expected, as a higher percentage of words are masked out, the loss function is higher. This is because when more words are masked out, there is less context in the sentence to base predictions on. Also as expected, for the training set, all 3 FedBERT models vastly outperform their counterparts, often with half of the loss compared to other models. Within the models, we see that FedBERT-BERT is the best, followed by FerBERT-Fin, then FedBERT-prm. We theorize that Fed speeches are structurally different from company reports, on which FinBERT is trained. We hypothesize that company reports are more similar to Fed statements than speeches, which we will discuss more later. On the other hand, due to the diverse nature of Wikipedia articles, BERT would be the better performer in this case. The prime variant actually outperforms FinBERT variant when few words are masked out, but gets worse as context disappears, this is due to the incomplete training due to lack of NSP that enforces sentence, grammar and syntax structure. As less context is given, the ability for a freshly trained prime model is challenged.

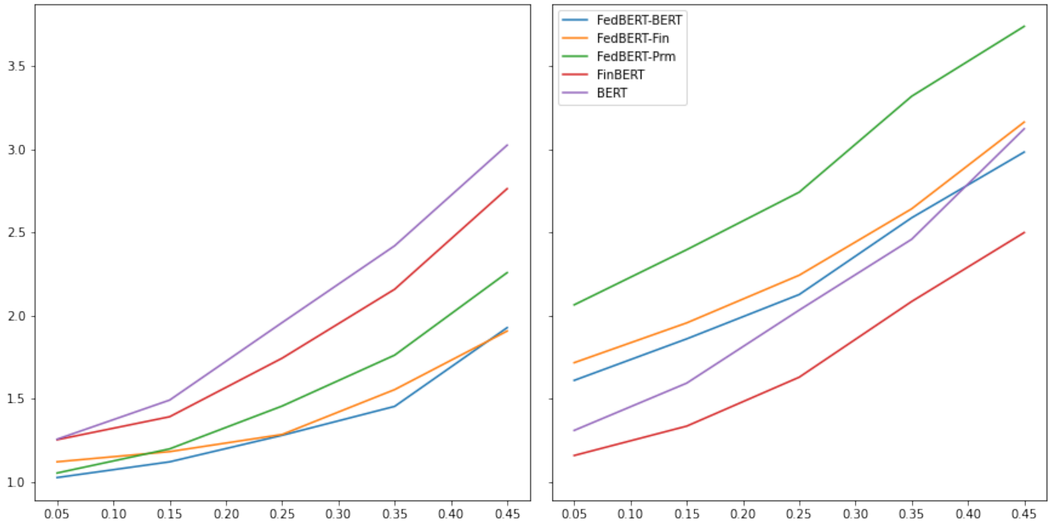


Figure 1: Loss as a function of mask probability on speeches (left) and statements (right)

When it comes to validation data, we are surprised to learn that the FedBERT models performs worse than Google BERT at lower mask probabilities and better at higher probabilities, and worse than FinBERT in general. This is especially true for the prime variant, which is drastically worse than the other two. Figure 2 shows the change in loss when we switch between training and validation data. Google BERT as a baseline did not change between the two datasets, while all FedBERT models performed much worse, especially when we increase the percentage of words masked out. While it is not expected, we think losing NSP fine-tuning has really damaged the performance of the language models, especially when it comes to sentence structures. What is more surprising, however, is that

FinBERT outperforms itself when validation data is used. This shows that the statements by the Fed is more similar to the training data used to train FinBERT. Since Fed statements are usually more formal and examined, there are more legalese involved and is more suited for FinBERT.

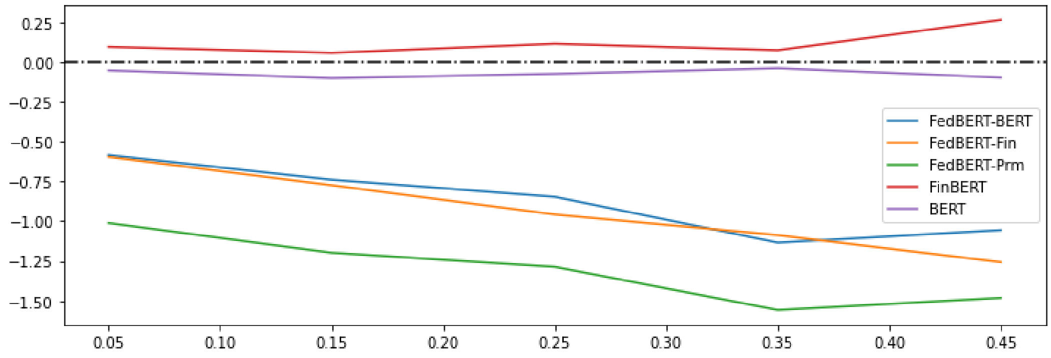


Figure 2: Change in loss between training and validation data

4.2 Recovering masked out words

It is one thing to perform well on an arbitrary masked out context, it is completely different to predict only the important words. It is possible that a language model works well for syntactical structures, but is totally lacking in excelling in specific contexts. For example, we picked this sentence out of a Fed statement (validation set) and masked out five crucial words to understand this sentence. The five words contain most of the information, so decoding them would be more important than if the model is able to identify peripheral words such as ‘its’ or ‘time’. The prediction given by the models are given in bold. Clearly all FedBERT variants outperform the baseline models including BERT and FinBERT. BERT’s answers are for the most part nonsensical, often filling the blanks with meaningless tokens like ‘-’ or ‘,’. FinBERT represents a step-up, but its predictions are still defined by its training corpus, including a lot of corporate finance verbiage. FedBERT solves all that problems in the sense that the output produced by FedBERT is the most Fed-like.

> Original: the federal reserve is committed to using its full range of **tools** to support the u.s. **economy** in this challenging time, thereby **promoting** its maximum **employment** and price **stability** goals
>
> BERT: the federal reserve is committed to using its full range of **assets** to support the u.s. **economy** in this challenging time, thereby **enhancing** its maximum **,** and price **-** goals
>
> FedBERT-BERT: the federal reserve is committed to using its full range of **tools** to support the u.s. **economy** in this challenging time, thereby **meeting** its maximum **employment** and price **stability** goals
>
> FinBERT: the federal reserve is committed to using its full range of **resources** to support the u.s. **economy** in this challenging time, thereby **meeting** its maximum **growth** and price **stability** goals
>
> FedBERT-Fin: the federal reserve is committed to using its full range of **tools** to support the u.s. **economy** in this challenging time, thereby **promoting** its maximum **employment** and price **stability** goals
>
> FedBERT-Prime: the federal reserve is committed to using its full range of **tools** to support the u.s. **economy** in this challenging time, thereby **maintaining** its maximum **employment** and price **stability** goals

As we can see, FedBERT performs really well on speech-like sentences, especially if they are macroeconomics related. However, Fed statements have a lot more varieties, including talking about businesses and currency. For the sake of completeness, we also present a case where FedBERT variances did not perform as well as BERT or FinBERT (see below). Clearly the output for BERT and FedBERT-BERT are suboptimal, where vocabulary from other corpus have had an effect on model predictions, so while their use is wide ranging, when it comes to specific tasks they fail compared to either FinBERT or FedBERT-Fin. FinBERT performed really well, but instead of financial system, it uses words from corporate word like 'delivery' system, something that could be seen in an Amazon quarterly report. FedBERT-Fin combines the best of both worlds, and it seems like the small number of training steps did not affect the performance.

> Original: And, while many of these changes have improved the efficiency of our **financial** system and lowered **costs** for **consumers**, it is only realistic to1 acknowledge that they also present new and sometimes daunting **tests** for community **banks**
>
> BERT: And, while many of these changes have improved the efficiency of our **financial** system and lowered **costs** for **consumers**, it is only realistic to1 acknowledge that they also present new and sometimes daunting **tests** for community **banks**
>
> FedBERT-BERT: And, while many of these changes have improved the efficiency of our **vote** system and lowered **lower** for **vote**, it is only realistic to1 acknowledge that they also present new and sometimes daunting **.** for community **.**
>
> FinBERT: And, while many of these changes have improved the efficiency of our **delivery** system and lowered **costs** for **residents**, it is only realistic to1 acknowledge that they also present new and sometimes daunting **challenges** for community **development**
>
> FedBERT-Fin: And, while many of these changes have improved the efficiency of our **financial** system and lowered **costs** for **consumers**, it is only realistic to1 acknowledge that they also present new and sometimes daunting **challenges** for community **banks**
>
> FedBERT-Prime: And, while many of these changes have improved the efficiency of our **financial** system and lowered **.** for **example**, it is only realistic to1 acknowledge that they also present new and sometimes daunting **areas** for community **banks**

4.3 Cosine similarity

One of the ways to assess the performance of the models is to use cosine similarity to gauge their understanding in embedding words with different contexts. For example, for the following two sentences, the meaning for the word 'bank' changes.

> The man was accused of robbing a bank
>
> The man went fishing by the river bank

If FedBERT has learned from its corpus, we should see that it is able to distinguish between these two banks, since bank is a usual suspect for Fed speeches. Indeed it is the case, when we compute the cosine similarities between the word 'bank' across different models, the cosine similarity between the word bank decreases for the FedBERT models, suggesting that the FedBERT models has learned when the financial institution bank is used in different contexts than river banks. Similarly, we provide the following sentences to validate this behaviour

> the existence of substantial economic shock
>
> the man being hit by the car is in shock

> the fed is promoting its long term price stability target
>
> it is the most purchased item at his local target

The effect is carried over through different models, with BERT the least able to distinguish the words, and FinBERT an improvement due to its training in financial datasets, and FedBERT performing best since the training corpus uses all three words heavily.

Table 3: Cosine similarity of same words across models

Model	bank	shock	target
BERT	0.97	0.92	0.95
FedBERT-BERT	0.89	0.86	0.89
FinBERT	0.91	0.8	0.88
FedBERT-Fin	0.78	0.7	0.76

5 Future work and conclusions

In this paper, we apply transfer learning to BERT using Federal Reserve Board text data. We show that the model, combined with pretrained models such as BERT or FinBERT works well in understanding Fed-specific languages. We find the pre-trained model improves the cross-topic utility of the model, and the pretrained model corpus has a big impact on model performance in out-of-sample scenarios. We perform cosine similarity analysis and masked-out word analysis to conclude that the performance of FedBERT is superior to the baseline models in Fed specific tasks and understanding, including the ability to differentiate words of the same meaning better in different contexts.

We acknowledge the shortcoming of our model, including the lack of next sentence prediction and its shortcomings including out of sample performances as masking probability increases. This model can be extended in many directions, including adding a bottleneck layer to extract features that might predict the next Fed action.One way to do this would be to train the model on only the data available before a fed meeting and use a simple sentiment analysis model to predict the interest rate decision. However, it should be noted that this process would require training the model several times, which would be computationally expensive. Ultimately, if a model can predict Fed action better in today's investment climate, it is more equipped to act more forcefully in pursuit of alpha.

References

Vinicio Desola, Kevin Hanna, and Pri Nonis. Finbert: pre-trained model on sec filings for financial natural language tasks. 08 2019. doi: 10.13140/RG.2.2.19153.89442.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, Sep 2019. ISSN 1460-2059. doi: 10.1093/bioinformatics/btz682. URL http://dx.doi.org/10.1093/bioinformatics/btz682.