Ada Kilic
Professor Kaisler and Professor Hasanov
11 July 2023

Midterm Report

SUMMARY

This research project aims to build an index that measures the overall sentiments in one of the most important US economic policy institutions, namely the FOMC meeting statements. The objective is that such an index can be used to learn overall sentiment by the policy makers on the economic and market conditions. I will call this index the FOMC Sentiment Index, (FOMC-SI) and explore its usefulness in providing information about behavior of some economic and market variables via regressions. To this end, the project aims to use the text data based on FOMC statements over time and build an index value for each statement since January 2000.

The sample of statements for this project covers all the available statements made by the FOMC since 2000. Prior to 2000, Chairs and FOMC did not issue a press statement. Since there are officially 8 FOMC meetings in a year, with some years more than 8 meetings depending on the developments in the economy, the data collected has 202 statements corresponding to the period between January 2000 and June 2023. These statements are publicly available and can easily be downloaded from https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm.

The data compilation is completed for this project and data is put in an excel. The preprocessing for this research is also completed. The statements are preprocessed by removing excess white spaces, hyperlinks, and quotation marks. Conceptually, FinBERT model is studied by reading the relevant literature and work is underway to develop a Python notebook to implement FinBERT in the FOMC statements data set in order to develop the FOMC-SI.

PROJECT PROGRESS UPDATE

Data Collection & Cleaning

My primary focus for the first half of the project has been collecting the data, cleaning the data, and learning FinBERT model architecture in order to apply it correctly on my project. As I have explained above the data collection and cleaning phase is completed. After my data collection, I cleaned the data. Cleaning and preprocessing text data is an important step before performing sentiment analysis. I started by converting all the text to lowercase. This helps to avoid having

multiple copies of the same words. I also discovered that I do not need stemming and lemmatization. These processes reduce words to their root form. I learned that BERT models are designed to understand word context without the need for stemming or lemmatization. Then I will explore tokenization. This process breaks the text into individual words or tokens. I also pre-processed the text of FOMC statements by removing excess white spaces and hyperlinks. It looks like statement texts are generally pretty standard and not much cleaning/process is needed in contrast to say twitter texts.

FinBERT & BERT

In addition to data collection and cleaning, I have read a number of papers that explains and discusses BET and FinBERT including the paper by Araci (2019) which is the original paper that discusses FinBERT in detail. Below I provide an overview of FinBERT and BERT based on my understanding of the relevant papers.

FinBERT is designed to conduct sentiment analysis of Financial text based on pre-trained language models. FinBERT is based on BERT (Devlin et al., 2018). BERT is a state-of-the-art pre-trained machine learning model capable of understanding sentences alongside the context in which they are being applied. BERT is pre-trained on the Toronto BookCorpus (containing 800M words) and Wikipedia articles (containing 2.5B words). BERT converts words into vectors, and reads the text bidirectionally to classify sentences given the context in which words are being used. This unique ability to understand contextual representation, and doing so in both directions of the text allows BERT to significantly outperform other machine-learning-based and dictionary based models in tasks like text prediction and sentiment calculation. Furthermore, it can be pre-trained further and then fine-tuned to better understand a desired context, like financial jargon.

FinBERT uses attention masks to mask the padding tokens in sequences that are shorter than the maximum sequence length in the batch. The mask is an array of 1s and 0s, with 1s for the actual tokens in the sequence and 0s for the padding tokens. This is automatically handled by the BERT tokenizer which is part of preprocessing.

The BERT model goes through 2 unsupervised learning steps. These are Mass Language Modelling (MLM) and Next Sentence Prediction (NSP). In MLM the model is given a random sentence with a few words masked or replaced by a token. BERT then tries to predict the embeddings for these words. In this way it learns the language and grammar. In NSP step BERT is given 2 sentences, Sentence A and Sentence B and told to predict whether sentence A precedes sentence B in order. This way BERT learns the context of the language.

BERT uses a method called WordPiece tokenization. This involves breaking words into smaller subwords. This is especially useful for dealing with out-of-vocabulary words and enables the model to handle almost any word it encounters, even if it's not in the model's vocabulary. I used the transformers library from Hugging Face to load the tokenizer that corresponds to the pre-trained model I am using

FinBERT, from Araci (2019), is a refined version of BERT that is designed to understand text in the context of Financial sentiment. FinBERT is pre-trained using a large corpus of financial texts and fine-tuned with a dictionary of financial words and phrases from Malo et al. (2014). One feature of FinBERT is that it was pre-trained using longer texts, so it splits sentences individually and then calculates sentiment on each one of them. This feature generally fits well with our data as our FOMC statements generally contain text for each statement date that has several sentences.

FinBERT produces five sentiment values. Three values represent the probabilities that the text is either positive, negative, or neutral. FinBERT also calculates a compound score as the positive probability minus the negative probability. Lastly, FinBERT provides trinary sentiment prediction which is based on the highest of the three probabilities.

FOMC-Sentiment Index

I plan to use FinBERT's compound score to assign a numeric value of sentiment for each sentence in a FOMC statement. FinBERT provides three probabilities to measure the odds that the analyzed text conveys positive, natural, or negative sentiment and also offers a compound sentiment score computed as the difference between the probability of the test having positive sentiment and the probability of the text having negative sentiment. Therefore, FinBERT should provide us with a sentiment score between -1 and +1 for each sentence in our sample of FOMC statements for a given statement date.

For the purpose of measuring the evolution of average sentiment for a statement in date, first a sentiment score of zero will be assigned to sentences that are labeled as neutral and then sum the sentiment score (compound score) of all remaining sentences in a statement and divide by the total number of sentences in the statement. For example, suppose that there are *n* sentences in a statement for a given date *t.* I will sum the sentiment score (compound score) of all sentences, excluding sentences that are predicted by FinBERT as "neutral" with a probability of say 0.5 and above, and divide this sum by the total number of sentences in the statement.

As an example, for the FOMC statement dated June 14, 2023, there are 15 sentences. My initial exploration suggests that FinBERT predicts that 11 of these sentences are “neutral” with a probability score above 0.5 and 3 scores are “positive” and 1 sentence is “negative. Ignoring the neutral sentences with probabilities above 0.5, and using only the probabilities for “positive” and “negative” predictions, I calculated the average FOMC-SI value to be 0.811. I can multiply this number by -1 to get -0.811 to reflect that lower values indicate more “positive” tone and hence, positive sentiment in the FOMC statements.

This proposed approach can allow me to average sentiment across time, say over the first half of say 2000 and compute FOMC sentiment values at lower time frequencies such as quarterly or semi0-annually etc. in addition to the FOMC meeting-specific sentiments..

Also note that in principle, the value of the FOMC sentiment index would vary by changes in the sentiment value of each sentence (i.e., intensive margin) in the statement or by the share of sentences with positive or negative sentiment (i.e., extensive margin). Therefore, I can also compare the above FOMC sentiment index with the share of positive vs. negative sentences for each meeting to understand to what degree the sentiment is driven either by the changes in the sentiment value of each sentence or by the share of sentences with positive or negative sentiment values.

Also I will use data visualizations as well. I will use a time series plot of sentiment index. This can be used because I am creating a historical sentiment index. This can show how sentiment changes over time. The x-axis can represent time (from 2000 to present), and the y-axis can represent the sentiment index. Each point on the graph represents the average sentiment value of each FOMC statement. Also I plan to use a heatmap of sentiment over time. A heatmap can help me show the sentiment of each FOMC meeting over the course of a year. The x-axis could represent months, the y-axis could represent years from 2000 to present. Finally, the color could represent sentiment with a gradient from negative to positive.

CHALLENGES AND SOLUTIONS

Some challenges I encountered was making sure that each of the sentences in the statements are evaluated. The statements in the column contain multiple sentences since I want to apply FinBERT on each sentence separately while keeping track of the date. I first split the statements into individual sentences in order to apply FinBERT on each sentence. The sent_tokenize function from the NLTK library that splits text into sentences. It uses a machine learning model that has been trained to recognize sentence boundaries in English text. This model takes into account a variety of factors, including punctuation (like periods, question marks, and

exclamation points), capitalization, and certain multi-word phrases that indicate the end of a sentence. This helped me ensure that each sentence is evaluated in FinBERT.

FUTURE WORK & CONCLUSION

My work is underway to develop a notebook that can be used to implement FinBERT in my data in order to calculate FOMC-SI for each FOMC meeting date. Once I complete the notebook and implement FinBERT in my data, I will construct the FOMC-SI and explore the index over time to link it to some major economic events such as COVID-19 induced recession in 2020 and the Great Financial Crisis during 2007-2008. If time left, I also plan to use FOMC-SI to link it to some key economic variables such as economic activity including GDP growth or stock market performance measures via regressions.

REFERENCES

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. CoRR abs/1908.1006. https://arxiv.org/pdf/1908.10063.pdf

Devlin, J., M. Chang, K. Lee, and K. Toutanova (2018). BERT: pre-training of deep bidirectional transformers for language understanding. CoRR abs/1810.04805.

Malo, P., A. Sinha, P. Takala, P. Korhonen, and J. Wallenius (2014, 04). Good debt or bad debt: Detecting semantic orientations in economic texts. Journal of the American Society for Information Science and Technology.