Sentiment Analysis Comparison of Federal Open Market Committee Press Releases and Meeting Minutes Leveraging Available Financial Language Models

Alexa Bagnard, Erin Stanton
abagnard@berkeley.edu, elstanton@berkeley.edu
U.C. Berkeley

December 6, 2020

Abstract

The Federal Open Market Committee (FOMC) is the branch of the Federal Reserve system (FED) that sets monetary policy in the United States based on their analysis of current global economic conditions and forecasts. The monetary policy decisions made in these meetings can have a significant impact on the US economy and financial markets. As such, investors tend to pay close attention to the press release statements (release notes) and meeting minutes put out by the FOMC following each meeting as these documents can provide insight into future economic and financial market health. Our initial hypothesis is that existing financial language models, including the Loughran and McDonald (L-M) sentiment lexicon and the FinBert model, which are trained off of general financial lexicon, do not perform as well given the specific language that the FED uses. Our research demonstrates that a fine-tuned FinBert and a fine-tuned BERT model which are directly trained off of the FOMC release notes and meeting minutes results in a better sentiment classification.

Keywords— n atural language processing, textual analysis, sentiment, federal reserve, FED, Federal Open Market Committee, FOMC, finance, BERT, FinBERT

1 Introduction

The Federal Open Market Committee (FOMC) is an important part of the United States and global economy. The Committee, made up by 12 members of the Federal Reserve (FED), meets at least eight times per year to discuss current economic conditions, forecast future economic conditions and decide the monetary policy in the US. It is in these meetings that the FOMC will set interest rates, create regulation and forecast future US economic conditions in hopes to maintain the stability of the nation's economy and financial system. After each meeting, the FOMC will release multiple forms of communication to the general public, including, but not limited to, press releases (release notes), meeting minutes, meeting transcripts and presentation materials. It is in these documents that the public can gain insight into the FOMC's sentiment on the US and global economy.

There are several resources available for analyzing the sentiment of financial documents which include the Loughran and McDonald (L-M) sentiment lexicon as well as the FinBERT model. Our hypothesis is that the FOMC has it's own specific lexicon that is not captured effectively by these more generalized financial language models. As part of our research we first looked at the sentiment captured by the L-M sentiment lexicon and the pre-trained FinBERT model and then further trained our own sentiment models based on a selection of the FED's public documents and a constructed label of the market's reaction.

2 Approach

2.1 Data

There were two main sources of data required for our research - FOMC communication documents and historical market data.

The FED's website, www.federalreserve.gov, includes public FOMC communication documents going back to the 1930. These documents include press release statements (release notes), meeting minutes, presentation documents and full meeting transcripts. Unfortunately, the format and online location of the published statements and minutes did not maintain a consistent format over time so a multi-faceted HTML text scraper was required to gather all of the statement and meeting minute data. We cap our historical FOMC data to 1993 based on a change invoked to provide even more transparency which resulted in a shift in overall meeting format and dialogue.

After scraping the online document data, we removed white spaces, punctuation and stop words (from nltk.corpus.stopwords) to create the data set used to train and evaluate our models.

Historical market data, specifically the SP500 index and 10-year treasury yields, was gathered using the Alpha Vantage https://rapidapi.com/alphavantage/api/alpha-vantage and Quandl APIs https://www.quandl.com/data/USTREASURY-US-Treasury, respectively. This data is used to construct the labels needed to train and compare the models analyzed in this project.

2.2 Labels

In order to measure the effectiveness of the FinBERT and BERT models, we constructed a label, based on changes in the SP500 index and 10-year treasury yields, in order to gauge the sentiment of the FED release notes and meeting minutes. As it is generally hard to isolate a singular market event, pre- and post- meeting date averages were taken to measure changes in general market movement. Please refer to 6 for additional details on the construction of our label.

Two issues were encountered when it came to constructing the label.

- The first was trying to come up with a measurement that captured the market's reaction to the FED's release notes and meeting minutes. The market can be affected by dozens of factors that are outside of the FED itself as demonstrated in Figure 1 which shows the SP500 over time; many of the sharp inclines and declines were independent of FED announcement dates.

Figure 1: SP500 index returns over time

- Another issue was the class skew of the label. As seen in Figure 2, the majority of labels were neutral. As a result, we split our test and validation data sets using stratified random sampling so that each data set best represents the sentiment class breakdown of all of the documents being

analyzed. The stratified random sampling process that we applied divided our documents into smaller groups based on their market label.

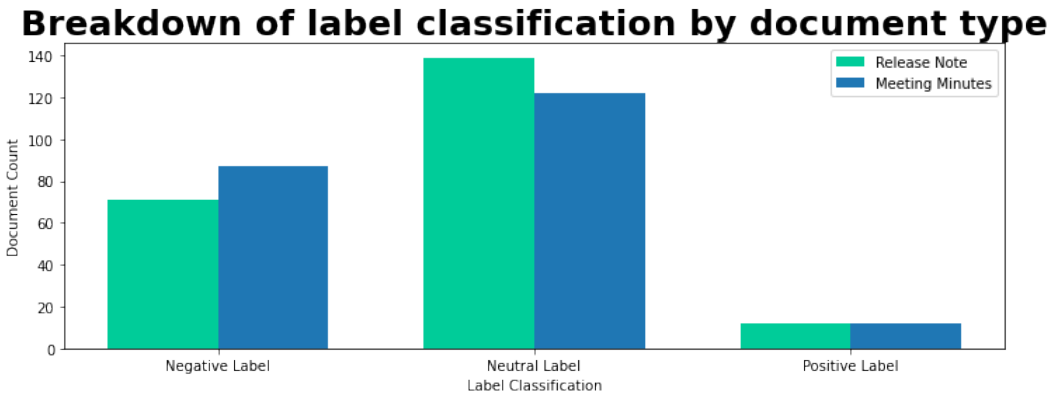

Figure 2: Label Classes

3 Models

In this project, we compared accuracy and weighted-F1 scores to determine which text-based sentiment prediction model is best at determining the FOMC's economic sentiment.

3.1 Loughran-McDonald Sentiment

The Loughran-McDonald (L-M) sentiment lexicon is the most popular financial lexicon used in natural language processing (NLP). This lexicon assigns positive or negative sentiment values to thousand of words appearing in financial document such as 10Ks, 10Qs and earning calls.

As a cursory effort to better understand if there were any language differences between the release statement and meeting minutes, we used the word cloud visualization on the top positive and top negative words as classified by the L-M sentiment lexicon.

As seen in Figure 3 and Figure 4, there is a clear negative language skew for the meeting minutes as compared to the initial release note.

Figure 3: Loughran-McDonald Sentiment, Top Negative and Positive Words for the Meeting Statements

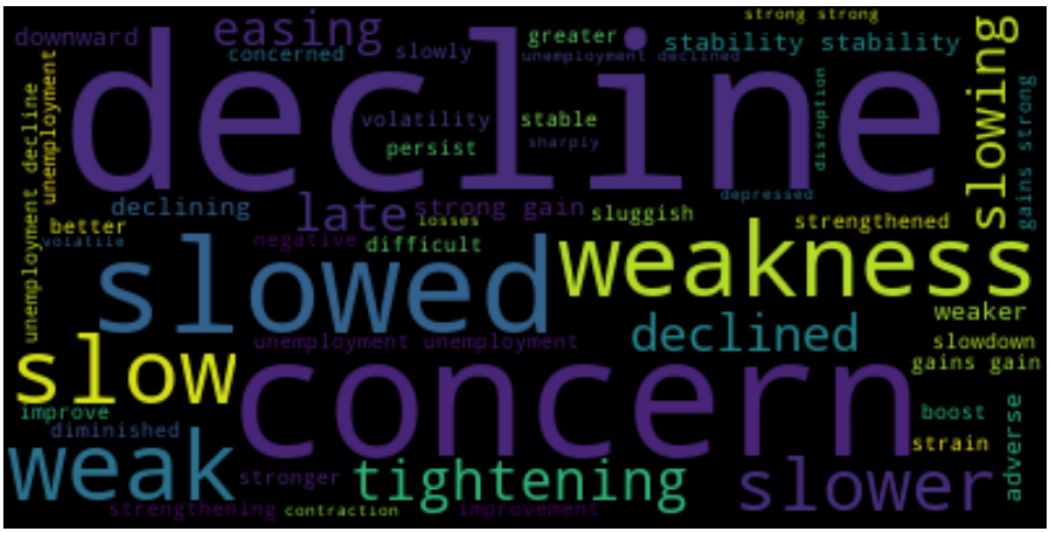

Figure 4: Loughran-McDonald Sentiment, Top Negative and Positive Words for the Meeting Minutes

Following the initial visual interpretation of lexicon used in the two document types, we further analyzed if there was an inherent sentiment bias between the release statement and meeting minutes. Continuing to leverage the L-M sentiment lexicon we calculated an aggregate sentiment score for each document, adding a value of one for each positive word and subtracting a value of one for each negative word. Figure 5 tracks the changes in the calculated L-M sentiment scores since 1993 by document type. This graphs shows a clear difference in sentiment scores between the release statements and meeting minutes; FOMC release statements have a more neutral sentiment, with a sentiment score closer to zero, while the associated FOMC minute notes have a negative sentiment skew.

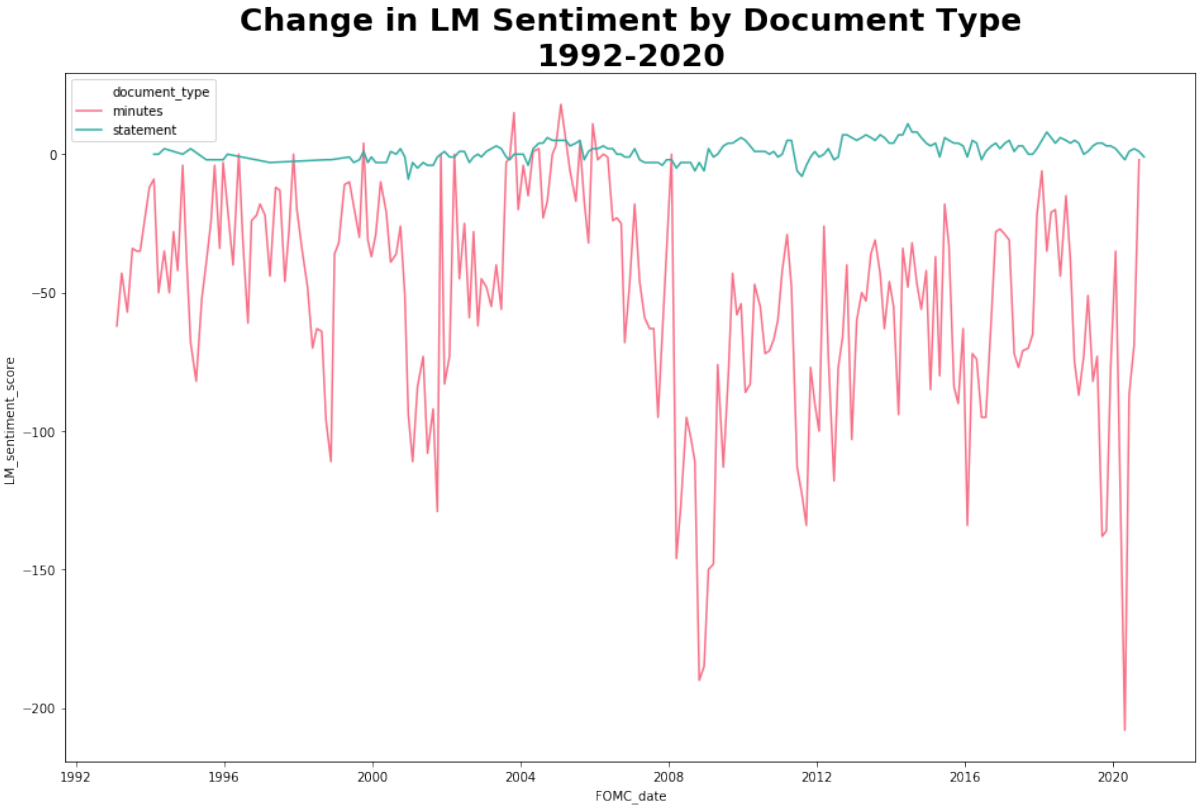

Figure 5: Loughran-McDonald Aggregate Sentiment per Entire Document

This skew is not entirely a surprise, as like many organizations, press release statements tend to be

shorter and include more conservative language. The meeting minutes, on the other hand, are typically released three weeks after the FOMC meeting and are much longer than the press release statements (on average 4,100 words vs 510 words, respectively).

An additional consideration for the initial sentiment comparison between the release notes and meeting minutes is that both the BERT and FinBERT models can only accept 512 tokens. While many of the press release statements were contained within 512 tokens with stop words removed, the FOMC meeting minutes contained an average of 4,100 words, significantly more than BERT's token limit. As a result, we decided to take the middle 512 words of each minute document and use these smaller sections to train our models. Our concern with splitting the longer meeting minutes into multiple 512-chunks is that we're only using a single label and different parts of the meeting would carry contrasting sentiments.

As a result, we recalculated the L-M sentiment score for each document using the smaller 512-length subsections. Figure 6 tracks the change in this new L-M sentiment score for the subset of each document over time. With the revised 512-document length, both the release notes and meeting minutes showed a more varied sentiment outcome. It should be noted that we're assuming the market interpreted the real sentiment put out by the FED on the actual meeting date so the label for both the release notes and meeting minutes is associated with the original press release date.

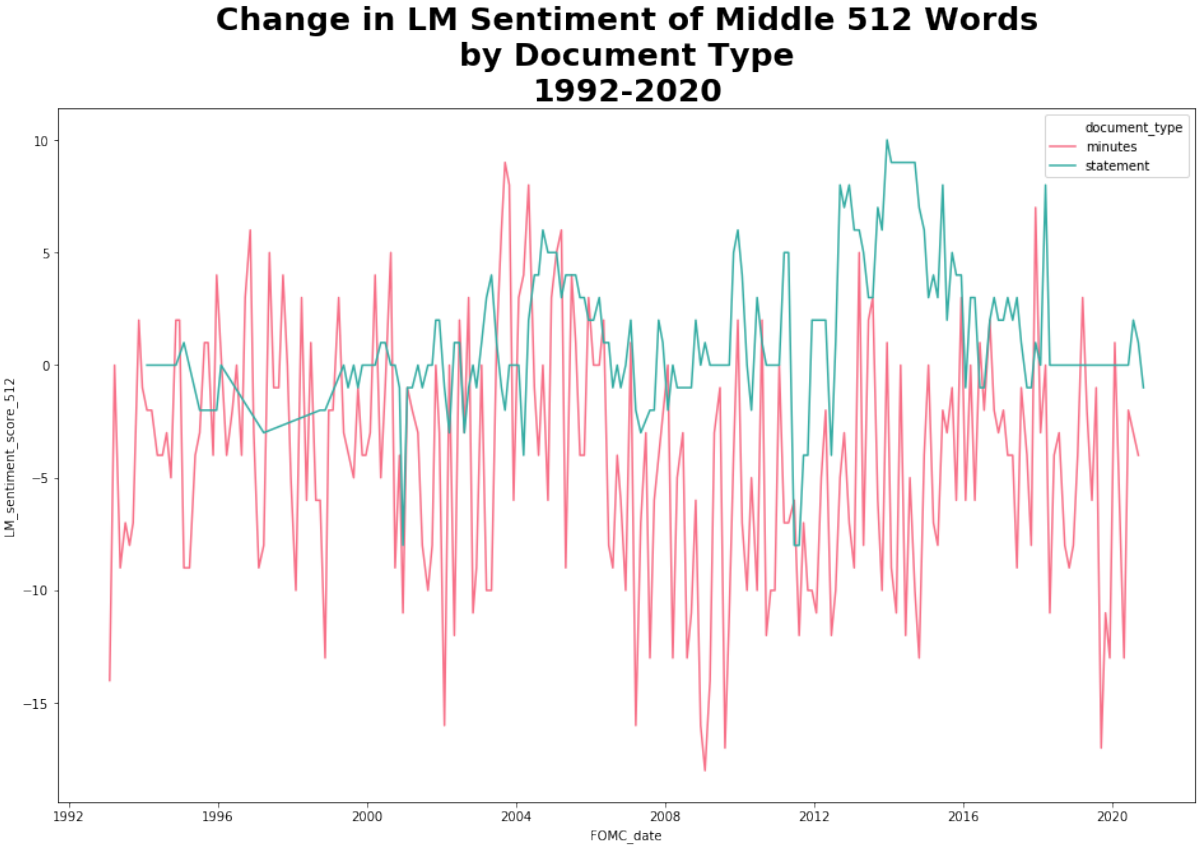

Figure 6: Loughran-McDonald Aggregate Sentiment per Document, middle 512 document subsection

To complete our analysis of the L-M sentiment lexicon on the smaller 512-length subsections, we prepared a confusion matrix comparing the L-M document sentiment label versus our market label. As seen in figure 7, the L-M sentiment lexicon does a poor job when the market captures a neutral sentiment reaction compared to a negative document classification. To further support the confusion matrix results, we calculated the accuracy and weighted-F1 score of the L-M sentiment lexicon model, 40% and 42%, respectively.

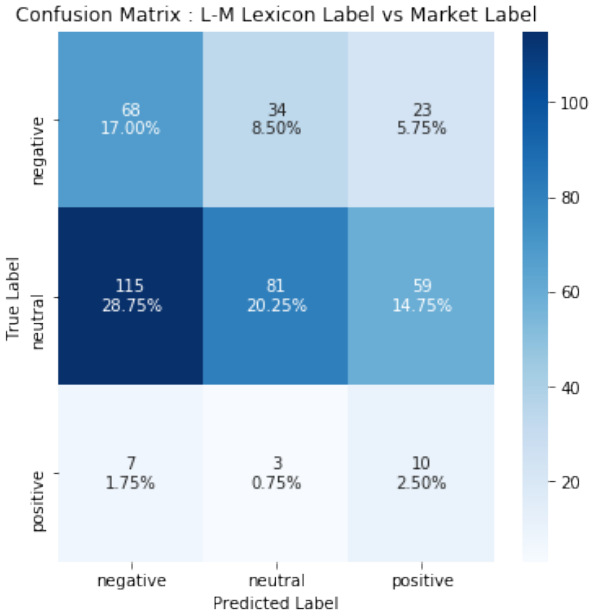

Figure 7: Confusion Matrix: L-M Sentiment Lexicon vs. Market Label, All Documents

	precision	recall	f1-score	support
negative	0.36	0.54	0.43	125
neutral	0.69	0.32	0.43	255
positive	0.11	0.50	0.18	20
accuracy			0.40	400
macro avg	0.38	0.45	0.35	400
weighted avg	0.55	0.40	0.42	400

Figure 8: Classification Report: L-M Sentiment Lexicon vs. Market Label, All Documents

3.2 FinBERT, out of the box

In 2019, Dogu Araci, introduced FinBERT, a language model based on BERT to tackle natural language processing (NLP) tasks specific to the financial domain. Financial sentiment analysis is a challenging task due to the specialized language and lack of labeled data. As a result, general purpose models are not as effective at determining sentiment within financial statements. Since the FOMC discusses a variety of finance related topics in each meeting, for example interest rates, we decided to use FinBERT to predict FOMC meeting sentiment. The FinBERT model that we leverage within our research is "ipuneetrathore/bert-base-cased-finetuned-finBERT."

We first ran the pretrained version of FinBert without any fine-tuning. As seen in Figure 9, the FinBERT model best categorized the neutral label, correctly labeling 48% of observations as neutral. The documents the FinBERT model had the hardest time categorizing were those with a negative sentiment misclassifying most of these documents as having a neutral sentiment.

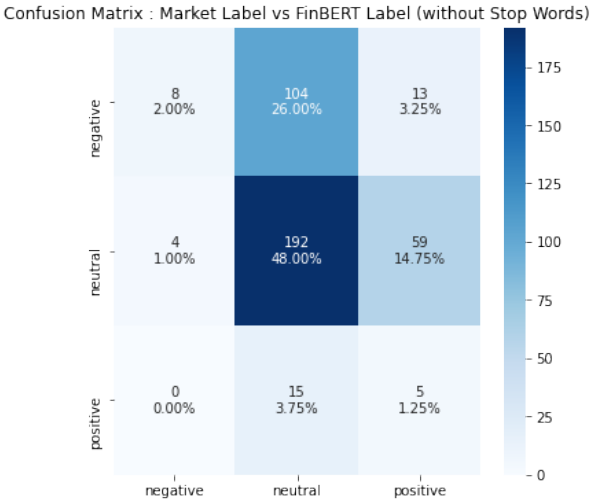

Figure 9: Confusion Matrix: FinBert vs. Market Label, All Documents

With an accuracy score of 51% and a weighted-F1 score of 47%, as seen in Figure 10, overall, the out of the box FinBERT model does a better job at determining sentiment within the FOMC meeting documents compared to using the L-M sentiment lexicon. However, for documents with a negative sentiment label, which make up over 31% of the total documents, the pretrained FinBERT model struggles, with a resulting recall score of only 6%.

	precision	recall	f1-score	support
negative	0.67	0.06	0.12	125
neutral	0.62	0.75	0.68	255
positive	0.06	0.25	0.10	20
accuracy			0.51	400
macro avg	0.45	0.36	0.30	400
weighted avg	0.61	0.51	0.47	400

Figure 10: Classification Report: FinBert vs. Market Label, All Documents

3.3 FinBERT with fine-tuning

After running the pretrained, out of the box, FinBERT model, we wanted to see if we could improve the accuracy and f1-score by further tuning the model. To do this, we used the Simple Transformer library to create a BERT classification model with the pretrained FinBERT architecture.

We fine-tuned the model by altering various parameters including: batch size, number of epochs, learning rate, early stopping and class weights. In addition to tuning model parameters, we also split our test and validation data sets using stratified random sampling, which also had a large impact on the model's results.

Compared to the out of the box FinBERT model, the fine-tuned model was more successful in classifying the FOMC meeting documents. Figure 11 clearly shows that the fine-tuned FinBERT model is much better at classifying documents with negative or neutral sentiment.

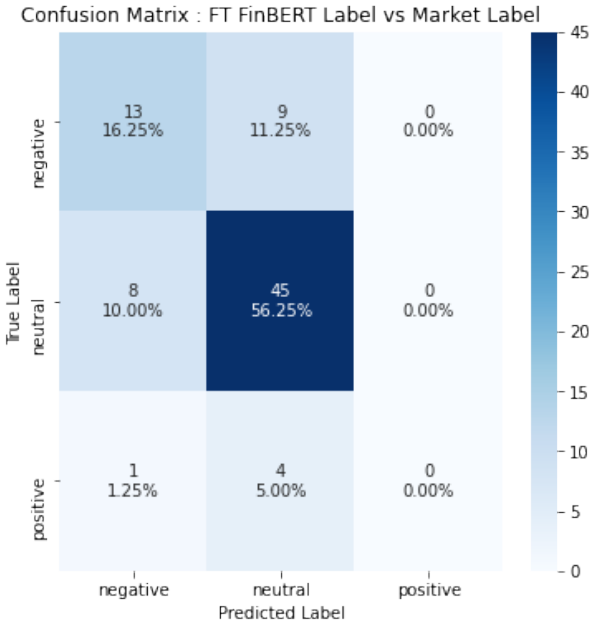

Figure 11: Confusion Matrix: Fine-tuned FinBert, Test Data Subset

Figure 12 provides additional results of our fine-tune FinBERT model. Overall, the fine-tuned FinBERT model had a weighted f1-score of 70% and an accuracy score of 73%. This is a significant improvement to the pretrained FinBERT model that had a weighted-f1 score of 47% and an accuracy score of 51%. In addition, when compared to the pretrained model, the f1-score for neutral labels improved, from 68% to 81% and more noticeably, the negative label class f1-score improved from 12% to 59%. Unfortunately, this model was not as good at classifying documents with positive sentiment, however, this could be explained by the small number of documents with a positive label in the validation set.

	precision	recall	f1-score	support
negative	0.59	0.59	0.59	22
neutral	0.78	0.85	0.81	53
positive	0.00	0.00	0.00	5
accuracy			0.73	80
macro avg	0.46	0.48	0.47	80
weighted avg	0.68	0.72	0.70	80

Figure 12: Classification Report: Fine-tuned FinBert, Test Data Subset

3.4 BERT with fine-tuning

The last model that we explored was a BERT model with fine-tuning applied. The model used here is pretrained BERT-based-uncased model, which is a deep neural network with 12 layers, 768 hidden units, 12 heads that was trained on the Wikipedia and BooksCorpus (not financial corpora).

In addition to covering finance related topics, like the stock market and interest rates, the FOMC reviews a variety of other economic indicators to determine future monetary policies. These topics include, but are not limited to, unemployment rates, geopolitical relationships and regulations. As such, we decided to fine-tune a BERT model and compare its results to the three prior models that were all trained on financial corpora.

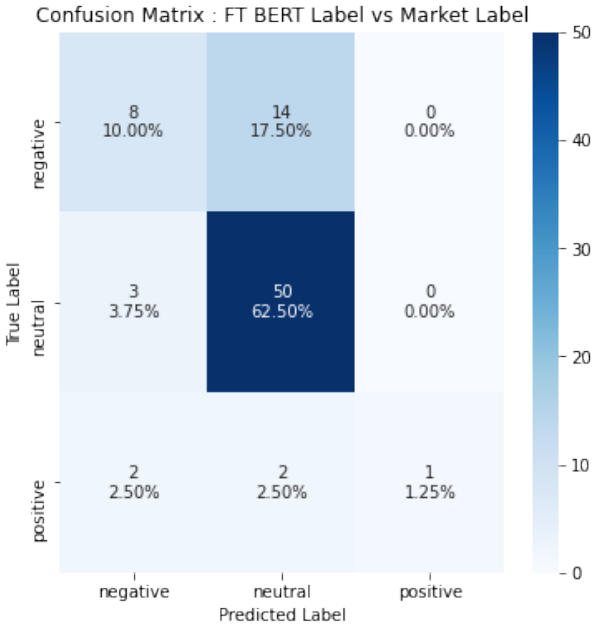

Figure 13: Confusion Matrix: Fine-tuned BERT, Test Data Subset

The fine-tuned BERT model produced very similar results as the fine-tuned FinBERT model. The accuracy score improved slightly (74% vs 73%), while the weighted f1-score remained the same at 70%. The fine-tuned BERT model did slightly better at classifying positive and neutral labeled documents, with f1-scores increasing from 0% to 33% and 81% to 84%, respectively. However, the BERT model did worse at predicting negative labeled documents as the f1-score dropped from 59% to 46%.

	precision	recall	f1-score	support
negative	0.62	0.36	0.46	22
neutral	0.76	0.94	0.84	53
positive	1.00	0.20	0.33	5
accuracy			0.74	80
macro avg	0.79	0.50	0.54	80
weighted avg	0.73	0.74	0.70	80

Figure 14: Classification Report: Fine-tuned BERT, Test Data Subset

4 Results

The following figures show the comparison of scores across the four models that we analyzed. While the overall weighted f1-score was the same for both the fine-tuned FinBERT model compared to the fine-tuned BERT model, we believe that the improvement in the positive label class vs. the similar negative/neutral label class scores makes the fine-tuned BERT model the best fit for the sentiment analysis for FOMC language documents.

Model	Weighted Precision	Weighted Recall	Weighted F1-Score	Accuracy
L-M Lexicon	0.55	0.4	0.42	0.4
Pretrained FinBERT	0.61	0.51	0.47	0.51
Fine-tuned FinBERT	0.68	0.72	0.7	0.73
Fine-tuned BERT	0.73	0.74	0.7	0.74

Figure 15: Model Comparison Results

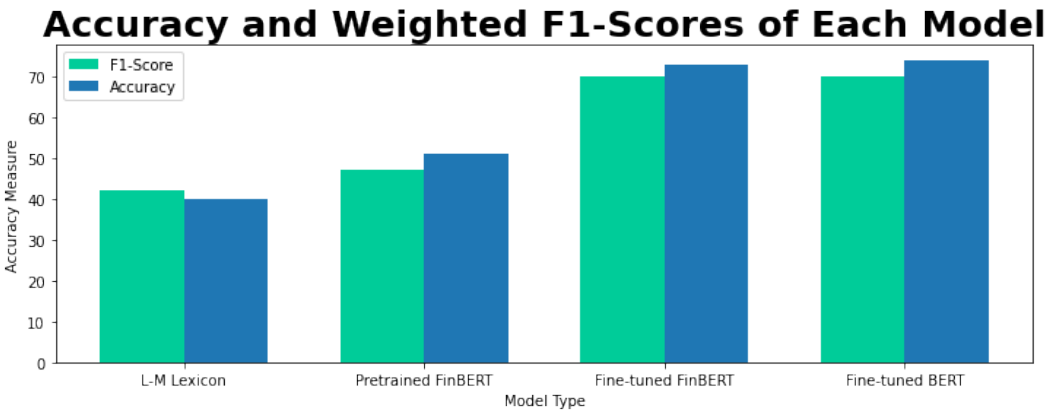

Figure 16: Accuracy and Weighted F-1 Scores of Each Model

5 Conclusion

In this project, we examined how closely the the sentiment of the FOMC's statement and minute documents related to market movements. To do this, we compared metric results between the Loughran-McDonald (L-M) sentiment lexicon, a pretrained FinBERT model, a fine-tuned FinBERT model and a fine-tuned BERT model.

Our initial hypothesis was that the FED had it's own 'language' which was different enough from general financial language lexicon. We saw poor sentiment classification results leveraging two of the most popular financial sentiment models; the Loughran and McDonald (L-M) sentiment lexicon and the pretrained FinBert model. Training our own FinBERT and BERT model using the specific FOMC release notes and meeting minutes resulted in a final sentiment analysis that both reduced document classification confusion and improved our overall class-specific weighted f1-score.

While we were able to improve the FinBERT accuracy score by 45%, we could not improve the accuracy score to more than 74%. We believe this is partly due to the following reasons:

- Lack of Training Data: Due to inconsistencies in the documents published by the FOMC, we were only able to use press release statements and meeting minute notes going back to 1993. As a result, our data set included 400 documents in total, 320 of which made up our training data. While we applied stratified random sampling to create our training data set and tuned model parameters, the models continued to overfit the training data and did not generalize well. In addition, some of the FOMC documents used in this analysis were significantly longer than BERT's 512-token limit, so we were unable to use the full documents to determine sentiment.

- Sentiment Classification Labels: FOMC documents do not come with a predetermined sentiment label. As a result, we created our own label, using the SP500 index and 10-year treasury yields, that would capture the market's reaction to the FOMC meeting documents. Unfortunately, financial markets can be affected by a variety of factors unrelated to the FED and FOMC, which can impact the accuracy of our classification labels. Additionally, we had a clear label imbalance in our data set, in which 64% of the documents were labeled as having neutral sentiment despite the fact that the SP500 index has gone up over 3,000 points since 1993.

In the future, we hope to improve our analysis by increasing the data set size and determining a more accurate FOMC sentiment label.

6 Appendix - Labels

The constructed label includes a combination of the SP500 index and 10-year treasury yields.

For the SP500 portion:

- An average of the close price on release note date - 2, -3, and -4 was compared to an average of the close price on release note date - 1. A second average of the the open price on release note date, close price on release note date, close price on release note date + 1 and close price on release note date + 2 was compared to the close price on minute release date + 3.
- If the comparison between the two metrics was positive then a positive value was passed into the label construction portion. If the comparison metric was negative then a negative value was passed into the label construction process.

For the label construction portion:

- If both the SP500 comparison metric was positive and the change in the 10-year treasury yield was positive then the label = 2 (BERT models require positive only-labels).
- If the SP500 comparison metric was negative and the change in the 10-year treasury yield was positive then the label = 1.
- If the SP500 comparison metric was positive and the change in the 10-year treasury yield was negative then the label = 1.
- If the SP500 comparison metric was negative and the change in the 10-year treasury yield was negative then the label = 0.

7 Appendix - Meeting Note/Minutes Text Scraping

To pull the FOMC's statements and meeting minutes from the FED's website, we used the library BeautifulSoup. Once we extracted the content from each specific URL, we selected a subset of the extracted subset based on the document type, to remove any website header and footer text. We then saved this data into .txt files.

After extracting and saving the FOMC's document data, we then cleaned the data and saved the results as a new .txt file. First we forced the text to be lowercase. Next, we removed all extra white space and punctuation. Lastly, we normalized the various names of the FOMC by renaming "federal open market committee" and "the committee" to "FOMC".

After cleaning the FOMC's document data, we then removed all the stop words listed in nltk.corpus.stopwords and saved this data as a new .txt file.

References

[1] Dogu Arac. *FinBERT: Financial Sentiment Analysis with Pre-trained Language Models*. University of Amsterdam, 2019.

[2] Adam Hale Shapiro and Daniel Wilson. *Taking the Fed at its Word: A New Approach to Estimating Central Bank Objectives using Text Analysis*. Federal Reserve Bank of San Francisco, 2019.

[3] Sam Cannon. *Sentiment of the FOMC: Unscripted* . Federal Reserve Bank of Kansas City, 2015.

[4] Sentiment Analysis Of FOMC Statements Reveals A More Hawkish Fed, *https://www.forbes.com/sites/alapshah/2017/09/22/sentiment-analysis-of-fomc-statements-reveals-a-more-hawkish-fed/*

[5] Aharish Gandhi Ramachandran and Dan DeRose Jr. *A Text Analysis of Federal Reserve meeting minutes*. Illinois Institute of Technology.

[6] Simple Transformers, *https://simpletransformers.ai/*

[7] Federal Open Market Committee Meeting Documents, *https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm*

[8] Alpha Vantage, *https://rapidapi.com/alphavantage/api/alpha-vantage*

[9] Quandl APIs, *https://www.quandl.com/data/USTREASURY-US-Treasury*