Ada Kilic

Report 2 (deadline: Jun 26, 23:59):

Research Strategy:

For my research I am primarily following a quantitative strategy. This involves using FinBERT for sentiment analysis of the text data. This is a quantitative technique because it will obtain results in numerical scores representing the sentiment conveyed in the text.

The project aims to build an FOMC sentiment index that relies on sentiment analysis of the text data of Chair's statements post FOMC meetings.

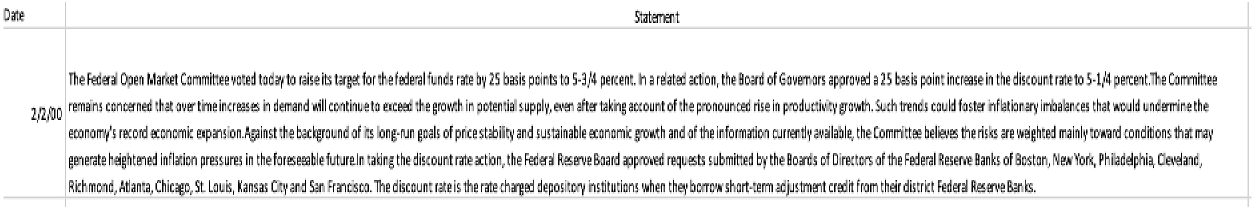
For each sentence in each statement in the sample of policy statements, I plan to measure sentiment using FinBERT, a language model developed by Araci (2019) from BERT (Devlin et al., 2018). FinBERT is specifically designed to measure sentiment of financial text. I plan to average sentiment values of each sentence in each of the policy statements and hence, construct a historical index of the FOMC sentiment index.

BERT is a state-of-the-art pre-trained machine learning model capable of understanding sentences alongside the context in which they are be- ing 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. For this project, I do not plan to pre-train the model as my text data is rather limited and hence, will use FinBERT instead. FinBERT 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 advantage of FinBERT for this project is that it was pre-trained using longer texts, so it splits sentences individually and then calculates sentiment on each one of them. This will allow me to average sentiment measures for each sentence in each policy statement and build an index which I will call the FOMC sentiment index. This index can be used for a number of purposes including investigating the link between the FOMC sentiment and the financial market conditions via regressions among other uses.

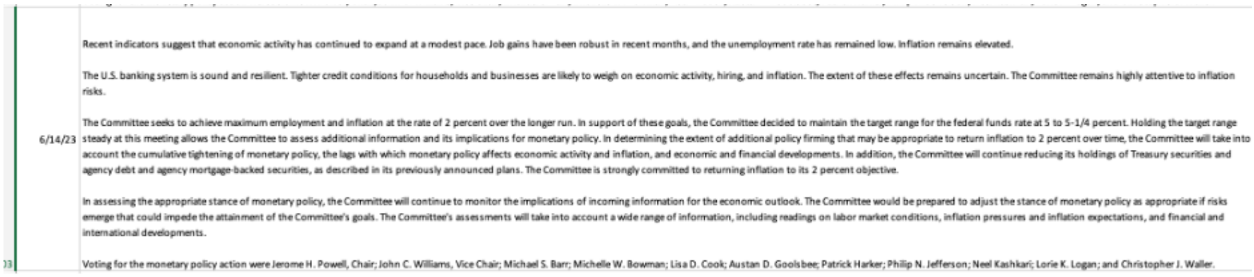
FinBERT produces five sentiment values. Three values represent the probabilities that a sentence is either positive, negative, or neutral. FinBERT also calculates a compound score as the positive probability minus the negative probability. Lastly, FinBERT provides trinary sentiment prediction which is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

Below are the steps in my flow diagram:

1) Data Collection: For the first step in my research project, I did data collection. I gathered all FOMC policy statements from January of 2000 to the latest statement (6/14/2023) from the Federal Reserve website. Please see an example screenshot below.



Date	Statement
2/2/00	The Federal Open Market Committee voted today to raise its target for the federal funds rate by 25 basis points to 5-3/4 percent. In a related action, the Board of Governors approved a 25 basis point increase in the discount rate to 5-1/4 percent.The Committee remains concerned that over time increases in demand will continue to exceed the growth in potential supply, even after taking account of the pronounced rise in productivity growth. Such trends could foster inflationary imbalances that would undermine the economy's record economic expansion.Against the background of its long-run goals of price stability and sustainable economic growth and of the information currently available, the Committee believes the risks are weighted mainly toward conditions that may generate heightened inflation pressures in the foreseeable future.In taking the discount rate action, the Federal Reserve Board approved requests submitted by the Boards of Directors of the Federal Reserve Banks of Boston, New York, Philadelphia, Cleveland, Richmond, Atlanta, Chicago, St. Louis, Kansas City and San Francisco. The discount rate is the rate charged depository institutions when they borrow short-term adjustment credit from their district Federal Reserve Banks.

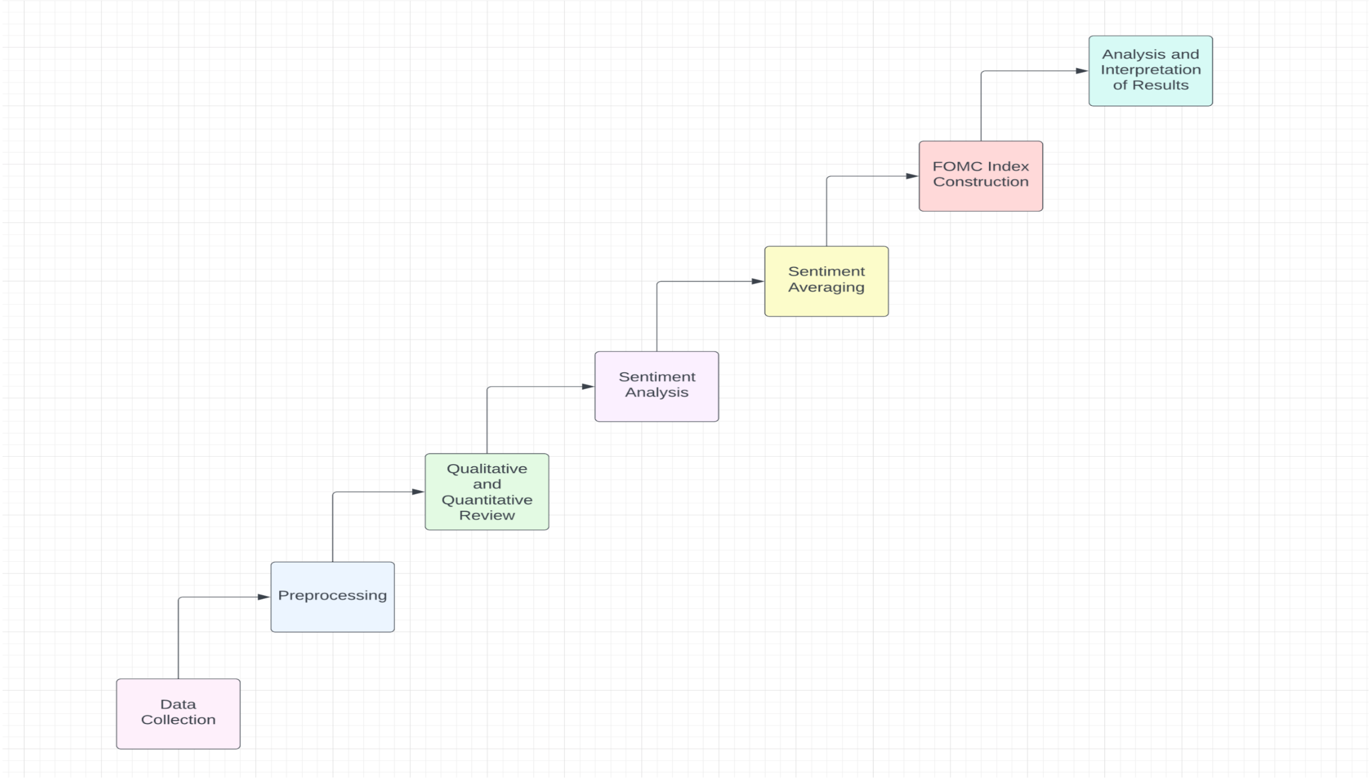


6/14/23	Recent indicators suggest that economic activity has continued to expand at a modest pace. Job gains have been robust in recent months, and the unemployment rate has remained low. Inflation remains elevated. The U.S. banking system is sound and resilient. Tighter credit conditions for households and businesses are likely to weigh on economic activity, hiring, and inflation. The extent of these effects remains uncertain. The Committee remains highly attentive to inflation risks. The Committee seeks to achieve maximum employment and inflation at the rate of 2 percent over the longer run. In support of these goals, the Committee decided to maintain the target range for the federal funds rate at 5 to 5-1/4 percent. Holding the target range steady at this meeting allows the Committee to assess additional information and its implications for monetary policy. In determining the extent of additional policy firming that may be appropriate to return inflation to 2 percent over time, the Committee will take into account the cumulative tightening of monetary policy, the lags with which monetary policy affects economic activity and inflation, and economic and financial developments. In addition, the Committee will continue reducing its holdings of Treasury securities and agency debt and agency mortgage-backed securities, as described in its previously announced plans. The Committee is strongly committed to returning inflation to its 2 percent objective. In assessing the appropriate stance of monetary policy, the Committee will continue to monitor the implications of incoming information for the economic outlook. The Committee would be prepared to adjust the stance of monetary policy as appropriate if risks emerge that could impede the attainment of the Committee's goals. The Committee's assessments will take into account a wide range of information, including readings on labor market conditions, inflation pressures and inflation expectations, and financial and international developments. Voting for the monetary policy action were Jerome H. Powell, Chair; John C. Williams, Vice Chair; Michael S. Barr; Michelle W. Bowman; Lisa D. Cook; Austan D. Goolsbee; Patrick Harker; Philip N. Jefferson; Neel Kashkari; Lorie K. Logan; and Christopher J. Waller.

2) Preprocessing: I needed to clean the data to ensure it is in a format that FinBERT can process. This includes removing unnecessary punctuation, tokenizing the text, lemmatization etc.

3) Qualitative and Quantitative Review: As part of the qualitative review, I reviewed a subset of the statements for understanding the context. This allows me to have the necessary background to interpret the results from the sentiment analysis. Additionally, the quantitative review involves using FinBERT for sentiment analysis of the text data. It's a quantitative technique because it results in a numerical score representing the sentiment conveyed in the text.
4) Sentiment Analysis: For this part, I will apply FinBERT to each sentence in each statement. Each sentence will receive a sentiment score.
5) Sentiment Averaging: Compute the average sentiment for each statement to create a single sentiment score for that statement.
6) Index Construction: For This I will use the average sentiment scores to construct the FOMC sentiment index.
7) Analysis and Interpretation: Finally, I will analyze the resulting sentiment index in the context of other economic and political factors. This might involve correlating the sentiment index with other economic indicators, or investigating how the sentiment index changes in response to specific events.

Please see the flow diagram below.



Also I plan to validate my research findings by using my qualitative review. This can guide me to further ensure the sentiment analysis and the patterns observed align reasonably with the qualitative understanding of the policy statements. The quantitative reviews will give an accurate understanding of the FOMC policy statements.

Data Collection Strategy:

The sample of statements for this project is planned to cover all the available statements made by the FRB Chairs since 2000. Prior to 2000, Chairs did not issue a press statement. Since there are 8 FOMC meetings in a year, this should give us roughly 180 texts of statements. These statements are publicly available and I download them from: https://www.federalreserve.gov/monetarypolicy/fomc_historical_year.htm.

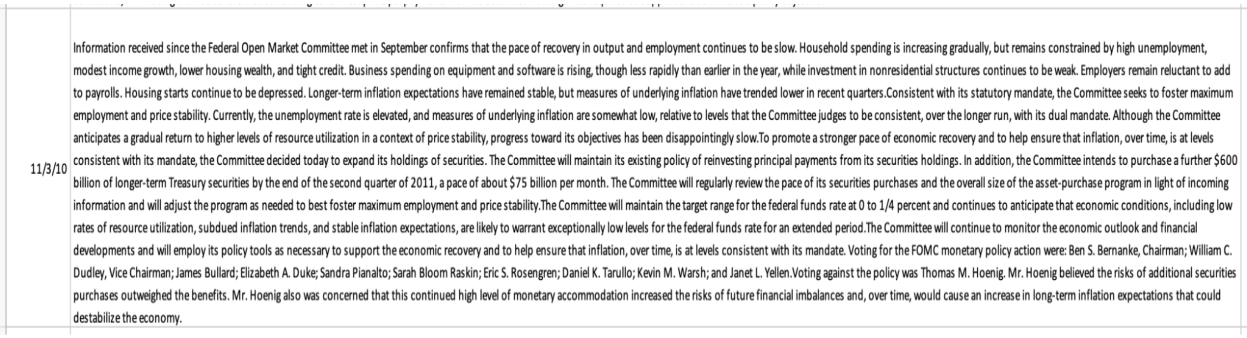
Link to my Dataset:
https://github.com/ADA-GWU/guidedresearchproject-ayda000/blob/main/fomc_statements.xlsx

Also below are more details on my data collection strategy:

1) Identify Source: The source for your data will be the Federal Reserve website, specifically the page that archives all FOMC policy statements.
2) Determined Range: I clearly defined the range of my data collection by collecting policy statements from 2000 onwards. To start, confirm the exact start date (2/2/2000- meeting of the year 2000) and end date (6/14/2023- the most recent FOMC meeting).
3) Download Statements: I accessed each FOMC policy statement individually. The Federal Reserve website typically provides these statements in a webpage (html) format. I downloaded the content of these statements manually.

4) Organize and Store Data: I stored the statements as an excel spreadsheet where each row corresponds to a different FOMC meeting and the columns include the date of the meeting and the full text of the policy statement.
5) Backup Data. I made sure to back up my data to prevent any loss.
6) Completeness and Quality Check: Once I collected all the statements, I verified the completeness of my data. I made sure all FOMC meetings within the date range are accounted for and that the text of each policy statement has been accurately and fully captured.

Also all the statements are organized in a similar format like below. They start and end in a similar manner even though the content differs.



11/3/10	Information received since the Federal Open Market Committee met in September confirms that the pace of recovery in output and employment continues to be slow. Household spending is increasing gradually, but remains constrained by high unemployment, modest income growth, lower housing wealth, and tight credit. Business spending on equipment and software is rising, though less rapidly than earlier in the year, while investment in nonresidential structures continues to be weak. Employers remain reluctant to add to payrolls. Housing starts continue to be depressed. Longer-term inflation expectations have remained stable, but measures of underlying inflation have trended lower in recent quarters.Consistent with its statutory mandate, the Committee seeks to foster maximum employment and price stability. Currently, the unemployment rate is elevated, and measures of underlying inflation are somewhat low, relative to levels that the Committee judges to be consistent, over the longer run, with its dual mandate. Although the Committee anticipates a gradual return to higher levels of resource utilization in a context of price stability, progress toward its objectives has been disappointingly slow.To promote a stronger pace of economic recovery and to help ensure that inflation, over time, is at levels consistent with its mandate, the Committee decided today to expand its holdings of securities. The Committee will maintain its existing policy of reinvesting principal payments from its securities holdings. In addition, the Committee intends to purchase a further $600 billion of longer-term Treasury securities by the end of the second quarter of 2011, a pace of about $75 billion per month. The Committee will regularly review the pace of its securities purchases and the overall size of the asset-purchase program in light of incoming information and will adjust the program as needed to best foster maximum employment and price stability.The Committee will maintain the target range for the federal funds rate at 0 to 1/4 percent and continues to anticipate that economic conditions, including low rates of resource utilization, subdued inflation trends, and stable inflation expectations, are likely to warrant exceptionally low levels for the federal funds rate for an extended period.The Committee will continue to monitor the economic outlook and financial developments and will employ its policy tools as necessary to support the economic recovery and to help ensure that inflation, over time, is at levels consistent with its mandate. Voting for the FOMC monetary policy action were: Ben S. Bernanke, Chairman; William C. Dudley, Vice Chairman; James Bullard; Elizabeth A. Duke; Sandra Pianalto; Sarah Bloom Raskin; Eric S. Rosengren; Daniel K. Tarullo; Kevin M. Warsh; and Janet L. Yellen.Voting against the policy was Thomas M. Hoenig. Mr. Hoenig believed the risks of additional securities purchases outweighed the benefits. Mr. Hoenig also was concerned that this continued high level of monetary accommodation increased the risks of future financial imbalances and, over time, would cause an increase in long-term inflation expectations that could destabilize the economy.

Data Cleansing Approaches:

The data preprocessing and cleansing is a crucial step in any natural language processing task especially sentiment analysis. Before proceeding with the sentiment analysis using FinBERT, I ensured my data is cleaned and properly structured.

Noise Removal: Remove unnecessary and irrelevant items (noise) from the text. I removed any sort of noise in the text using Python's regex (re) library to clean the text.

Tokenization: Tokenization is the process of splitting text into individual words or tokens. Breaking the text into individual sentences. Each sentence will be a separate document to be fed into FinBERT. Python's NLTK library has a utility for sentence tokenization. This is necessary

because NLP models don't process entire sentences or documents all at one time. BERT models don't take raw text as input and they require tokenizing input text into tokens. The specific tokenization method depends on the model. BERT does use WordPiece tokenization.
Also BERT expects special tokens that show the beginning and separation/end of sentences.

Stop Word Removal: Remove stop words (eg.a and) to focus on the more meaningful words. For my case, FinBERT is already trained to handle stop words in financial text.

Stemming and Lemmatization: This step involves reducing words to their root form. Stemming is a rudimentary rule-based process of stripping the suffixes.
Lemmatization is an organized & step by step procedure of obtaining the root form of the word.

Lowercasing is when converting all text to lowercase can help ensure that the model doesn't treat the same word with different cases as different words for instance "Federal" vs "federal"

Removing Punctuation can be important as it can confuse NLP models. The punctuation is not specifically relevant to my analysis so I can remove it.

Data Integration: Also I am keeping data in common format.

Case Normalization -> Convert all text to lower case. This is done to prevent duplication of words due to case difference.

I need to check for special characters. I need to decide how to handle special characters like dollar signs, percentage signs, ampersands, etc.

I will have a tabular data structure where each row represents a unique statement from the FRB Chair, with associated metadata like date, chair's name, and the cleaned text.

Since new statements are released every year the data cleaning process will be done in a reproducible way. I will make sure to capture in detail the data cleaning process into a function or script to easily re-run it in the future as needed. Also I will keep a raw copy of the original data before starting the cleaning process.

These all are executed prior to feeding data into the FinBERT model.

BERT-based models, including FinBERT require the cleaning steps like the ones explained above.

Also other steps to consider may include using an attention mask for a sequence of 1s and 0s indicating which tokens are padding and not. Also all sentences fed into BERT must be of the same length. Padding is done to match the length of the sentences in a batch.

I will have to depend on an NLP library to do these preprocessing steps before using FinBERT. I found that for Python, a common choice is the Transformers library which includes built-in tokenizers for BERT and many other models.

Also next are some Data visualizations I am considering includes:

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.
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.

The data cleaning for this dataset was quite organized. Also the primary challenge is only the large volume of data. Since I collected data from about 180 policy statements, the data cleaning process is time-consuming.