Predicting the Federal Funds Target Rate Using FOMC Minutes: A Comparison Between Bag of Words and Sequential Models (BERT)

Vinicius Ono Sant'Anna, Trung Le
Department of Mathematics and Statistics
Grinnell College, Grinnell, IA 50112, USA

Grinnell STA-395 Final Project
May 16, 2023

{onosanta,letrung}@grinnell.edu
Project Repository

Abstract

This paper explores the application of two different text mining paradigms - Bag of Words (BoW) and Sequential Models (BERT) - for predicting changes in the Federal Funds Target Rate using the Federal Open Market Committee (FOMC) minutes. The aim is to determine which paradigm offers better predictive performance in this context. Our methodology involves tokenizing and vectorizing the minutes, training separate models based on BoW and BERT, and then comparing their performance. Our results reveal a surprising outcome: Sequential models underperformed in comparison to our best Bag-of-Words model.

Keywords: NLP; FOMC; Federal Funds Rate; BERT; BoW

1 Introduction

The Federal Funds Rate, a crucial indicator of the U.S. economy, plays a significant role in decision-making processes for investors, policymakers, and businesses (Bernanke & Kuttner, 2005). Predicting changes in this rate is valuable for financial planning, investment strategies, and policy formulation.

The minutes of the Federal Open Market Committee (FOMC) meetings offer rich information for predicting the Federal Funds Rate. These minutes provide detailed accounts of policy discussions, reflecting the Fed's assessment of economic conditions and monetary policy stance (Rosa, 2013). However, leveraging this text data for predictive analysis poses challenges due to the complexity of natural language.

This study aims to explore advanced natural language processing techniques for predicting changes in the Federal Funds Rate using the FOMC minutes. Specifically, we compare the efficacy of Bag of Words (BoW) and Bidirectional Encoder Representations from Transformers (BERT) approaches. We hypothesize that the FOMC minutes not only provide insights into impending rate changes but also that the contextual elements significantly impact their predictive potential. Through quantitative and qualitative analyses, this study aims to uncover the relationship between FOMC communications and policy actions, contributing to a nuanced understanding of monetary policy decisions.

A major challenge of this project is the limited number of time-series observations, with only 240 documents released from March 1993 to May 2013. While models with fewer parameters are less prone to overfitting this small dataset, they might not capture the complex dynamics of interest rate movements. Our methodology involves encoding words using pre-trained BERT models (DistilBERT and FinBERT) and then constructing Bag of Words models (BoW) through Grid search cross-validation. We optimize the hyperparameters of each of the following BoW models: Logistic Regression, Random Forest, XGBoost, and Multinomial Naive Bayes. Surprisingly, contrary to our initial hypothesis, the sequential models underperformed compared to our best Bag-of-Words model.

2 Data

2.1 Data Sources

Our two data-sets were the FOMC minutes and the Federal fund's Target rate. The FOMC minutes are publicly available online that can be accessed via The Federal Reserve. Through web scraping (using the Beautiful Soup library), we have compiled the FOMC minutes from March 1993 - May 2023. The historical Federal funds rate is then collected using the Federal Reserve Bank of St. Louis (FRED) API. After successfully collecting both datasets, we then merged by date, ensuring a one-to-one match between the FOMC minutes and the corresponding Federal Funds Rate. Hence, our raw dataset was a 3-column table with dates as the index, FOMC text for the given meeting, and the federal funds rate decision for the given FOMC. We then created a variable indicating the next target federal funds rate, which is the lead variable from the current target rate. After completing our data wrangling process, we categorized our FOMC next target rate variable into three groups: "Unchanged", "Increase", and "Decrease". This involves creating a variable delta $Rate_{\Delta}$, which refers to the change in the rate at time $t+1$ and the rate at time t, expressed as $(Rate_{\Delta} = Rate_{t+1} - Rate_t)$. Based on the value of $Rate_{\Delta}$, we assign labels to indicate whether the current rate at time t changed positively $(Rate_{\Delta} > 0)$, negatively $(Rate_{\Delta} < 0)$, or remained the same $(Rate_{\Delta} = 0)$. This labeling process is part of our methodology and helps us categorize the rate changes in our analysis.

2.2 Data Pipeline/ Cleaning

Upon structuring our dataset, the next vital step was cleaning. Given that our dataset was entirely composed of text, this was a crucial phase in our pipeline. Effective cleaning is pivotal to ensure that our models do not learn from potential data noise, which could skew the results. Therefore, we meticulously cleaned our dataset, striving for the highest accuracy possible in our subsequent analyses. Text cleaning involved lowercasing, punctuation removal, tokenization, and lemmatization using the Natural Language Toolkit (NLTK) (Bird, 2006). Tokenization and lemmatization are particularly important. Tokenization dissects text into 'tokens' or individual words, converting unstructured text into a format that models can analyze. Each token essentially becomes a model input. Conversely, lemmatization condenses words to their root form, enabling models to identify different forms of the same word as a single concept. Additionally, we removed any stop words, which are commonly used words such as "and", "the", "is", etc. This step is important as these words often do not carry significant meaning and tend to clutter the text data. Finally, following the approach of (Ye, 2020), we removed the standard introduction from the FOMC minutes by removing any text preceding the first match of the word "unanimous" or "manager". Here are some summary statistics of our pre-processed data:

- **Number of FOMC documents**: 240
- **Number of words**: $4,622,140$
- **Number of unique words**: $7,935$

3 Methodology

3.1 Feature engineering

Prior to initiating our model building, we devised sentiment scores for each set of FOMC minutes based on five key categories: positive, negative, uncertainty, litigious, and constraining. These categories were chosen as they effectively capture the various tones and sentiments present in financial discourse, which can significantly impact monetary policy decisions. The sentiment scores were computed using Loughran and McDonald's financial dictionary, a highly regarded financial text analysis resource (2011). This dictionary is specifically tailored for the financial domain, differentiating it from generic sentiment dictionaries. It contains a comprehensive list of words categorized according to their sentiment, specifically calibrated to capture the nuances of financial language. For instance, a word considered 'negative' in a financial context might not necessarily be negative in a general context. By applying this dictionary to the FOMC minutes, we translated the qualitative information in the text into quantifiable data, thus enabling us to analyze the sentiment of each document. The sentiment scores provide us with insightful data points, revealing the dominant sentiment conveyed in each FOMC minutes and helping us predict changes in federal fund rates more accurately. To summarize our approach, we developed functions that counted the occurrences of specific words within each category and converted these counts into proportions relative to the total number of words in each FOMC text. It's important to note that the scores we obtained range from 0 to 1. As a result of this process, we added several new variables to our dataset. These variables include the following: positive_score, negative_score, uncertainty_score, litigious_score, and constraining_score. Each of these scores represents the proportion of words in the corresponding category relative to the total number of words in the FOMC text. By incorporating these scores into our dataset, we can gain insights into the prevalence and relative strength of positive, negative, uncertain, litigious, and constraining language used within the FOMC texts. This information can be valuable for understanding the sentiment, risk, and potential regulatory implications conveyed in these documents.

It is important to note that this feature engineering process was specifically applied to our Bag of Words (BoW) models. This is because the pre-trained BERT models we used can only take text as input and do not require explicit feature engineering. Therefore, the steps we described for counting words and generating proportions were not applicable to the pre-trained BERT models, as they handle the text input differently.

3.2 Model Building

We leverage two most popular paradigms in natural language processing to process our textual FOMC meetings' minutes: the Bag of Words (BoW) model and the sequential model represented by BERT (Bidirectional Encoder Representations from Transformers).

3.2.1 Bag of Words Models (BoW)

Bag of Words (BoW) is a widely used NLP technique that simplifies text data by converting words into numeric vectors based on their frequency. It disregards word order and focuses on capturing word occurrence in each document. We implement BoW using tf-idf vectorization (Ramos, 2003), which considers word frequency within a document and its significance across the corpus. Words that are common in a document but rare overall receive higher weights. This process involves calculating term frequency (TF), normalizing it, and calculating inverse document frequency (IDF) to measure word rarity. Multiplying TF and IDF gives the tf-idf representation, a numerical vector suitable for machine learning.

After obtaining the tf-idf matrix, we scale the engineered features using Min-Max Scaler, especially we scale our scores generated through Loughran and McDonald's financial dictionary. Then, we train several machine learning models on the combined dataset, including Logistic Regression, Random Forest, XGBoost, and Multinomial Naïve Bayes. These models are selected for their effectiveness in classification tasks and suitability for our dataset.

To find the best hyperparameters for each model, we employ grid search, which systematically evaluates different parameter combinations for cross-validation of 10 folds, since our dataset was small. The hyperparameters are specified in **Table 1**. We use the F1 score, a balanced measure of precision and recall, as the evaluation metric for a grid search. This is particularly useful for imbalanced datasets, providing an accurate assessment of model performance.

3.2.2 Sequential Models

Sequential models are an important class of algorithms that process words in the order in which the words appear. This allows them to capture important syntactic structures and semantic nuances in English sentences. These capabilities may allow these models to capture context of the words and sentences in the FOMC minutes, extracting predictive information for the next target federal funds rate. to limitations of simple sequential models like Recurrent Neural Networks (RNNs), which can only use the current state and the current word to predict the next state cannot capture long-term context dependency, we employ Transformer-based models, like BERT, to combat that shortcoming.

Instead of processing the sentence word by word, BERT (Devlin, Chang, Lee, & Toutanova, 2019) uses a mechanism called self-attention to consider all words in the sentence simultaneously and weighs their influence on each other to generate a contextualized representation for each word. This allows BERT to capture both short-term and long-term dependencies between words, and to consider the full context of the sentence in both directions (left to right and right to left).

In our study, we leverage two specific variants of BERT: DistilBERT and FinBERT:

- **DistilBERT (Sanh, Debut, Chaumond, & Wolf, 2020):** This is a distilled version of BERT. It retains 95% of original BERT performance while being smaller and faster using a process called knowledge distillation, in which the distilled model predicts the output of the original model. The lightweight architecture of DistilBERT allows faster training time for our application.
- **FinBERT (Araci, 2019):** This is a domain-specific version of BERT that was pre-trained on large-scale financial text data. FinBERT is expected to perform better than general BERT models on tasks involved with financial text data and our particular task, given that these minutes contain a significant amount of financial jargon and complex economic concepts.

For each variant, we first tokenize and encode the FOMC minutes into an input format suitable for BERT models. These encoded inputs are then fed into the BERT model, which outputs contextualized representations of each word. We leverage the representation of the special [CLS] token, which aggregates sentence-level information with contextual awareness, and pass it through a fully connected layer to predict the change in the Federal Funds Rate. We then use Stochastic Gradient Descent to update the models' parameters and fine-tune the models. We report the same metrics used in BoW for our BERT models on training and test data.

4 Results

Tables 2 and **3** present the prediction accuracy, precision, recall, and F1-score of the optimal models from each classifier on the training and testing datasets respectively. Notably, in **Table 2**, among all Bag-of-Words (BoW) models, the Naïve Bayes model demonstrated the best performance. However, we observed potential overfitting in both the Random Forest and XGBoost models. Interestingly, both DistilBERT and FinBERT achieved similar scores, indicating comparable performance despite differences in their pre-trained data. In Summary, the Multinomial Naïve Bayes model outperformed all other models with an impressive F1-score of 0.9575. Surprisingly, even compared to all BoW models, the two BERT variants exhibited inferior performance. On the testing dataset (**Table 3**), consistent with the training results, the sequential models performed worse compared to other models. Both (BERT, FinBERT) with an F-1 score of 0.7396 in comparison to 0.739696 with the best-performing model from the BoW batch.

5 Discussion

The poor performance exhibited by the BERT and FinBERT models can be attributed to two primary factors. Firstly, it is important to acknowledge the presence of a class imbalance within our dataset. **Graph 1** demonstrates that out of our 240 observations, 145 are labeled as "no change," 45 as "increase," and 26 as "decrease." Research by Sun, Huang, and Qiu (2019) highlights that an imbalanced dataset tends to result in suboptimal performance for BERT models. Therefore, it is reasonable to conclude that the poor performance can be attributed to the high class imbalance within our data.

Secondly, our decision to collect a sample of the FOMC minutes spanning from 1993 to 2023 has led to a dataset consisting of 240 observations. By dividing this dataset into a 80-20 train-test split, we are left with a relatively small training dataset of only 192 observations. Liu et al. (2019) discuss the limitations of BERT when applied to small datasets and emphasize that BERT cannot be effectively explored or fine-tuned with such limited sample sizes.

In conclusion, the underperformance of the BERT and FinBERT models can primarily be attributed to the class imbalance within the dataset and the small training dataset size. However, it may also be that BoW models are just better in the context of the FOMC, which implies that getting the context or semantic meaning of sentences for FOMC minutes might not be as relevant. In addition, BoW can handle out-of-vocabulary words, words that were not seen during training of BERT models, making them more resilient at financial jargon and economic textual data.

6 Conclusion

Contrary to our initial hypothesis, our results reveal a surprising outcome: Sequential models underperformed in comparison to our best Bag-of-Words model. This counter-intuitive finding challenges our initial assumption about the superior predictive power of Sequential models. It suggests that the simpler, less context-dependent Bag-of-Words approach can, in fact, more accurately forecast changes in federal fund rates based on the information extracted from FOMC minutes. This unexpected outcome urges a reconsideration of how we perceive the relationship between the context of FOMC communications and their predictive potential for federal fund rate changes. Indeed, the Federal Reserve designs the FOMC minutes to maintain a certain degree of generality, aiming to prevent markets from prematurely predicting their future actions. Consequently, the context of FOMC minutes may not yield substantial predictive information. It might, therefore, be more advantageous to employ a Bag-of-Words model that focuses on individual word patterns, as this approach could more effectively capture the nuances of these critical financial communications. However, as we only used BERT variants for this task, we were limited to the BERT architecture. To fortify this conclusion, we would want to employ newer models that employ more recent word embeddings methods and transformer layers.

Future research could explore various model architectures, such as LSTM and GRU, and employ techniques like gradient clipping to potentially enhance stability and performance (Pascanu, Mikolov, & Bengio, 2013). Additionally, integrating key economic indicators, such as CPI, Unemployment Rate, and economic growth variables, alongside text data could provide a more comprehensive understanding of the factors influencing federal fund rate changes. By conducting rigorous hyperparameter tuning, including batch size, learning rate, and training epochs, it could be possible to optimize model performance and potentially improve the accuracy of predictions based on FOMC communications.

References

Araci, D. (2019). *Finbert: Financial sentiment analysis with pre-trained language models.*

Bernanke, B. S., & Kuttner, K. N. (2005). What explains the stock market's reaction to federal reserve policy? *The Journal of Finance*, *60*(3), 1221–1257.

Bird, S. (2006, July). NLTK: The Natural Language Toolkit. In *Proceedings of the COLING/ACL 2006 interactive presentation sessions* (pp. 69–72). Sydney, Australia: Association for Computational Linguistics. Retrieved from https://aclanthology.org/P06-4018 DOI: 10.3115/1225403.1225421

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *Bert: Pre-training of deep bidirectional transformers for language understanding.*

Federal Reserve Bank of St. Louis. (2023). *FRED API.* https://fred.stlouisfed.org/docs/api/fred/.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... Stoyanov, V. (2019). *Roberta: A robustly optimized bert pretraining approach.*

Loughran, T., & McDonald, B. (2011). When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance*, *66*(1), 35-65. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2010.01625.x DOI: https://doi.org/10.1111/j.1540-6261.2010.01625.x

Pascanu, R., Mikolov, T., & Bengio, Y. (2013, 17–19 Jun). On the difficulty of training recurrent neural networks. In S. Dasgupta & D. McAllester (Eds.), *Proceedings of the 30th international conference on machine learning* (Vol. 28, pp. 1310–1318). Atlanta, Georgia, USA: PMLR. Retrieved from https://proceedings.mlr.press/v28/pascanu13.html

Ramos, J. E. (2003). Using tf-idf to determine word relevance in document queries..

Rosa, C. (2013). The financial market effect of fomc minutes. *Economic Policy Review*, *19*(2). Retrieved from https://ssrn.com/abstract=2378398

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.*

Sun, C., Huang, L., & Qiu, X. (2019, June). Utilizing BERT for aspect-based sentiment analysis via constructing auxiliary sentence. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 380–385). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1035 DOI: 10.18653/v1/N19-1035

The Federal Reserve. (2023). *FOMC Minutes.* https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm.

Ye, Y. (2020, March). *Can fomc minutes predict the federal funds rate?* Stanford CS224N Custom Project.

Appendix

Tables

Table 1: Hyperparameters considered for the Grid Search for Each BoW model

Model	**Hyperparameters and Values**
Logistic Regression	Penalty: l1, l2 C: 1.0, 0.5, 0.1
Random Forest	n_estimators: 100, 200 max_depth: None, 10, 20 min_samples_split: 2, 5, 10
XGBoost	n_estimators: 100, 200 learning_rate: 0.1, 0.01 max_depth: 3, 5, 7
Multinomial Naive Bayes	alpha: 1.0, 0.1, 0.01 fit_prior: True, False

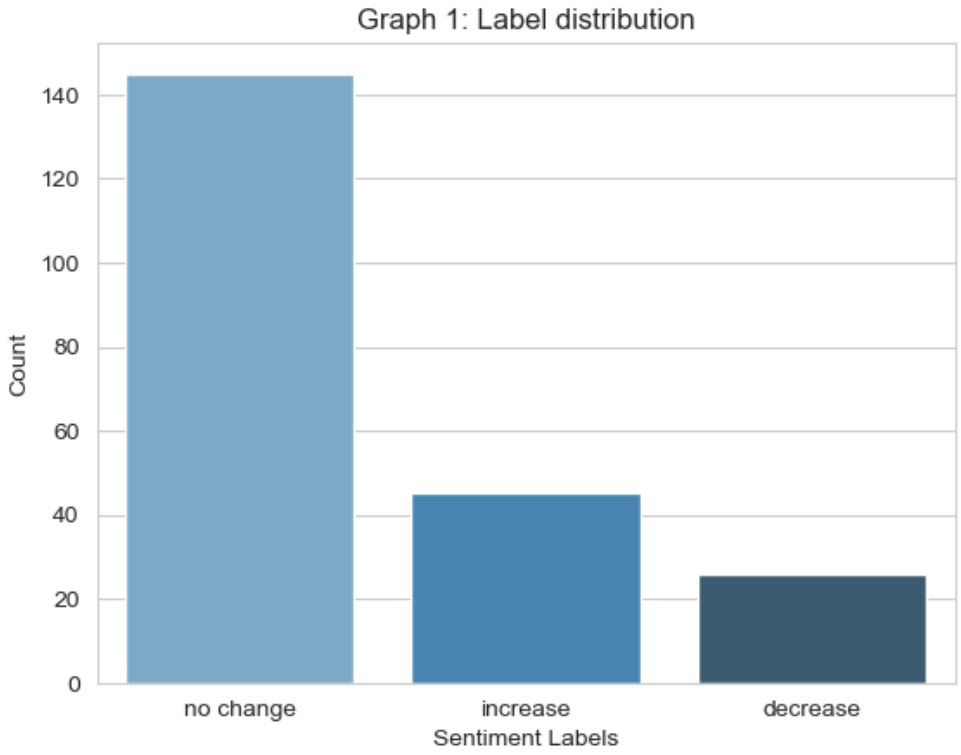
Table 2: Performance of Non-sequential vs Sequential Models on Training data

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.8333	0.8665	0.8333	0.8142
Random Forest	1.000	1.000	1.000	1.000
XGBClassifier	1.000	1.000	1.000	1.000
Multinomial Naive Bayes	**0.9583**	**0.9608**	**0.9583**	**0.9575**
DistilBERT FT [Sanh *et al.*,2019]	0.6667	0.4444	0.6667	0.5333
FinBERT FT [Araci, 2019]	0.6667	0.4444	0.6667	0.5333

Table 3: Performance of Non-sequential vs Sequential Models on Test data

Model	**Accuracy**	**Precision**	**Recall**	**F1-score**
Logistic Regression	0.7917	0.6267	0.7917	0.6996
Random Forest	0.7917	0.6267	0.7917	0.6996
XGBClassifier	0.7500	0.6196	0.7500	0.6786
Multinomial Naive Bayes	**0.7500**	**0.8075**	**0.7500**	**0.7396**
DistilBERT FT [Sanh *et al.*,2019]	0.7500	0.5625	0.7500	0.6429
FinBERT FT [Araci, 2019]	0.7500	0.5625	0.7500	0.6429

Graphs


Graph 1: Label distribution
Count
140
120
100
80
60
40
20
0
no change
increase
decrease
Sentiment Labels