NLP of Fed. Reserve Speeches To Predict Stock Market Trend

JP Baselj

1 Introduction

The global economy continues to grapple with the impacts of the Covid-19 pandemic, with the US Federal Reserve largely responsible for shaping economic stability. Federal Reserve chairman Jerome Powell periodically makes public statements on monetary policy, including federal interest rates, which “The Fed” ultimately controls. Since the inception of Covid-19 in the US, these speeches of federal monetary policy have incited pronounced volatility and trading activity in the stock market.

This project harnesses Natural Language Processing (NLP) to analyze Powell’s speeches and predict the S&P500’s trend over the following week. Our objective is to develop a model that, on average, would yield profitable results when investing in S&P500 index or reverse index funds based on predicted market direction. This report outlines the project’s methodology and findings covering data wrangling, Exploratory Data Analysis (EDA), feature engineering, and machine learning modeling.

2 Dataset

This analysis is based on two key datasets: speech transcripts from Chair Jerome Powell and daily S&P500 stock market data. The speech transcripts, sourced from the Federal Reserve’s website, were initially in PDF format and had to be converted to plain text files. This text was then cleaned by preprocessing methods including converting all letters to lowercase, removing punctuation and common english stop words, and tokenizing the text.

The S&P500 historical records data were sourced from investing.com and required light preprocessing including the removal of empty fields and conversion of numeric values to suitable data types. Fields representing the Maximum and Minimum price for each trading day were used to create a new field, daily span, which serves as a proxy for daily market volatility. The macroscopic trend of the Price and Daily Span of the market over the target time frame is plotted in **Figure 1**, along with vertical markers denoting the date of each speech in consideration.

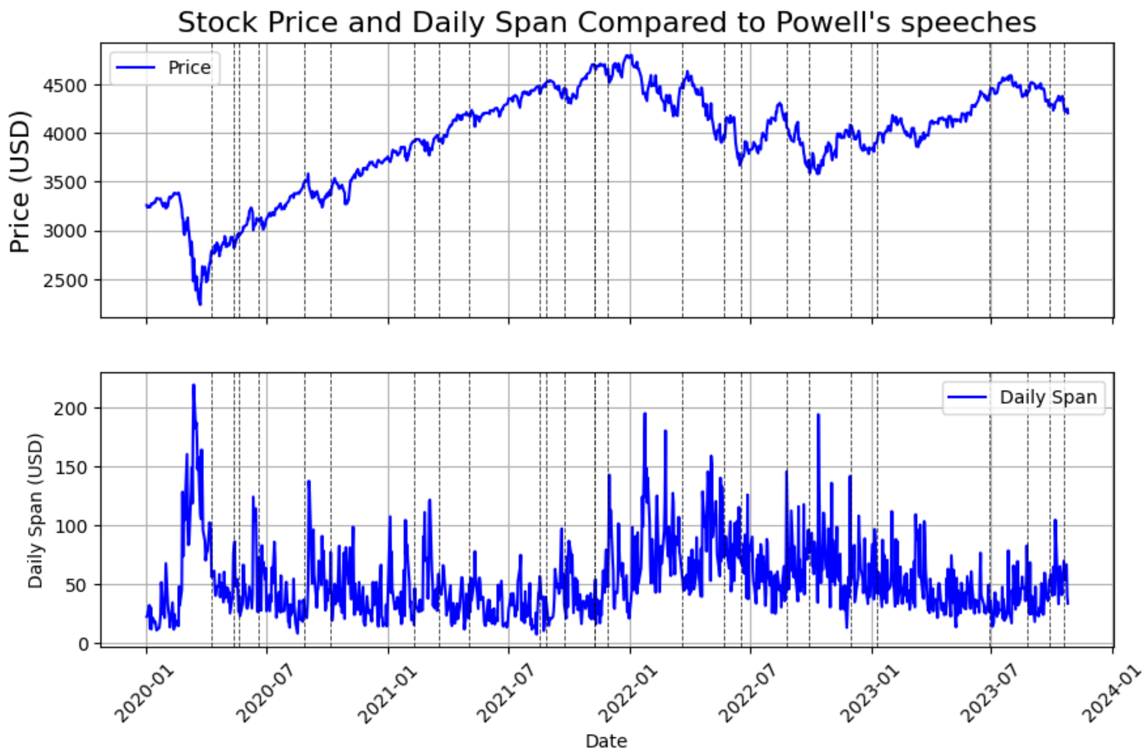

Figure 1

Additionally, the stock market data was aggregated on a weekly basis to create features that represent the market value, volatility, and returns for any 7 day span. The S&P data was then merged with the data from Chair Powell's speeches so the market features became linked to the weeks preceding and following each speech.

3 EDA and Feature Engineering

The analysis performed aimed to foster of how the features of the speech text interact with the features of the stock market data. First, sentiment analysis was performed on Chair Jerome Powell's speeches using the VADER module from Python's Natural Language Toolkit (NLTK) suite, made for processing textual data. Each text was evaluated to produce four different sentiment score features, which represent the sentiment for the beginning, middle, and end one-thirds of the speech, as well as the full speech overall.

The rolling stock market features surrounding each speech were visualized to identify what trends, if any, existed in the speeches' influence of the market over time. **Figure 2** illustrates the change in average market value from the week before to the week after each speech, with vertical dashes lines indicating the dates of speeches.

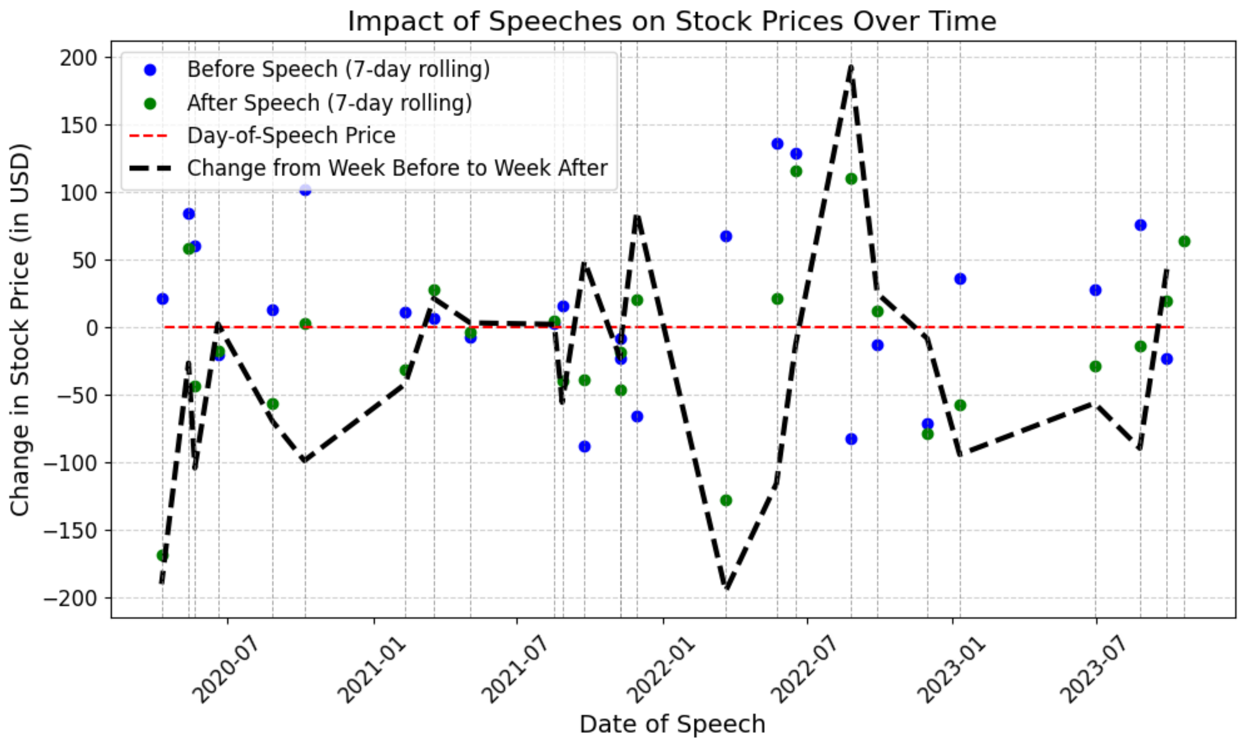

Figure 2

Features relevant for modeling were identified to form a dataset, considering their significance in capturing the interaction between speeches and market dynamics. Notably, stock market data representing the past for each speech entry was retained, while data from the "future" was removed, acknowledging that future data will not be available for real-world model application. In addition, a target variable for modeling was created: a binary variable representing whether the S&P500 index increased in value from the week before to the week after a given speech.

The speech-text data was vectorized into a format compatible with the standard Machine Learning models available in the Scikit-Learn Library. Two distinct vectorization techniques, TF-IDF and Bag of Words, were each used to vectorize speeches, to later be compared for their performance in capturing relevant features for predicting market trends. The two resulting datasets were oversampled to create balanced-class training sets, while also preserving the natural/imbalanced class sets to enable a comparative assessment of their efficacy in predicting market outcomes. Each of the 4 data preparation combinations were randomly split into training and testing sets for modeling.

4 Modeling

To establish a baseline for model performance, Logistic Regression (LogReg), Random Forest (RF) Classifier, and Support Vector Machine (SVM) models were trained and evaluated using cross validation for each of the 4 data input combinations. Assessment of each of the 12 combinations of model type and data input type yielded that the best data for predicting stock market performance was using a Bag of Words text vectorizer, and training the model using balanced classes. The RF and LogReg models performed better than the SVM model, so each was further hyperparameter tuned and evaluated, resulting in the tuned RF model being selected as the final model. Model metrics including feature importances can be found in the "Model Metrics" document in the same filepath as this report.

In the final modeling phase, the Random Forest Classifier model was evaluated using all available data, including the holdout validation dataset. Because of the relatively small amount of total available data limited by number of input speeches, cross validation techniques were used to calculate a model accuracy of about 72%.

A simulation was performed to model how applying a predictor with 72% accuracy may behave for our use case of predicting market moves following a speech, and the results are visualized in **Figure 3.** The first plot in figure 3 represents the pseudo-random walk of the cash value of an investment had a model with 72% accuracy been applied to predict the stock trend over Powell's previous speeches. The second plot represents the distribution of the results of running the same simulation 1000 times.

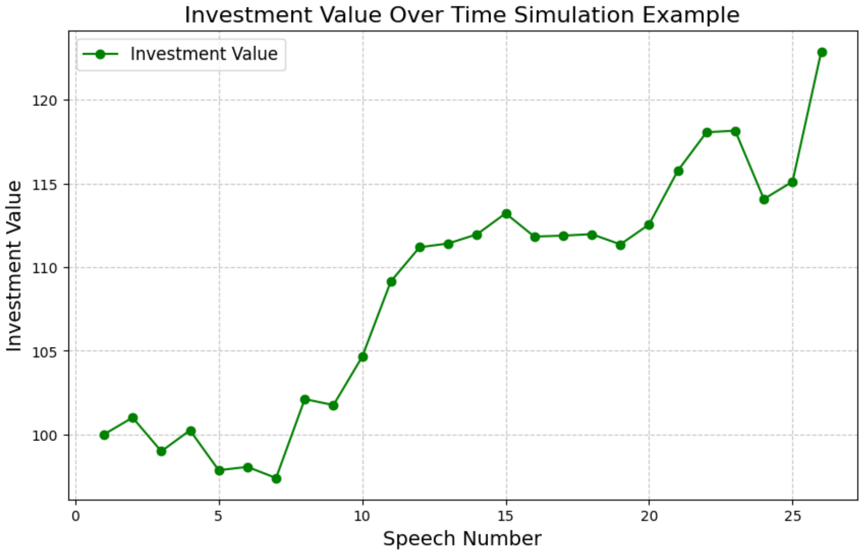

Figure 3A

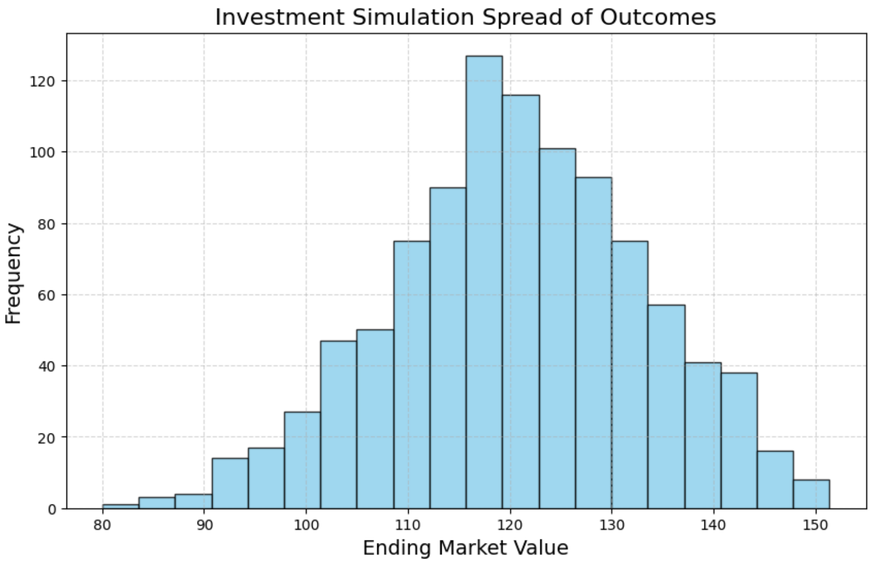

Figure 3B

5 Conclusion

This project demonstrates the feasibility of applying Natural Language Processing techniques to relevant public speeches to predict future stock market trends. The limitations of this analysis, however, lie in the relatively small data set used for model training with only 25 historic public speeches being considered. In a future analysis, it may be beneficial to build upon these results by including larger datasets. This could be achieved in various ways including:

1) Collect data from a longer time period, including all historical speeches by the chairman of the Federal Reserve, including and before Chair Powell
2) Collect data from more diverse sources within the same time period, for example include published speeches from other members of the Federal Reserve or the greater US government and evaluate how various speakers' words impact the market.

I would recommend using the RF model that has been re-trained on 100% of the data available here to perform market predictions for future speeches, and iteratively re-training to include all historical speeches for each prediction moving forward from there. As we move forward, continued refinements and expanded data sources hold promise for further enhancing the model's predictive capabilities. This project represents a demonstration that applying data science to the factors that move financial markets can help create a market edge, and offers a framework for future investigations in this domain.

6 References

Federal Reserve (speeches): https://www.federalreserve.gov/newsevents/speeches.htm
Stock MArket Data Source: https://www.investing.com/indices/us-spx-500-historical-data