

Sentimental Analysis of Reddit for
the Prediction of the Federal Funds Rate

Gokul Ganesan, Devesh Jaiswal, Derek Sun

Abstract

Macroeconomic changes tend to have a major influence on the financial, economic, and business world. Interest rate prediction has been a priority for stakeholders, investors, policymakers, and individuals. It is essential since it allows these different individuals to make more informed decisions regarding borrowing, lending, financial planning, and monetary policy. It is a measure that affects the overall health of the economy nationally and other economies abroad as the US is a large open economy that does international trade, participates in exchange markets, and is embedded in economic interdependencies. However, interest rate prediction has been difficult due to the influence of disruptive events and other macroeconomic variables. Furthermore, interest rates tend to fluctuate daily expressing excess volatility. There is a clear need to define an efficient algorithm for short-term interest rate prediction. Through previous research, it has been proven that social media sentiment analysis can provide predictive power in economic and financial variables. The goal here is to see if we can use the sentimental analysis of Reddit, incorporating multiple economic variables, to predict the federal funds rate. We will use a regression model involving and incorporating sentimental analysis to get the most accurate prediction for our model.

Introduction:

The federal funds rate is an interest rate controlled by the central bank of the United States which is the federal reserve. This is a macroeconomic variable that can be thought of as the measure of the cost of borrowing or return from lending specifically regarding depository institutions.

The changes made by the federal reserve of the federal funds rate tend to affect other interest rates like bond yields, mortgage rates, prime rates, business loan rates, and more. Therefore the prediction of the federal funds rate tends to be an area of interest to individuals, policymakers, investors, and stakeholders. It would allow them to make more informed decisions regarding financial strategy. There is this need for an efficient algorithm for federal funds rate prediction but it tends to not be easy.

In recent years we have further seen that this can be affected by disruptive events. For example, during covid-19 the federal funds rate had reached a sudden low since the financial crisis of 2008 which was another key event that caused the federal funds rate to drop. Sentiment analysis is applicable here as it has been proven to be helpful for works like event detection and trends of financial markets. By using sentiment analysis we try to gauge consumers during these events before changes are made by central banks and for an overall idea of how consumers feel regarding the economy.

Furthermore, the change the federal reserve decides to make regarding the federal funds rate is also affected by many macroeconomic variables such as inflation, consumer expectations, unemployment, gross domestic product, net capital outflows, and more. Due to the various amounts of macro variables that may have a direct/indirect effect on the federal funds rate we had to devise a set of them which may be of most importance. To do this we used related works similar to ours and concepts taught from economic courses such as the quantity theory of money.

By incorporating sentiment analysis and historical macroeconomic data into our regression we will get a gauge of how social media data can help us in our prediction of the federal funds rate.

Literature Review:

Before embarking on this project we searched for related works to make sure our project idea was viable. For instance, in the article “An Efficient Deep Learning Based Model to Predict Interest Rate Using Twitter Sentiment” and “Interest Rate Prediction with Twitter Sentiment” Twitter was used along with historical data to predict daily interest rates. The goal was to test which model would make the best use of Twitter data to calculate interest rates. However, these researchers also specified periods of disruptive or what they called mega events such as the US elections or Brexit. Furthermore, they applied it to countries around the world such as the UK or Switzerland. They compared support vector regression, linear regression, and deep learning models. They found that linear regression was not able to take advantage of the Twitter sentiment data. Despite this, they were able to prove that the deep learning model provided the most accurate results in their predictions. In addition, social media proved helpful in the predictions as they were able to show that interest rates are responsive to the social media response of investors.

Another example was done by two federal reserve researchers who wanted to investigate how Twitter can help detect policy changes made by the federal reserve. The article is called “More than Words: Twitter Chatter and Financial Market Sentiment:” and was published in June of 2023. The recency of this article proved that our idea was viable. Essentially they created a Twitter Financial Sentiment Index using keyword clustering and groupings based on semantic similarities from Twitter and used FinBert to get sentiment value for each tweet. The model that was used was a multivariable time regression model they were able to find that the Twitter Financial Sentiment Index helped predict the size of restrictive monetary policy surprises, while it is uninformative on the size of easing shocks. Furthermore the closer the day was to an FOMC meeting the more accurate the model was.

Therefore, due to the multiple related works found we were reassured that our idea should be pursued. We took a few motivations regarding how to approach social media data and what other historical data may be useful in our prediction.

Datasets:

As discussed numerous variables may influence the federal reserve's decision in making a change in the federal funds rate. Based on related works that others have done by inspecting the data used by them and concepts that are taught we were able to get a set of 10 macroeconomic variables that may help in our prediction.

Federal Funds Rate Data: This data represents the variable we want to predict. This is therefore known as the label since it is the category being predicted. This is historical monthly data that came from the Federal Reserve Bank of St.Louis. It contains monthly data from 1954 to 2023.

3-Month Treasury Bond Yield: This data is one of the macroeconomic variables that we included to predict the federal funds rate. It is a measure of the 3-month interest rate (short-term rates) of bonds. This was historical data that was daily ranging from 1954 to 2023. This data is from the Federal Reserve Bank of St.Louis as well.

Consumer Price Index(CPI): This data is another macroeconomic variable used to predict the federal funds rate. It is a measure of the average change of prices over time of a basket of goods and services specifically for urban consumers. This data was historical data that was monthly from 1913 to 2023 that came from the Bureau of Labor Statistics.

Inflation: This dataset corresponds to macroeconomic variables known to be highly related to interest rates. Inflation is a measure of the rate of increase in prices over time where there is a decrease in the purchasing power of money. This dataset is monthly data from 1914 to 2023 from Ycharts.

Inflation Expectations: This dataset is another macroeconomic variable derived from the quantity theory of money. It is a measure of the consumer's expectations of inflation where if there are sharp increases in expectations the federal reserve tends to increase rates. This dataset is monthly data from 1978 to 2023 and is from the Federal Reserve Bank of New York.

Investor Sentiment: This dataset is a macroeconomic variable that was included to predict the federal funds rate. This variable is a measure of the bull-bear spread which is a good indicator of how investors feel regarding the state of the economy. This dataset is every three days from 1978 to 2023 from Ycharts.

Net Capital Outflows: This dataset is a macroeconomic variable that refers to the international state of the US economy. It is a variable that measures the interaction or trade between the US and other countries explicitly about money. It is a monthly dataset from 1978 to 2023 from the Federal Reserve Bank of St.Louis.

Personal Consumption Expenditure(PCE) Price Index: This is a dataset similar to CPI but used as a main index by the federal reserve. It is a variable that measures the total value of goods and services purchased by individuals and households within a country over a specific period. It is a monthly dataset from 1960 to 2023 from the Federal Reserve Bank of St.Louis.

Unemployment: This dataset is another macroeconomic variable that was used to help predict the federal funds rate. It is a measure of the percentage of unemployment in the United States. This dataset is a monthly dataset from 1948 to 2023 from the Federal Reserve Bank of St.Louis.

Gross Domestic Product(GDP): This data is a very important macroeconomic variable that was included in the prediction. It is a variable that measures consumption, investment, government spending, and net exports. It is monthly data from 1992 to 2023 from Ycharts.

Reddit WallStreetBets Subreddit Data: This dataset contains all comments posted on the popular retail stock tracking subreddit r/wallstreetbets from April 2012 to June 2021.

Data Preprocessing and Feature Selection

To use the historical data for our model many preprocessing techniques had to be applied to come out with the best outcome. We started with 10 datasets of different macroeconomic variables.


First looking at the datasets we noticed that each started at different periods and this would have to be normalized to represent the same range of time. The way we approached this at first was to find the highest minimum date among all the datasets. Then go through each dataset and delete rows that were earlier than this date. However, this ended up reducing the number of data points we had dramatically so instead we took a different approach. We found the lowest minimum date amongst all datasets and then for each set we added the necessary rows up to their earliest date so they started on the same date. This ended up adding more data points which were good for us since our historical data sets were already small. After this first step, due to the added rows, we had to deal with missing values. As learned we could replace it with the average of the column, minimum of the column, maximum of the column, or zero. We ended up replacing these missing values with zeros. This is because we were planning to convert the raw values of our macroeconomic variables to percent changes from the previous period.

Therefore for the added rows with missing values, this would end up becoming zero regardless.

Next, we noticed that some of our data were monthly, weekly, daily, every three days, etc. Specifically the federal funds rate we are trying to predict is monthly so we wanted all the datasets to be monthly. So to do this we calculated average monthly values for each dataset to be normalized according to the federal funds rate data.

Furthermore, once we had the cleaned and normalized version of each dataset we applied feature selection. The problem we anticipated was that many of the macroeconomic variables we chose may be highly correlated with each other. This would make our model prone to overfitting and create a bias in the accuracy result. At first, we were thinking of using dimensionality reduction techniques such as PCA, however, interpretability was important to our model. Instead, we used a high feature correlation threshold. Essentially we calculated the Pearson correlation between each dataset. Then we calculated the

average correlation amongst all features which was set as the threshold. Therefore, for features that had correlations higher than this threshold they were removed while others were kept. This helped reduce highly inter-correlated variables and reduced our 10 datasets to 7. Then we integrated our processed dataset and this yielded one final dataset that represented the features and the class label that we wanted to predict.

 final (1).csv

```
Date,CPI,3 month Tbill Rate,Inflation,Unemployment,US GDP,Net Capital Outflows,Fed Funds Rate(Class Label)
2023-04-01,0.2518435,6.26262626,-17.903576,8.82352941,0.49710034,102.5,4.83
2023-03-01,-0.5033574,5.76923077,-1.0972919,-2.8571429,0.61049186,-84.70781,4.65
2023-02-01,-0.3299005,-0.8474576,-17.40672,-2.7777778,0.22383822,540.20979,4.57
2023-01-01,-0.5551124,3.05676856,-5.8423426,5.88235294,0.20949847,-86.579071,4.33
2022-12-01,-0.7931945,6.51162791,-0.6863535,-2.8571429,1.13580624,18.4546971,4.1
2022-11-01,0.3079546,0.70257611,-9.2246459,-2.7777778,-0.3106335,482.200647,3.78
2022-10-01,0.10110476,5.17241379,-8.1997056,-2.7027027,0.80120562,-88.788099,3.08
2022-09-01,-0.4040106,26.0869565,-5.5634807,5.71428571,0.82720561,79.5439739,2.56
2022-08-01,-0.2146169,12.195122,-0.7382575,-5.4054054,-0.2650603,594.570136,2.33
2022-07-01,0.03545249,22.6495726,-3.0745589,5.71428571,1.58266923,-87.890411,1.68
2022-06-01,0.01181331,40.9638554,-5.9052076,-2.7777778,-0.155976,13938.4615,1.21
2022-05-01,-1.3549953,46.9026549,5.57361767,0,1.15572152,-99.128686,0.77
2022-04-01,-1.0903331,36.1445783,3.90986366,0,0.89933372,-8.2410824,0.33
2022-03-01,-0.5551539,62.745098,-3.3233831,0,0.14678319,-44.731475,0.2
2022-02-01,-1.3175469,37.8378378,0.52993864,-5.2631579,0.92231881,-661.45038,0.08
2022-01-01,-0.9051305,54.1666667,5.23001644,-5,0.88302321,-123.40331,0.08
2021-12-01,-0.834436,300,6.302939,2.56410256,-0.0255078,56.5734266,0.08
2021-11-01,-0.3063106,20,3.33969746,-7.1428571,0.88212393,-633.58209,0.08
2021-10-01,-0.4889404,0,9.43602437,-6.6666667,0.42437329,-129.45055,0.08
2021-09-01,-0.8239662,25,15.4277127,-6.25,1.97222636,-27.777778,0.08
2021-08-01,-0.2708614,0,2.64696361,-7.6923077,0.64969893,300,0.09
2021-07-01,-0.2061652,-33.333333,-2.128413,-3.7037037,1.13621722,-70.08547,0.1
2021-06-01,-0.4787493,20,-0.4822406,-8.4745763,0.37566229,4.0513834,0.08
2021-05-01,-0.9205141,400,7.98765199,1.72413793,0.68712481,-30.874317,0.06
2021-04-01,-0.7953342,0,20.0254826,-4.9180320,1.08733737,181.652893,0.07
2021-03-01,-0.8151909,-66.666667,58.7792961,0,0.61661461,-31.702728,0.07
2021-02-01,-0.7033453,-25,56.2939983,-1.6129032,2.94606884,-17816.667,0.08
2021-01-01,-0.5444577,-33.333333,19.745678,-1.5873016,-0.7835762,-100.28024,0.09
```


Sentiment Analysis

Originally, we intended to use data picked from Twitter that contained a list of keywords, but due to Twitter's current API pricing, this was found to be unfeasible. Twitter's free tier allows for far too few tweets to obtain a dataset spanning the course of years. Additionally, one of Twitter's changes had broken the most commonly used web scrapers. The best alternative, Reddit, had also begun charging for API access. Because of this, the best publicly available dataset that suited our purposes was an archive of Reddit's WallStreetBets subreddit. The dataset contained many different columns, including upvote count, created_utc, awards, usernames, etc. For our purposes, the only two relevant columns were the "created_utc" column and the "body" column, which contained the text. There were also a significant amount of unavailable comments from deleted accounts, which were removed. After converting the utc date to a usable date and deleting all other irrelevant columns, the data was ready for sentiment analysis.

For sentiment analysis, we used VADER. VADER was the best selection for us because it has already been trained on social media text, and is able to take emoji usage and common internet slang into account, a significant benefit for us as WallStreetBets users do commonly use slang and emojis. VADER rates each comment's sentiment between -1, for most negative, and 1, for most positive. After each comment was analyzed, we obtained the mean value for each month and placed it into the dataset along with the data selected from feature selection.

Regression Analysis

To determine whether the sentiment analysis improved the ability of a regression model to predict the federal funds rate, we ran a multivariate regression on the dataset without the sentiment analysis, and then compared the results to a regression on the dataset that does have the sentiment column.

OLS Regression Results						
=====						
Dep. Variable:	FFR	R-squared:	0.019			
Model:	OLS	Adj. R-squared:	-0.038			
Method:	Least Squares	F-statistic:	0.3328			
Date:	Sat, 19 Aug 2023	Prob (F-statistic):	0.918			
Time:	00:39:34	Log-Likelihood:	-129.56			
No. Observations:	111	AIC:	273.1			
Df Residuals:	104	BIC:	292.1			
Df Model:	6					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	0.7118	0.092	7.764	0.000	0.530	0.894
CPI	-0.0548	0.260	-0.211	0.834	-0.571	0.461
3 month Tbill Rate	-0.0004	0.001	-0.506	0.614	-0.002	0.001
Inflation	0.0005	0.001	0.467	0.641	-0.002	0.003
Unemployment	-0.0039	0.006	-0.678	0.500	-0.015	0.008
US GDP	-0.1155	0.097	-1.192	0.236	-0.308	0.077

We can see that the R-squared and F-statistic for the regression without the sentiment analysis are very low. Additionally, the RMSE for the model's prediction is 0.77747.

OLS Regression Results						
=====						
Dep. Variable:	FFR	R-squared:	0.329			
Model:	OLS	Adj. R-squared:	0.284			
Method:	Least Squares	F-statistic:	7.219			
Date:	Sat, 19 Aug 2023	Prob (F-statistic):	5.03e-07			
Time:	00:39:24	Log-Likelihood:	-108.46			
No. Observations:	111	AIC:	232.9			
Df Residuals:	103	BIC:	254.6			
Df Model:	7					
Covariance Type:	nonrobust					
=====						
	coef	std err	t	P> t	[0.025	0.975]

const	1.4203	0.128	11.111	0.000	1.167	1.674
CPI	0.1284	0.218	0.590	0.557	-0.303	0.560
3 month Tbill Rate	-8.379e-05	0.001	-0.139	0.890	-0.001	0.001
Inflation	0.0001	0.001	0.165	0.869	-0.002	0.002
Unemployment	-0.0101	0.005	-2.071	0.041	-0.020	-0.000
US GDP	-0.1736	0.081	-2.145	0.034	-0.334	-0.013
Net Capital Outflows	8.521e-06	3.25e-05	0.262	0.793	-5.59e-05	7.29e-05
Sentiment	-6.2276	0.902	-6.902	0.000	-8.017	-4.438

For the regression containing the sentiment analysis, we can see a significant improvement in the R-squared and F-statistics, and the RMSE has been lowered to 0.642882. Additionally, we can see that the Sentiment column has a p-value of 0.00, a signifying highly significant variable. We can conclude that the sentiment has improved the quality of the prediction significantly.

Limitations

With the current increase in social media scraping restrictions and high API prices for social media platforms like Twitter and Reddit, acquiring niche social media data on economic opinions posed an obstacle as there wasn't previously collected data on the topic. To get around this, we used an archive of the r/wallstreetbets subreddit. Twitter was the ideal data source as it provides a large variety of opinions and is more generalizable to the population, which would allow for a more accurate analysis of the public economic perception that would likely better suit our model since the 'wallstreetbets' subreddit comments may be more biased in terms of sentiment. Additionally, the fact that our data was collected from an archived subreddit means that it could not include more recent comments/events and so would have to be based on outdated opinions. Nonetheless, these comments allow us to create a sort of "proof-of-concept" model for understanding the efficacy of incorporating Natural Language Processing in the analysis of interest rate trends.

Identifying the ideal method of collating the social media data with the historical data analysis to best predict short-term inflation rates also proved to be a difficult task since it requires matching the dates of the historical data with the dates of the comments and identifying the significance/weight of the comments on our predictive model. This is made even more difficult because the comments - which may be a response to the economy and inflation rates - may be offset from the historical data by some measure of time. Additionally, given the smaller sample size of our social media data, there may be large fluctuations or margins of error in our social media analysis as the sentiment of the data could largely vary based on a relatively small number of comments in any given month, making it a difficult task to account for these arbitrary fluctuations in sentiment.

Conclusion

In our exploration, we aimed to identify the efficacy of using public opinion as a feature in predicting short-term interest rates through natural language. This predictive analytics investigation provides a better understanding of the place that natural language processing has - even in quantitative contexts like analyzing interest rates.

Our findings from this exploration are encouraging. When comparing evaluations of the two models, we can see the model that incorporates natural language has a reduced RMSE of 0.642882 when compared to the original 0.77747, suggesting that the sentiment played a role in making more accurate predictions - improving on the base model. Additionally, the p-value of the sentiment feature is negligibly small which further encourages the idea that sentiment as a feature (from the data we collected) is closely tied to the variation in interest rates as the low p-value shows that variations in sentiment and interest rates from our data has $p < 0.000$ chance of being sampled as a random occurrence. This thoroughly establishes our notion that social media data holds valuable information in explaining a portion of the variation in interest rates. This is further supported by the magnitude of the coefficient of the sentiment feature (-6.2276) in relation to other features which range between an absolute value of 0 - 1.5, showing that the sentiment feature contributes significantly to the predictive power of the model. To go along with this, the R-squared statistic rises from 0.019 to 0.329 and the F-statistic rises from 0.3328 to 7.219, meaning that the addition of sentiment in the second model helps explain a much larger portion of the variation in the interest rates data and is therefore better fit to the interest rate data than the base model, in agreement with all the other supporting evidence that sentiment is a great feature in the interest rate prediction model.

Through this statistical evaluation, we conclude that the addition of sentiment to our base model provided massive value to the reliability and accuracy of its predictive capabilities. This result builds our understanding of - not only the implications of public opinion on short-term interest rates - but also the practical advantage that sentiment analysis can have in enhancing predictive models.