Stock Price Prediction using Central Bank Speeches

Mickael HAGEGE
Paris Dauphine - PSL University

Abstract

Central banks worldwide regularly issue speeches, offering insights into their analysis of the global financial landscape. These speeches are closely watched by financial stakeholders globally, and they hold significant sway over financial market trends. By employing various Machine Learning (ML) techniques, particularly using Natural Language Processing (NLP), the correlation between these speeches and movements in two key stock market indices, the VIX and the EURUSD 1M, has been effectively established. An accuracy (Acc) rate of 0.5850 has been achieved in predicting market trends and the values of these indices have been accurately forecasted, with a root mean square error (RMSE) of 0.3494. This successful integration of ML and NLP underscores the potential of advanced data analytics to provide valuable insights into financial market dynamics.

1 Introduction

The goal of this project, in collaboration with Natixis, was to predict the evolution of two financial market indices, the VIX and the EURUSD 1M, based on speeches delivered by the Federal Reserve (FED) and the European Central Bank (ECB) over the past 10 years, as well as daily data from various key financial markets. The VIX serves as an indicator of market volatility in the US, while the EURUSD 1M reflects the volatility of the USD/EUR exchange rate.

These two indicators appear to be closely correlated with the prerogatives of central banks and thus with the content of their speeches.

Indeed, central bank communications can impact various key economic factors, from interest rates to monetary policy, inflation expectations, credit, debt, and overall financial leverage for both private and public sectors. Given that these speeches can influence key macroeconomic factors and move financial markets, the ability to decipher and interpret central bank jargon has become a key area of interest for financial analysts and economic actors.

Two tasks were assigned for this challenge: a regression task to predict the value of the indicator on the day following a 20-day period and a binary classification task to predict whether it would increase or decrease. These tasks were respectively evaluated using the root mean square error (RMSE) and accuracy metrics.

For each stock index, training data in JSON format containing several examples, each with a 20-day history of stock prices and speeches, was provided.

The first challenge, common to all NLP projects, was to convert these strings into a format usable by a computer, and the second, specific to this project, was the length of these speeches. Indeed, the most common pre-trained NLP model, Bert, can only accommodate up to 512 tokens. The speeches had a median word count of 2,833 words each, which would have led, after data processing, to a token count exceeding Bert's capacity.

To address these challenges, it was ultimately decided to retain only the speeches from the last day (J20), those from a week before (J13), and to use the TF-IDF weighting method.

2 Model

Initially, it was necessary to adjust the data format and organize them into a dataframe that grouped the indicators and speeches by day, making them usable for the algorithms.

For the regression task, only the daily stock index values were retained, excluding the speeches, to be fed into the models. The selected model yielding the lowest root mean square error (RMSE) was a Stacked Long Short-Term Memory Network (LSTM), trained on the dataset for 100 epochs using a Rectified Linear Unit (ReLU) as the activation function. The resulting RMSE score on the test set was 0.3494.

For the classification models, it was decided to apply them to both the speeches and the financial indices. The focus was on the last speech for each period, either on the 20th day or, in cases where no speech was available, the last speech issued by a bank. Subsequently, data cleaning was performed on these speeches, including removing non-English speeches, punctuation, casing, stopwords, and conjugations. Following this preprocessing step, a TF-IDF transformation was applied to the cleaned speeches. Once the term importance scores were obtained, the stock index prices over the 20-day period were incorporated. For prediction, a Random Forest model with 200 estimators, a maximum depth of 10, and a fixed seed of 0 was employed. The model achieved an accuracy of 0.5850.

3 Work Phase

3.1 Data

First and foremost, it was crucial to examine the data structure: were there more upward or downward trends? Did the speeches exhibit similarities? The training data provided is in the form of two .json files, one for each indicator, totaling 1,254 lines. Each line contains 20 stock values along with speeches from both the FED and the ECB over these 20 days. It was observed that the distribution of stocks for both market indicators was similar, as well as their proportions of upward trends.

- Distribution of the target stock on Eurus [1]
- Distribution of the target stock on VIX [2]
- Proportion of upward trends on Eurus [3]
- Proportion of upward trends on VIX [4]

A closer look was taken at the speeches, including the writing style and the most frequently used words. Firstly, word clouds were generated for the ECB speeches:

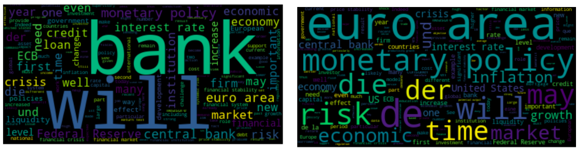


Figure 1: ECB Wordcloud

Subsequently, word clouds were generated for the FED speeches:

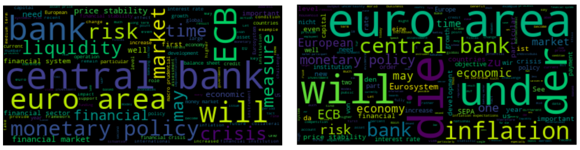


Figure 2: FED Wordcloud

The presence of speeches written in languages other than English was also noted. Using the `langdetect` package, speeches in Italian, German, and French were identified. Therefore, the decision was made to remove the cases containing these speeches to optimize the TF-IDF analysis.

3.2 Results, Errors, and Analysis

For this project, various models were prepared and submitted to improve accuracy and reduce RMSE. A summary table of these different approaches and models, along with the results of submissions, has been compiled [5]. Through this table, the evolution of the approach throughout the project and the various methods tested can be tracked.

4 Discussion

Significant challenges were encountered when attempting to utilize pre-trained models designed for handling long texts. Regardless of the optimizations and tricks employed, models like BigBird, LongBird, and LongTransformers consistently ran into the limitations of computers' capacities (RAM). As for Bert, the limitation to 512 tokens yielded less than satisfactory results, falling short of those obtained with TF-IDF.

Another issue faced was overfitting. In a model described in Table [5] using a Random Forest with the same parameters as the final model but leveraging the entire concatenated speech set for TF-IDF, the accuracy on the `train` file was 0.84. However, when this model was tested against the `test` file from the platform, an accuracy close to 0.53 was obtained. A probable explanation could be the potential difference in term frequency between the `train` and `test` files.

Ultimately, it was observed that the model could detect certain indicators in central bank speeches reflecting the financial reality of the market. However, it was not reliable enough to be used in financial decision-making. Considering the final values of Accuracy (0.5850) and RMSE (0.3494), it can be concluded that the model is not yet sufficiently effective.

5 Conclusion

This project has underscored the significant impact of Natural Language Processing (NLP) within the finance domain. It has illuminated that speeches hold more predictive power for classification than the numerical values of financial indices. Furthermore, it has emphasized the critical role of data preprocessing through various cleansing techniques, including language identification for speeches. Throughout the model implementation phase, the necessity of computational resources, such as processors (including GPUs and TPUs available on Google Colab), and the RAM of our machines for program compilation, has been underscored.

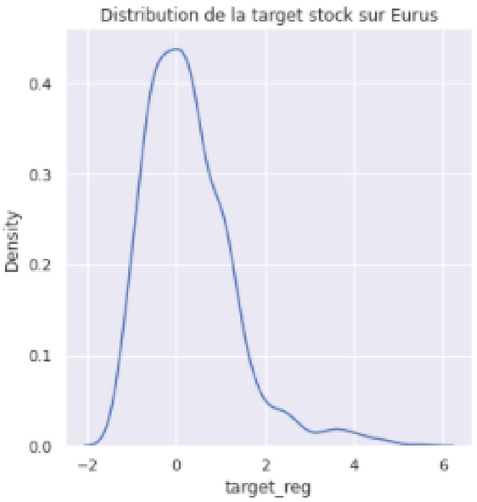
A prominent avenue for enhancing future endeavors lies in the effective utilization of pretrained models specifically tailored for processing lengthy textual data.

Despite the challenges encountered, leveraging the TF-IDF weighting method, coupled with a Random Forest for classification and a Stacked LSTM for regression, yielded commendable results, achieving a precision of 0.5850 and a root mean square error of 0.3494, respectively.

Acknowledgments

Gratitude is extended to Ms. Meziani for her instrumental role in initiating this project in collaboration with Natixis. This invaluable opportunity enabled the application of academic knowledge acquired throughout our tenure at Paris Dauphine University to real-world datasets. Appreciation is also owed to Syrielle Montariol for generously sharing her expertise in NLP and machine learning methodologies during collaborative sessions. Finally, sincere thanks are extended to Natixis for generously providing access to real-world data for this project, and to our esteemed professors whose dedicated instruction facilitated the effective utilization of this invaluable resource.

Appendices



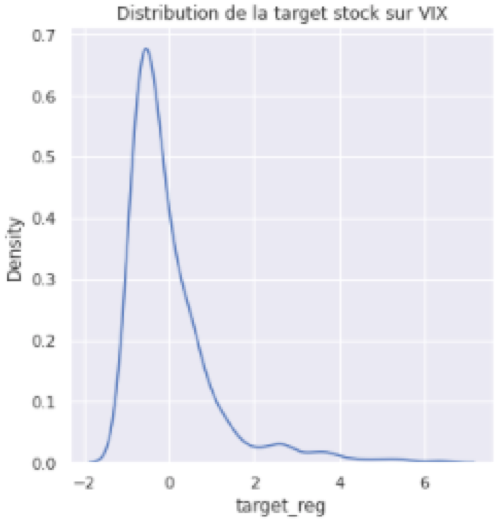
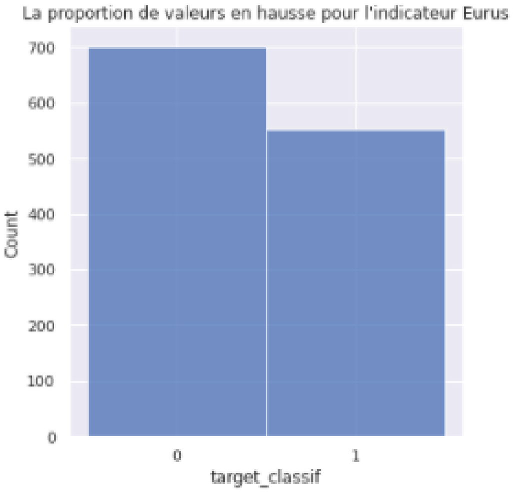


Table 1: Distribution of stocks for the Eurus indicator

Table 2: Distribution of stocks for the VIX indicator



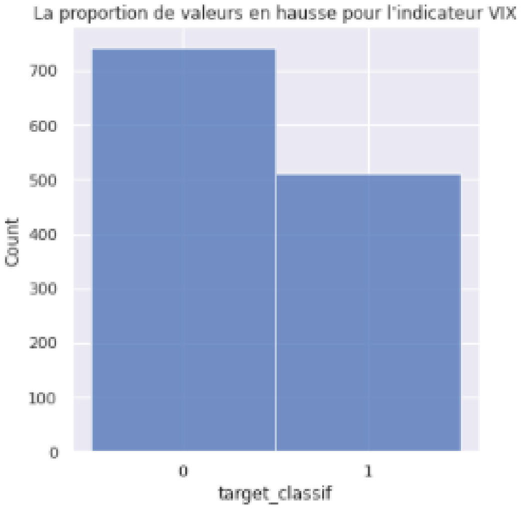


Table 3: Proportion of value increase for the Eurus indicator

Table 4: Proportion of value increase for the VIX indicator

Method used			Results		
Speech Processing	Classification	Regression	Accuracy	RMSE	Analysis
TF-IDF on all concatenated texts of a line regardless of the issuing bank. However, the number of words is limited to 10,000.	RandomForest	RandomForest	0.6425	0.9031	We can believe that the high number of texts and therefore words in the `train` file is the main reason explaining these poorer results on the `test` file.
TF-IDF on all concatenated texts of a line regardless of the issuing bank. However, the number of words is limited to 10,000.	Gradient Boosting	Gradient Boosting	0.5925	0.4382	Here, we can observe that the root mean square error has decreased, which is positive. However, we also notice a decrease in accuracy. Consequently, we decided not to choose Gradient Boosting.
TF-IDF on the last text. We decided to consider only the top 5,000 recurring words.	RandomForest	StandardScaler	0.685	0.3716	Here, we see that the results are similar to the final model, but the RMSE is slightly higher (0.3494 for the final model)

Table 5: Summary table of different tests and results.