Market Mood Analysis: Analysing the Impact of U.S. Central Bank Statements on Stock Market Trends

Konstantin Zuev
Department of Computer Science
Dalhousie University
Halifax, NS, Canada
kn905954@dal.ca

Mohammed Usama Jasnak
Department of Computer Science
Dalhousie University
Halifax, NS, Canada
mh659974@dal.ca

Pallavi Singh
Department of Computer Science
Dalhousie University
Halifax, NS, Canada
pall.singh@dal.ca

Rutvik Vengurlekar
Department of Computer Science
Dalhousie University
Halifax, NS, Canada
rt762740@dal.ca

Abstract

This study explores the relationship between the sentiment of U.S. Federal Reserve communications and stock market trends, with a focus on market volatility. Leveraging a comprehensive dataset of over 1,100 Federal Reserve announcements, we employed a range of Natural Language Processing (NLP) techniques, including rule-based (VADER), machine learning (Logistic Regression, Random Forest), deep learning (LSTM, GRU), and transformer-based models (DistilRoBERTa) to classify the sentiment of these communications as positive or negative. Our findings reveal that transformer-based models perform better than traditional approaches in accurately capturing sentiment nuances. However, the sentiment derived from Federal Reserve statements shows no impact on broad market indices like the S&P 500, while a minimal negative effect is observed on Bank of America's stock performance. These results suggest that although advanced NLP models can effectively analyse sentiment, the influence of central bank communications on financial markets is subtle and intertwined with multiple factors. Future research should consider more nuanced sentiment classifications and incorporate additional contextual variables to further understand this complex relationship.

1 Introduction

The stock market is inherently volatile and is influenced by numerous economic and political factors, including the actions, economic announcements and policy statements from influential institutions. Among these, the U.S. Central Bank stands out because its announcements seem to have a significant influence on both domestic and global markets. Policy announcements and public communications by these officials are closely scrutinised by market participants, as they often signal future monetary policies, rate adjustments, and economic outlooks. The precise mechanisms by which central bank communications affect stock market trends remain an open question. Traditional analyses often focus on explicit policy changes such as interest rate adjustments, leaving the more subtle but potentially impactful role of sentiment largely unexplored.

How does the sentiment in Federal Reserve speeches correlate with stock market volatility? Can this relationship be quantified to provide actionable insights for investors and policymakers? And most importantly, how can advances in Natural Language Processing (NLP) and machine learning be used to address these questions effectively? This gap in understanding motivates the current study,

Preprint. Under review.

which aims to explore the relationship between the sentiment of the U.S. Central bank speeches and stock market behavior, specifically stock volatility. The objective of this study is to bridge this gap by analysing the sentiment of Federal Reserve communications and examining its impact on stock market behavior. Specifically, we aim to achieve the following:

1. Develop a framework for sentiment analysis of Federal Reserve speeches, classifying them into categories such as "positive" and "negative".
2. Investigate the relationship between sentiment and stock market trends.
3. Quantify the significance of sentiment-driven market behavior using statistical analysis. We will also consider the role of macroeconomic factors to isolate the specific impact of sentiments.

The significance of this research lies in its potential applications for traders, financial analysts, and policymakers. For traders and investors, the ability to anticipate market reactions based on the sentiment of Federal Reserve statements can offer a competitive edge, enabling more informed investment decisions. For policymakers, understanding how market participants interpret their communications can help in crafting messages that minimise unintended volatility. This study can also help in creating communication strategies that minimise market disruptions. Moreover, the relationship between central bank communications and the stock market may provide valuable insights into the broader economic impacts of such statements, with the stock market serving as a proxy for overall economic health.

2 Related Work

2.1 Assessing Impact of Central Bank Communications

Central bank communications are not limited to explicit policy announcements. They often include subtle signals embedded in the tone, choice of words, and overall sentiment of the statements. For instance, a slightly more optimistic tone in a Federal Reserve speech could boost investor confidence, leading to increased stock prices, while a cautious tone might trigger market sell-offs. This implicit power of sentiment to move markets has been the subject of extensive academic and practical interest. Studies like those by Hayo and Neuenkirch (2013) have demonstrated that Federal Reserve communications influence market expectations even without concrete policy actions. Subtle shifts in tone or sentiment within these communications can lead to market reactions, altering stock prices, trading volumes, and volatility levels [1]. Similarly, Jansen and de Haan (2005) explored how the European Central Bank's (ECB) statements impacted the euro-dollar exchange rate, concluding that central bank communication, even without accompanying policy actions, can cause significant market movements [2].

To identify the impact of Federal Reserve Communications on Financial Markets, we plan to utilise different Natural Language Processing (NLP) approaches for sentiment analysis and classify communications into categories such as "positive" and "negative", providing insights into how market agents respond to central bank messaging. There have been a few prior studies using various NLP and machine learning approaches to evaluate the sentiments of textual data.

2.2 Rule-Based Sentiment Analysis

Rule-based sentiment analysis (lexicon-based approach) offers a straightforward method for classifying text using predefined linguistic rules. This approach involves defining sets of words or phrases that represent positive, negative, or neutral sentiments and applying them to classify text based on the presence of such indicators [3]. Rule-based methods can serve as a foundational approach to text analysis, providing a structured way to parse sentiment in financial communications while reducing computational complexity. For instance, VADER model is another rule based model which computes sentiment scores based on a combination of lexical features (like words and punctuation) and grammatical rules [4].

However, rule-based approaches may struggle with domain adaptation, e.g. the word "tight" might be negative in general but could be positive in an economic context.

2.3 Machine Learning: Shallow and Deep Learning-Based Sentiment Classification

A more flexible approach to sentiment classification involves shallow models such as Naive Bayes, Random Forest, or XGBoost, which offer adaptability to complex language. There are a number of feature extraction techniques like N-grams, Bag of Words (BoW), and Term Frequency-Inverse Document Frequency (TF-IDF) that can be applied alongside these machine learning methods to make them suitable for processing text data [5].

Deep learning models provide several benefits for sentiment classification tasks. They excel at capturing intricate patterns in large datasets and can automatically learn relevant features, reducing the need for extensive feature engineering. Furthermore, deep learning architectures, such as Long Short-Term Memory Recurrent neural networks (LSTM), can be effective for domain-specific tasks like sentiment analysis of financial texts. These models can recognise hierarchical patterns in the data, leading to improved performance over traditional machine learning approaches. Then can also encode for bidirectional dependencies in text. This architecture has been successfully used for text classification problems. LSTM variants were found to outperform all shallow ML models in predicting sentiments from user tweets [6]. They were also found to be more precise than traditional models in classifying emotionally charged texts [7].

Additionally, transformer-based models such as BERT (Bidirectional Encoder Representations from Transformers) are often used for sentiment analysis of speeches [8]. BERT's architecture allows it to capture contextual meaning and dependencies between words more effectively, potentially improving classification accuracy. One of the key advantages of BERT is its ability to leverage transfer learning; it can be pre-trained on a vast corpus of text and then fine-tuned on our specific dataset. This transfer learning capability means that even with limited domain-specific labeled data, we can achieve robust performance, as the model has already learned a wealth of language patterns and structures during pre-training. This makes BERT particularly advantageous for our study, as it allows us to harness the power of deep learning while addressing the challenges of domain specificity in financial communications. Many transformer based models are also available on HuggingFace which are fined tuned on the financial datasets. `DistilRoberta-financial-sentiment` is a finetuned version of distilroberta model. It is finetuned on the financial phrasebank dataset and predict sentiments.

3 Problem Definition and Methodology

As discussed earlier, financial markets are complex systems influenced by a host of factors, with central bank communications being one potential determinant of market sentiment. The primary objective of this study is to investigate whether sentiments of Federal Reserve speeches correlate with stock market volatility. Specifically, we seek to address the following questions:

1. Can the sentiment of Federal Reserve speeches be accurately classified using NLP techniques?

2. Does the sentiment conveyed in these speeches have a measurable impact on stock market volatility?

3. How significant is the relationship between sentiments and market behaviour if we account for macroeconomic factors?

By developing a sentiment analysis framework and evaluating its impact on stock market dynamics, this study seeks to provide actionable insights into the role of such sentiments in market decision-making.

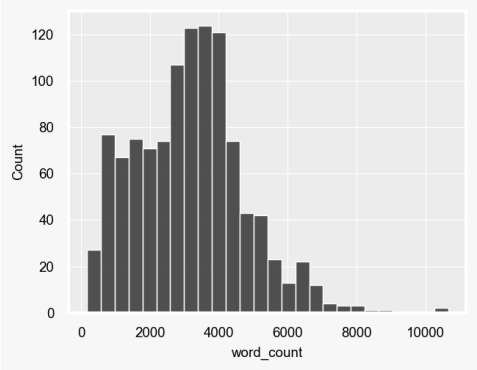
To address the research questions, the study is structured into two main components: first, to analyse and classify sentiments of Federal Reserve speeches using NLP techniques, and second, to examine the relationship between this sentiment and stock market behavior. The approach integrates data collection, data preprocessing, sentiment analysis, and statistical evaluation to ensure a comprehensive understanding of the problem. This section outlines the methodology in detail. Each component involves a series of steps utilising NLP techniques and statistical methods.

3.1 Data Collection

We utilised ‘selenium‘ framework to scrape announcements from the official Federal Reserve website (https://www.federalreserve.gov). The scraping process resulted in a dataset with 1,113 announcements providing us with a corpus of public communications by Federal Reserve officials, starting from 2006.

The dataset includes a total of 1,830,844 words and 29,441 unique words and thus provides us with rich textual data for analysis.

The distribution of the length of the text in each collected announcements are illustrated in Figure 1a, highlighting the volume and diversity of the dataset. Terms like inflation, risk, and price, were among the top 15 most frequently occurring words, which further highlighted the economic focus of these announcements (refer to Table 1b).



(a) Distribution of announcements lengths.

Word	Frequency
Bank	18,743
Financial	18,230
Market	15,018
Policy	13,06
Rate	12,419
Inflation	11,412
Federal	10,719
Risk	10,697
Reserve	8,818
Economy	8,025

(b) Most frequently occurring words.

Figure 1: Text distribution and word usage trends.

To further understand the sentiment distribution within the text, we applied a lexicon-based approach to words with frequencies exceeding 100 using TextBlob. This analysis yielded the following results: 2,114 words as “neutral”, 165 words as “positive”, and 82 words as “negative”. Unsurprisingly, a majority of the language used lacks strong emotional connotations with a smaller but noticeable presence of favorable or optimistic expressions.

3.2 Data Labelling

The dataset of news articles was reviewed and annotated independently by three team members. Each member assigned one of the sentiment categories — positive or negative to each article based on its content and tone. After all annotations were completed, the majority vote approach was used to determine the final label for each article. This method ensured that the labels reflected a consensus among the annotators, reducing individual biases and improving the reliability of the annotations. These labels were then used for training and evaluating the sentiment classification models. The final distribution was roughly 3:1, with statements labelled as "positive" being the majority class.

It is important to note that classifying central bank statements into positive or negative sentiment can be challenging. Central bank statements often contain balanced phrasing and hedging terms, which can obscure clear sentiment indicators. For instance, phrases like "moderate economic growth" or "potential downside risks" might be interpreted differently depending on the context, making it difficult to assign a definitive positive or negative label. Hence, the final label may change depending on how a statement is interpreted. In the future, it might be more accurate to use a scale system rather than binary labels.

3.3 Data Preprocessing

The steps undertaken for preprocessing are as follows:

3.3.1 Initial Cleaning

1. **Removal of Redundant Lines:** Speeches were truncated at lines containing disclaimers like "The views expressed here" or "These views are my own", ensuring that only relevant content was retained.
2. **Elimination of Unwanted Text Patterns:** Text artifacts, such as accessibility instructions and unrelated metadata (e.g., video playback instructions), were identified and removed.
3. **Exclusion of References and Footnotes:** Sections containing "References" or "Footnotes" were removed.

3.3.2 Content Filtering

1. **Bullet Point Removal:** Lines beginning with numbers or bullet points were stripped to focus on narrative content.
2. **Specific Pattern Matching:** Frequent phrases such as "Accessible Version" and "View speech charts and figures" were systematically removed using regex patterns.
3. **Short Speech Exclusion:** Speeches containing fewer than 100 words were excluded from the dataset to maintain a minimum threshold for meaningful analysis.

3.3.3 Normalisation

1. **URL Removal:** Links embedded within the text were stripped.
2. Converting to Lowercase: All text was converted to lowercase for uniformity during tokenisation.
3. **Removing Special Characters:** Non-alphanumeric characters (except for punctuation like ".", "!", "?") were removed to standardise the text.

3.3.4 Tokenisation and Text Processing

1. **Sentence Tokenisation:** The text was segmented into individual sentences using NLTK's sentence tokeniser for finer-grained analysis.
2. **Word Tokenisation:** Following that, words were extracted from the text using NLTK's word tokeniser.
3. **Stopword Removal:** Common English stopwords were eliminated to reduce noise in the data.
4. **POS Tagging:** Part-of-speech tagging was applied to identify the grammatical roles of words, enhancing contextual understanding.
5. **Stemming and Lemmatisation:** Words were optionally reduced to their root forms using either stemming or lemmatisation, depending on the analysis requirements.

The preprocessing steps were applied to the entire dataset. Later, some of these steps were evaluated to determine their necessity based on the out-of-sample performance of the Machine Learning models.

3.4 Sentiment Analysis

This section focuses on classifying Federal Reserve speeches into categories such as "positive" or "negative". To achieve this, we employed both traditional and more advanced methods, including rule-based models, machine learning (ML) algorithms, deep learning (DL) architectures, such as transformer-based models.

3.4.1 Rule-Based Sentiment Analysis: Vader Model

As a baseline approach, we used the VADER (Valence Aware Dictionary and sEntiment Reasoner) model which is a rule-based sentiment analysis tool specifically tuned to sentiments expressed in social media, such as Twitter, but it's also effective for general sentiment analysis tasks. It computes sentiment scores based on a combination of lexical features (like words and punctuation) and grammatical rules. The major advantage of this model is that it is quick and computationally efficient.

The VADER model generates sentiment scores ranging from -1 to 1, where the conventional thresholds classify scores below -0.05 as negative, scores between -0.05 and 0.05 as neutral, and scores above 0.05 as positive. In our study, however, we required a binary classification system with only "positive" and "negative" labels. So, we experimented with different approaches for interpreting the VADER scores. Initially, we considered using a threshold of 0 to distinguish between the two classes. Additionally, we experimented with normalising the VADER scores to a 0 to 1 range and applying a 0.5 threshold for classification. Unfortunately, these alternative approaches did not yield satisfactory results. So at the end, we adopted a simplified rule where any score below 0 is classified as negative and any score above 0 is classified as positive.

3.4.2 Machine Learning-Based Sentiment Classification

The ML based approach to sentiment classification in this study involves using supervised learning algorithms to classify Federal Reserve speeches. ML models at times overcome the limitations of rule-based models, such as their inability to capture nuanced linguistic patterns and domain-specific contexts and thus offer adaptability to complex language, which should be particularly useful in the context of classifying long and complicated announcements.

For feature extraction, various techniques were employed to represent the textual data effectively. TF-IDF (Term Frequency-Inverse Document Frequency) was applied to highlight the importance of words within individual speeches relative to the entire corpus, reducing noise from frequently occurring but less meaningful terms. Additionally, we experimented with creating TF-IDF for different ngrams, but for this task unigrams worked the best.

Three shallow machine learning models were implemented for sentiment classification:

1. **Logistic Regression (LR)**, chosen as a baseline for its simplicity. We used L2 regularisation to prevent overfitting and `class_weight == 'balanced'` that helps in handling imbalanced data by employing a weighted log-loss with inversely proportional to class frequencies.
2. **Naive Bayes (NB)**, which leverages a probabilistic framework.
3. **Random Forest (RF)**, an ensemble method effective at capturing complex relationships while mitigating overfitting. Random Search was used to tune hyperparameters. Similarly to LR, `class_weight == 'balanced'` was employed. It assigns weights inversely proportional to class frequencies, which affects the metric calculation during node splitting.

3.4.3 Deep Learning-Based Sentiment Classification: RNN Models

Deep learning-based sentiment classification leverages neural network architectures to capture complex patterns and contextual relationships within the text of Federal Reserve speeches. To handle the sequential nature of text data, two recurrent neural network (RNN) architectures employed are as follows:

1. **Long Short-Term Memory (LSTM)** as they have the ability to retain long-term dependencies through gated mechanisms and are particularly effective at understanding nuanced sentiment shifts across lengthy speeches.
2. **Gated Recurrent Unit (GRU)** as they are a streamlined variant of LSTMs, normally providing similar performance but with fewer parameters.

Both models were trained using GloVe word embeddings to encode semantic relationships between words, enabling them to understand contextual sentiment more effectively than traditional methods [9]. We used smaller embeddings with a dimension of 50, and the embeddings were not frozen during training.

To prevent overfitting, dropout layers were incorporated. AdamW was also used to introduce additional regularisation in the form of weight decay.

In the case of LSTM, we experimented with incorporating both the hidden and cell states in the classifier layer. Three strategies were adopted: using only the hidden state, combining the hidden and cell states through a subsequent linear layer, or summing them.

To combat class imbalance, we used weighted cross-entropy. Additionally, we introduced label smoothing to prevent the model from underestimating or overestimating probabilities. A larger penalty was given to the minority class, as it can help in improving model performance under class imbalance [10].

The models were trained over multiple epochs with early stopping criteria and a small batch size of 64.

Lastly, during training, the window length was limited to 150 tokens, whereas for inference, RNN models were allowed to classify statements based on the entire sequence length.

3.4.4 Deep Learning-Based Sentiment Classification: Transformers

We used the DistilRoBERTa fine-tuned financial news sentiment analysis model from Hugging Face which is trained for sentiment classification in financial text. The model is fine-tuned on the Financial PhraseBank dataset and can handle the unique challenges of financial language, namely domain-specific jargon and nuanced sentiment shifts. Being a lightweight version, it offers the computational efficiency of a smaller model while maintaining comparable performance to its larger alternatives. Since this model classifies text into three categories: positive, negative, and neutral, we made some changes to it's configuration to handle binary classification.

Since this transformer based model was fine-tuned on financial data, our initial approach was to utilise it directly. However, we observed that the model's performance was suboptimal in this configuration. As a result, our next step involved fine-tuning the model using our own dataset. Due to the model's limitation of processing a maximum of 512 tokens, we could not use some of the speeches contained more than 1,000 tokens. To address this issue, we implemented two approaches:

1. **Splitting:** We splitted the speeches into smaller segments of 500 tokens each, ensuring that each chunk retained the corresponding labels. The model was then fine-tuned using these segmented inputs.
2. **Truncation:** Alternatively, we employed the model's default truncation parameter. This involved keeping only the first 512 tokens for fine-tuning or prediction.

The results of both approaches are discussed in the subsequent section 4.

All shallow ML models were implemented using `scikit-learn`, and all DL models were implemented using either `PyTorch` or `Hugging Face`.

3.5 Stock Market Analysis and Correlation

Following the classification of speeches into sentiment categories, we intend to evaluate the influence of these sentiments on market behaviour.

Two-sample t-tests are conducted to determine first whether there is a statistically significant difference in means between announcements classified as "positive" or "negative". Before applying the tests, stock prices are differenced to remove the stochastic trend, calculated as $y_t - y_{t-1}$.

While statistical tests can determine the overall difference, they provide only a general measure of significance. To gain a more nuanced understanding, we will also account for numerous macroeconomic factors using regression analysis. This approach allows us to isolate the specific impact of sentiment on market behavior by controlling for other relevant variables, offering a more comprehensive view of how Federal Reserve communications affect stock market dynamics.

The following macroeconomic factors were collected:

1. **Macroeconomic Activity and Growth:** GDP, UNRATE (unemployment rate).
2. **Inflation and Price Levels:** CPIAUCSL (U.S. consumer price index).
3. **Monetary Policy and Interest Rates:** 10Y_Treasury_Yield (debt obligations).
4. **Market Sentiment and Risk:** VIX (Fear Gauge), UMCSENT (consumer sentiment).
5. **Currency and Exchange Rate:** DXY (U.S. dollar index).
6. **Housing Market Data:** CSUSHPINSA (U.S. housing price index).

7. **Commodities Prices:** Oil, Gold, Copper, Natural_Gas

Plots for these factors can be found in Appendix (Figure 3).

Since we were only interested in quantifying the effects of Federal Reserve announcements, all macroeconomic factors were included in the form of principal components (PC) to decorrelate these variables. All variables were standardised before Principal Component Analysis (PCA) was used.

To assess the impact of positive and negative sentiments, we used regression with ARIMA errors [11]. We opted for this hybrid approach, as it combines the strengths of both regression and time-series analysis. It allows for the inclusion of explanatory variables, such as macroeconomic factors (e.g., interest rates, GDP growth or inflation), to capture their direct impact on stock prices. Meanwhile, ARIMA errors account for the autocorrelation and time-dependent patterns often present in stock price data. The best model, with respect to ARIMA terms, was selected based on the Akaike Information Criterion (AIC).

The primary target of our analysis of central bank announcement sentiments was the S&P 500. Being a broad market index makes it a robust proxy for the overall U.S. equity market. This helps capture the aggregate market response to central bank actions, which often aim to influence macroeconomic conditions like interest rates, inflation, and growth.

While individual stocks can be heavily influenced by company-specific factors such as earnings reports, management decisions or industry trends, which may obscure the broader market's reaction to central bank signals, we decided to include the two largest U.S. banks: J.P. Morgan (JPM) and Bank of America Corporation (BAC). Analysing banks can also be insightful due to the direct and significant influence monetary policy has on the financial sector. Central bank policies, such as changes in interest rates or adjustments to reserve requirements, often have a profound impact on banks' profitability.

4 Results

4.1 Sentiment Analysis

The results of our sentiment analysis on Federal Reserve speeches using various classification models are summarised in Table 1. The performance of each model is evaluated based on precision, recall, F1 Macro score, balanced accuracy, and Area Under the Curve (AUC).

Model	Precision		Recall		F1 Macro	Balanced Accuracy	AUC
	-	+	-	+			
LR	0.493	0.869	0.630	0.792	0.691	0.711	0.841
RF	0.667	0.831	0.407	0.935	0.693	0.671	0.816
Vader	0.333	0.762	0.074	1	0.484	0.513	0.565
NB	0	0.757	0	1	0.431	0.500	0.613
LSTM	0.531	0.873	0.630	0.821	0.711	0.726	0.775
GRU	0.708	0.813	0.315	0.958	0.771	0.636	0.799
DistilRoBERTa (split)	0.500	0.816	0.388	0.875	0.641	0.632	0.766
DistilRoBERTa (trunc)	0.750	0.855	0.500	0.946	0.749	0.723	0.876

Table 1: Model performance scores, broken down by class: "negative" (-) and "positive" (+).

VADER served as our baseline model and it demonstrated reasonable precision for the "positive" class (0.762), its performance for the "negative" class was notably poor, with a precision of only 0.333 and a recall of 0.074. This indicates that VADER struggled to accurately identify negative sentiments in Federal Reserve speeches, likely due to the nuanced and domain-specific language used in such communications.

Among the shallow ML models, Random Forest (RF) exhibited the highest precision for the "negative" class (0.667) and maintained a solid precision for the "positive" class (0.831). Logistic Regression (LR) also performed well, particularly in identifying positive sentiments with a precision of 0.869 and an AUC of 0.841. However, Naive Bayes (NB) underperformed across all metrics, failing to

correctly identify any instances of the "negative" class (precision and recall of 0), which show that it performs poorly in handling the complexity of financial text.

Deep learning models, in general, performed better than shallow ML models. While the GRU achieved a better F1 Macro score than the LSTM, it had a very low recall of only 0.315 for the minority class (negative sentiments). Achieving a better precision-recall balance for both classes was more desirable, as both classes were equally important for our analysis. Hence, in this respect, the LSTM model was a better fit, which was also reflected in its highest balanced accuracy score.

When experimenting with model architectures, we concluded that the hidden layer size did not significantly impact the performance of the RNN models. Additionally, in the case of the LSTM, stacking hidden and cell states with a dense layer before the classification layer appeared more effective than dropping the cell state, though the difference was minor. Perhaps more data are needed to observe a practical difference in architecture choices. Furthermore, we did not consider data augmentations when training models, which, in theory, could provide substantial improvement.

One of the variants of transformer-based models was the top performer for the sentiment analysis. We first experimented with manual splitting of long sequences, version DistilRoBERTa (split). Sequences exceeding 512 tokens were split into smaller chunks of 512 tokens each. Each chunk was independently labelled with the sequence's overall classification label and used as input during training. Our second variant was DistilRoBERTa (trunc). Instead of splitting, sequences were truncated to include only the first 512 tokens using the `truncation=True` option in the Huggingface library. Interestingly, the truncation strategy performed significantly better during evaluation. This could have happened due to the fact that assigning the same label to all chunks assumes that every part of the sequence contributes equally to the overall label. However, in many cases, only certain portions of the sequence may be relevant for classification. This assumption can introduce noise in the training process. Other strategies are also possible. The labels for each chunk can be aggregated, and then the mode can be calculated. Alternatively, the training procedure itself can be modified: the pooler outputs of each chunk could be averaged before the classification layer.

To summarise the results, the DistilRoBERTa (trunc) model achieved the highest precision for the 'negative' class (0.750), while attaining the second-best F1 Macro score and by far the best AUC of 0.876 among all models tested. This demonstrates the effectiveness of transformer-based models in capturing deep contextual relationships and handling the complex linguistic patterns inherent in financial texts.

4.2 Stock Market Analysis and Correlation

We started by assessing the overall difference in the mean values of detrended stocks between announcements classified as "positive" or "negative" using a two-sample t-test (Table 2).

Detrended Stock	p-value
S&P 500	0.524
JPM	0.497
BAC	0.046

Table 2: Two-sample t-test results for detrended stocks.

Positive and negative sentiments are compared. Type I error: $\alpha = 0.05$.

Based on the results in Table 2, statistically significant findings were observed only for BAC. However, since regression analysis can reveal more intricate patterns, we still included the S&P 500 in the subsequent EDA steps.

Table 3 shows the loadings for each variable in the respective principal component. The first PC has strong positive correlations with the U.S. Dollar Index (DXY), GDP, Consumer Price Index (CPIAUCSL), and U.S. housing price index (CSUSHPINSA) and accounts for over 37% of the variance in the data. This PC can be interpreted as a measure of overall economic strength or macroeconomic health, capturing key indicators that collectively reflect the state of the U.S. economy.

Figure 2 depicts biplots for the first three principal components, with each observation marked as either "positive" or "negative". Over 45% of all negative sentiments are clustered in the left half of

Variable	PC1	PC2	PC3
10Y_Treasury_Yield	-0.05	0.45	0.30
CPIAUCSL	0.44	0.02	-0.12
DXY	0.39	-0.15	0.16
GDP	0.45	0.00	-0.06
UMCSENT	-0.01	-0.33	0.42
FEDFUNDS	0.15	0.34	0.39
CSUSHPINSA	0.42	0.14	0.03
Copper	0.15	0.37	-0.27
Gold	0.34	-0.05	-0.38
UNRATE	-0.27	-0.10	-0.41
Oil	-0.07	0.43	-0.26
Natural_Gas	-0.18	0.44	0.09
VIX	-0.09	-0.02	-0.27

Table 3: Loadings of the first three principal components for macroeconomic variables.

All variables were standardised first.

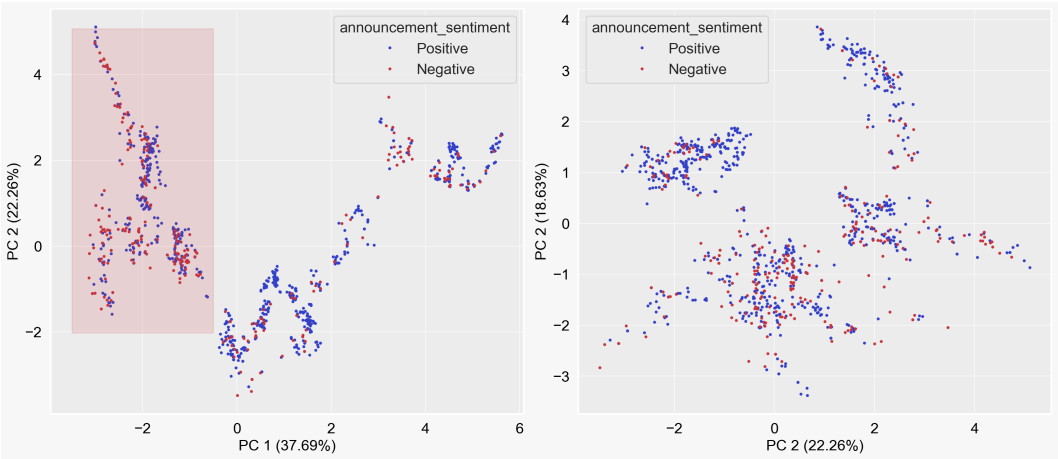


Figure 2: Biplots for the first three principal components.
The transparent area highlighted in red contains over 45% of all negative sentiments.

the first biplot, suggesting that worse economic conditions result in more messages with negative connotations being conveyed by the U.S. Central Bank.

We built regression models with ARIMA errors for the S&P 500 and BAC. To account for macroeconomic factors, we included the first four principal components, as they explained over 85% of the variance in the data. Moreover, including additional principal components did not improve model performance.

To assess the impact of U.S. Central Bank announcements, we included a categorical variable created during data labelling with the following categories: "positive" ($Sentiment(-)$), "negative" ($Sentiment(+)$), and "no announcement" (no announcement made), with the latter set as the reference level. Additionally, we incorporated two further variants of this variable: $Sentiment\ lead\ 30$ and $Sentiment\ lead\ 90$, which represent 30 and 90 days after an announcement, respectively. This was done to account for potential lag effects, as market participants may not react immediately, or if they do, there could be a delayed effect.

Results can be found in Table 4, and residual analysis is presented in Appendix, Figures 4 and 5. It can be seen that for S&P 500 none of the coefficients for the sentiment variable are statistically significant, which suggests that the sentiment surrounding U.S. Central Bank announcements does not have a strong impact on the index. For BAC, negative sentiment (variable $Sentiment(-)$) is statistically significant at $\alpha = 0.05$ with a coefficient $\beta = -0.04$. Despite this statistical significance,

Variable	β	SE	t-statistic	p-value
S&P 500: Regression with ARIMA(3,1,2)(0,0,2)[7] errors				
ARIMA terms omitted	-	-	-	-
Sentiment (-)	-0.12	0.77	-0.15	0.88
Sentiment (+)	0.17	0.49	0.34	0.74
Sentiment lead 30 (-)	-0.19	0.77	-0.25	0.80
Sentiment lead 30 (+)	-0.18	0.50	-0.36	0.72
Sentiment lead 90 (-)	0.49	0.77	0.64	0.52
Sentiment lead 90 (+)	-0.45	0.50	-0.91	0.36
PC1	-44.33	7.27	-6.09	0.00
PC2	5.08	3.74	1.36	0.18
PC3	-37.91	3.52	-10.77	0.00
PC4	106.74	2.01	53.00	0.00
BAC: Regression with ARIMA(2,1,2)(1,0,2)[7] errors				
ARIMA terms omitted	-	-	-	-
Sentiment (-)	-0.04	0.02	-2.71	0.01
Sentiment (+)	0.01	0.01	0.76	0.44
Sentiment lead 30 (-)	0.00	0.02	0.29	0.77
Sentiment lead 30 (+)	0.00	0.01	0.26	0.79
Sentiment lead 90 (-)	0.00	0.02	-0.18	0.86
Sentiment lead 90 (+)	0.00	0.01	0.05	0.96
PC1	-1.22	0.15	-7.96	0.00
PC2	0.49	0.08	6.23	0.00
PC3	-1.11	0.07	-14.92	0.00
PC4	1.20	0.04	28.51	0.00

Table 4: Regression results for models with ARIMA errors.

All ARIMA terms were omitted for simplicity. $Sentiment(-)$ denotes announcements labeled as negative, and $Sentiment(+)$ denotes those labeled as positive. $Sentiment\ lead\ x$ represents announcements made x days earlier.

the coefficient is very small, suggesting it has a minimal impact on the model, particularly when compared to the combined effects of macroeconomic factors, which are represented as principal components.

We identified several reasons why practically no effect was found:

1. **Lack of Surprises:** If the Central Bank's announcements are expected or align with market forecasts, the actual announcement may not introduce any new information. In such cases, the market may not react strongly because the announcement does not deviate from expectations.

2. **Announcement Scope:** Since we included all announcements made by the U.S. Central Bank without categorising them into relevant groups, the overall effect may be diluted. For instance, some announcements may not pertain to the state of the economy and/or could be very targeted, such as those addressing issues specific to a particular state.

3. **Announcement Specificity:** Using broad categories such as "positive" or "negative" may not allow for a sufficiently nuanced analysis of impact.

4. **Data Source:** Supplementing the original data with information from other sources, such as testimonies, news articles, and economic overviews, can help explain more of the variance in the data.

5. **Preemptive Mitigation:** In anticipation of impacting the market, the Central Bank may release statements that are inherently neutral or leaning toward the positive side. This is also reflected in the distribution of sentiments, with speeches classified as "positive" being the majority category.

5 Limitations

When conducting this study, we identified several key limitations that could be addressed in the future. First, using binary categories for classifying central bank communications may be overly simplistic, potentially resulting in a loss of nuanced information. Additionally, the study could benefit from incorporating other forms of communication and grouping them into subcategories based on their relevance to macroeconomic conditions, e.g., inflation, employment, or monetary policy. Addressing these limitations could provide a more comprehensive understanding of the nuanced messages conveyed by central banks.

6 Conclusion

We believe this study demonstrates that advanced deep learning models can be effectively used to classify central bank communications, allowing for more in-depth analysis of how their sentiments might influence market participants, including the stock market. While no significant impact was found in our analysis, it is possible that results may change once the limitations of our study are addressed, which highlights the need for future work in this area.

References

[1] Bernd Hayo and Matthias Neuenkirch. Do federal reserve communications help predict federal funds target rate decisions? *Journal of Macroeconomics*, 32(4):1014–1024, 2010.

[2] David-Jan Jansen and Jakob De Haan. Talking heads: the effects of ecb statements on the euro–dollar exchange rate. *Journal of International Money and Finance*, 24(2):343–361, 2005.

[3] Seongik Park and Yanggon Kim. Building thesaurus lexicon using dictionary-based approach for sentiment classification. In *2016 IEEE 14th International Conference on Software Engineering Research, Management and Applications (SERA)*, pages 39–44, 2016.

[4] C. Hutto and Eric Gilbert. Vader: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225, May 2014.

[5] Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, and Lifang He. A survey on text classification: From shallow to deep learning, 2021.

[6] Wajdi Aljedaani, Furqan Rustam, Mohamed Wiem Mkaouer, Abdullatif Ghallab, Vaibhav Rupapara, Patrick Bernard Washington, Ernesto Lee, and Imran Ashraf. Sentiment analysis on twitter data integrating textblob and deep learning models: The case of us airline industry. *Knowledge-Based Systems*, 255:109780, 2022.

[7] Jeow Li Huan, Arif Ahmed Sekh, Chai Quek, and Dilip K. Prasad. Emotionally charged text classification with deep learning and sentiment semantic. *Neural Computing and Applications*, 34(3):2341–2351, Feb 2022.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global vectors for word representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics.

[10] Ravid Shwartz-Ziv, Micah Goldblum, Yucen Lily Li, C. Bayan Bruss, and Andrew Gordon Wilson. Simplifying neural network training under class imbalance. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

[11] George E. P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, and Greta M. Ljung. *Time Series Analysis: Forecasting and Control*. Wiley, Hoboken, NJ, 5th edition, 2015. See Chapter 9.5, Regression Models with Time Series Error Terms, p. 339.

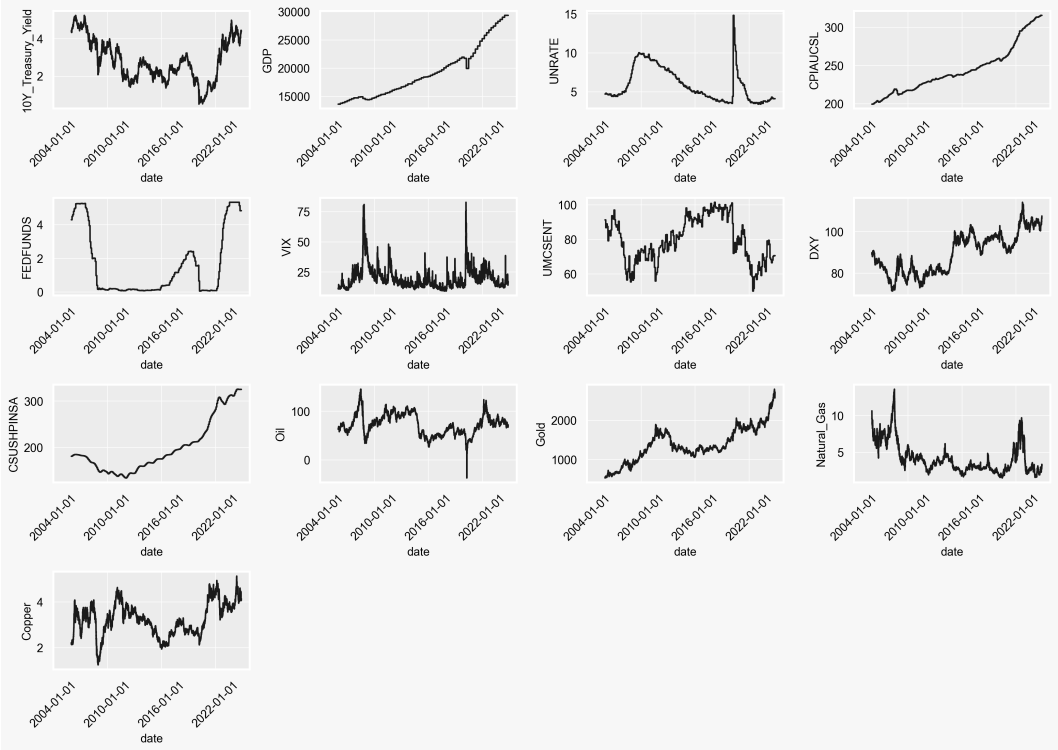


Figure 3: Macroeconomic factors included in regression models

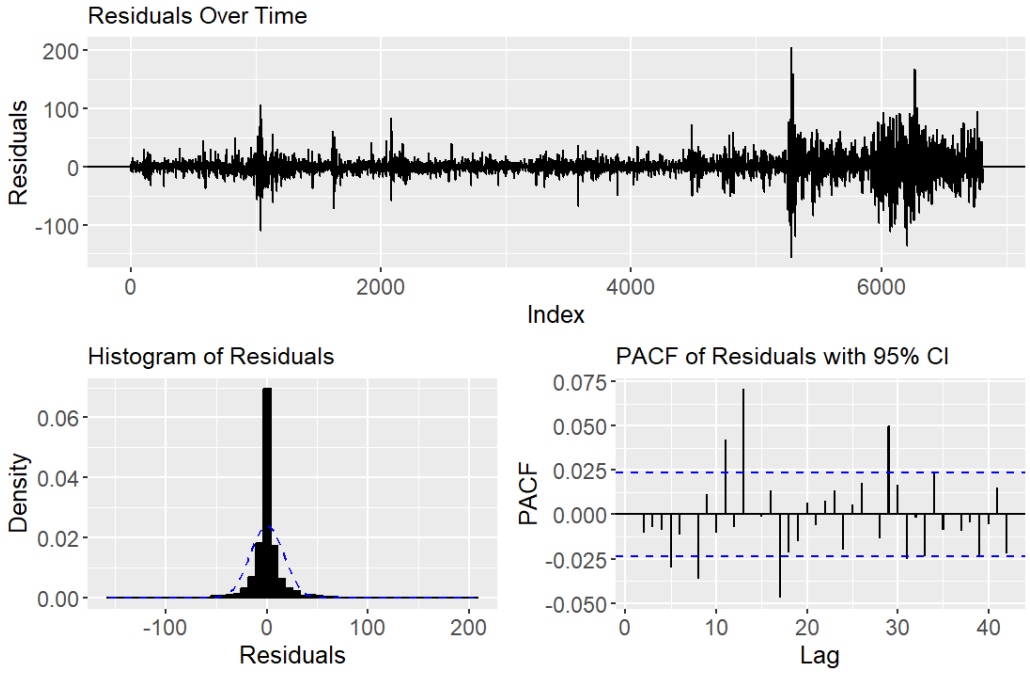


Figure 4: Residual analysis for S&P 500 model.

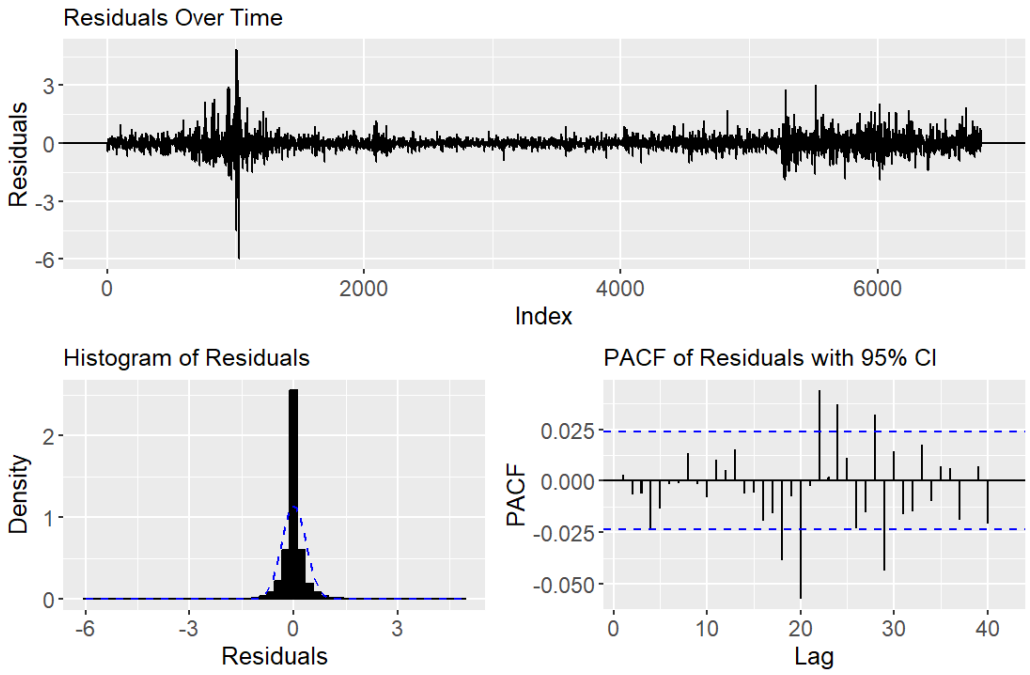


Figure 5: Residual analysis for BAC model.