Sentiment analysis of FOMC meeting minutes

A neural network approach

Ollie Sun *　　Stella Zhu †

May 1, 2023

Abstract

In this paper, we explore the application of Long Short-Term Memory (LSTM) networks for sentiment analysis of documents published by the Federal Open Market Committee (FOMC). FOMC meeting minutes play a crucial role in influencing economic assessments and beliefs about future policy, impacting the global economy and financial markets. LSTM networks, with their ability to capture long-range dependencies in sequential data, are well-suited for this task. We demonstrate the efficacy of LSTM networks in sentiment analysis of FOMC meeting minutes. Furthermore, we attempt to extend the model to predict financial market movements based on the sentiment analysis results.

1 Introduction

Over the past few decades, neural networks have experienced significant advancements, becoming increasingly powerful tools for a wide range of applications, from computer vision to natural language processing (LeCun et al., 2015; Goodfellow et al., 2016). These deep learning models, inspired by the structure and function of biological neural networks, have consistently outperformed traditional machine learning techniques,

*Email: olliesun@bu.edu; Questrom School of Business, Boston University
†Email: jyszhu@bu.edu; Questrom School of Business, Boston University

establishing new benchmarks and transforming the landscape of artificial intelligence research (Schmidhuber, 2015; Hinton et al., 2012).

In this paper, we present the application of a specific kind of neural network Long Short-Term Memory (LSTM) network in pursuit of conquering sentiment analysis tasks in the finance domain. LSTM is well-suited for sentiment analysis tasks due to its ability to capture and model long-range dependencies in sequential data, such as text (Hochreiter and Schmidhuber, 1997). Sentiment analysis often requires understanding the context and semantics of phrases, sentences, and even entire paragraphs to accurately infer the underlying sentiment, so we use LSTM networks which utilize memory cells and gating mechanisms to selectively store, update, and retrieve information across long sequences, to better capture the contextual relationships between words and phrases within a text (Hochreiter and Schmidhuber, 1997; Gers et al., 2000).

The target of analysis is the documents published by the Federal Open Market Committee (FOMC), which is a policy-making body within the Federal Reserve System responsible for making key decisions related to monetary policy that not only influence the United States but also has a significant impact on the global economy and financial markets. These documents can influence the public's economic assessment and beliefs about future policy. The most important documents are brief statements released shortly after meetings and meeting minutes released three weeks later, which contain more information and nuances. Examining the minutes offers a more independent analysis compared to policy statements as they do not overlap with announcements about monetary policy changes.

The paper is organized as follows: In Section 2, we first introduce the role and importance of FOMC meeting minutes and define sentiment analysis as the task we want to accomplish. Then we describe the methodology and architecture of our deep learning model using LSTM in Section 3. We also provide a detailed description of the dataset used in our study and the implementation of our model in Section 4. Finally, we present our test results of sentiment analysis and extend the model to predict the financial market movement based on the sentiment analysis in Section 5.

2 FOMC, Meeting Minutes and Previous Research

In this section, we provide an overview of the function of FOMC and the meeting minutes. Here we visit the previous research utilizing the FOMC minutes and sentiment analysis. Through these existing studies, we can observe the impact of FOMC minutes on the capital market and economy. Additionally, we also display the past usage of sentiment analysis toward stock price prediction. By referencing previous studies, we show how our project extends current works and contributes to filling research gaps.

2.1 The Role of FOMC and its Meeting Minutes

The Federal Open Market Committee (FOMC), branched out from the Federal Reserve System and is the committee that oversees national open market operations and conducts monetary policy. As a key part of the Federal Reserve System, FOMC makes the decision about the nation's money supply growth rate, which influences the availability of credits and the level of interest rates. These changes would further trigger a series of events that make short and long-term economic impacts. As such, the activities and wording of the FOMC are closely monitored by financial institutions, policymakers, and general investors. The FOMC committee consists of twelve members and has eight regularly scheduled meetings each year. Additional meetings are scheduled when the economic conditions require more discussion and attention. At each of these meetings, members of the committee review the financial and economic conditions to assess the long-term price stability risks and determine the proper federal funds rate. The structure of a typical meeting involves reviewing the recent developments in the financial and foreign exchange markets, providing an outlook for the domestic and global economy, and making proposals and voting for policy decisions based on this information.

The FOMC publicly announces the meeting decision at around 2 p.m. on the last day of the meeting and holds a press conference immediately after. Approximately three weeks after the meeting, the committee releases the meeting minutes. The minutes provide a more complete explanation of the views and decisions and are written for better general understanding and social awareness. Prospects of the meeting committee and their sentiment can be observed through these carefully constructed minutes which can influence the public's overall economic assessment and beliefs.

2.2 FOMC Minutes Impact on Markets and Economy

Research based on the FOMC meetings and statements has branched out into various directions. Much research has been dived into investigating the impact of FOMC statements on asset prices and the economy. However, there is still much to explore about how the FOMC meeting minutes influence the capital market. Some existing literature has proved the uniqueness and importance of FOMC minutes and their impact on asset prices and economic outlook. As explained by Rosa (2013), the FOMC minutes differ from FOMC statements and press conference records since it contains more detailed information and nuances of the policy meeting. Rosa (2013) shows that the release of FOMC minutes significantly increases market volatility and trading volume. Nechio and Wilson (2016) showed through semantic-based measures that the FOMC minutes release have sizable impacts on Treasury bond yields, and the effect is more prominent when there exists a difference in sentiment between the statement and the minute of the same meeting. Jubinski and Tomljanovich (2013) used an intraday analysis and find that the release of FOMC minutes has a significant impact on individual equity volatility but the effect on individual equity returns are less notable. They also documented that the magnitude of the effect of FOMC minutes is positively correlated with firm size. Apergis (2015) further extends the work of Rosa (2013) and found that the release of FOMC minutes has the tendency to significantly impact the mean and volatility of high-frequency asset returns, but this effect would be diminished after a significant crisis. This paper also points out that the FOMC minutes allowed market participants to update their expectations of future asset price paths and beliefs about the economic environment.

2.3 Sentiment Analysis and Previous Methods

Sentiment Analysis is a common method adopted to analyze and predict stock market movements in literature. Much current research in natural language processing (NLP), computational linguistics, and text mining is centered around sentiment analysis. In the field of financial studies, researchers use sentiment analysis to understand investor decisions and activity given textual information. The use of sentiment analysis builds the bridge between behavioral studies and quantitative financial analysis. Incorporating AI and machine learning algorithms, institutional and individual investors can "scrap" information from websites and social media to get market opinion allowing predictions of stocks and economic environments.

There is abundant literature that utilizes sentiment analysis on social media data. In particular, given the time efficiency consideration, Twitter data is one of the most common datasets this method is applied. Pagolu et al. (2016) used Word2vec and N-gram to analyze the public sentiment from Twitter and demonstrate the strong correlation between sentiment observed and individual stock movements. Khedr et al. (2017) constructed an effective model to predict future stock price trends based on sentiment analysis of historical prices and financial news. They used the naïve Bayes algorithm to perform news sentiment analysis and used the KNN algorithm to predict future stock price behavior.

Previous studies also attempt to use sentiment analysis and text mining techniques on FOMC datasets. Cannon (2015) used the FOMC transcripts data and employed the financial dictionary of Loughran-McDonald and the bag-of-words count method. He concludes that the FOMC meeting transcripts provide abundant sentiment information and are strongly related to economic activities. More recently Huang and Kuan (2021) used an adaptive Bayesian approach to build sentiment indicators on the FOMC minutes data. They improved the existing mining techniques and showed that FOMC minutes are valuable in predicting economic variables and generating superior out-of-sample forecasts. Ke et al. (2022) attempted to create a robust Bayesian analysis model based on the current Latent Dirichlet Allocation model, tested the model on the FOMC transcript data, and reported the "conformity" of committee members. Handlan (2020) analyzed the impact of FOMC statement announcements on fed funds futures prices using neural network methods. He found that the magnitude of changes in forward guidance on real interest rates doubled when using the text shocks. The adoption of neural networks allowed for incorporating complex features of the text.

Beyond the methods mentioned above, there are also attempts to standardize this procedure by assigning sentiment scores to the given text documents. Lucca and Trebbi (2009) constructed a linguistic measure

“Factiva semantic orientation score” (FSO) using discussions of FOMC statements from Dow Jones Factiva news database. They proposed to assign automated measurement of sentiment to individual sentences. With their novel approach, they concluded that short-term yields mainly respond to changes in policy rates, while longer-term treasury yields mainly respond according to FOMC statements. Given the certainty that investor sentiment affects the capital market, there are firms that specialize in providing a semantic assessment of textual materials, and these ready-to-use assessments are also used in research relevant to FOMC communications impact. For example, Nechio and Wilson (2016) used the sentiment scores provided by Prattle Analytics LLC and found that under this measure the hawkish (dovish) minutes would lead to hawkish (dovish) market reactions. Nechio and Wilson also attempted to use the previously mentioned FSO score which shows that the tone of minutes and market reactions positively correlates.

As shown in Figure 1, the methods of sentiment analysis have continued to branch out with new approaches.

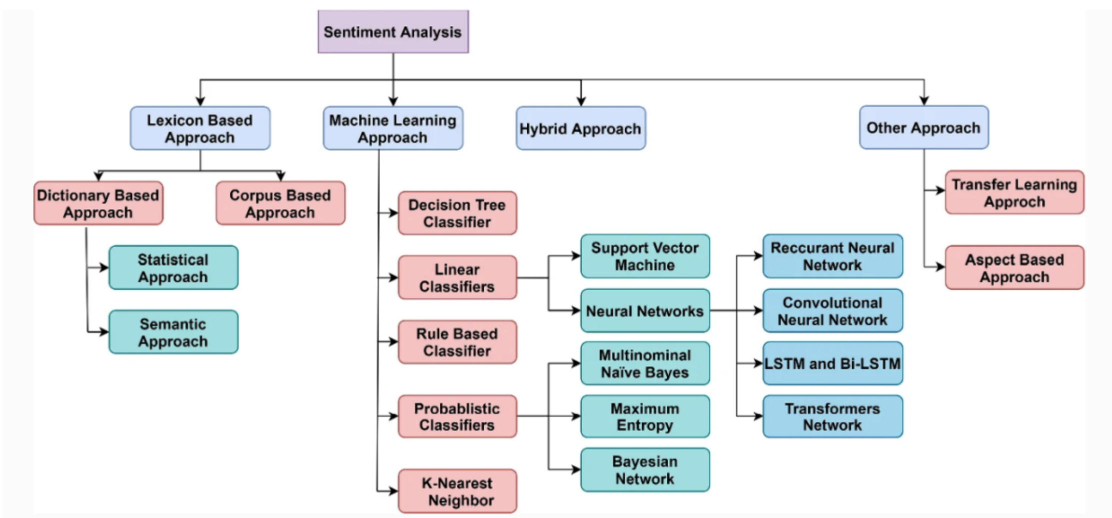

Figure 1: Sentiment analysis approaches. From Yilmaz (2022)

3 Neural Network Approach

In this section, we describe a novel methodology for sentiment analysis and market predictions, namely the Long Short-Term Memory (LSTM) network. LSTM is a type of recurrent neural network (RNN), that has proven to be highly effective in handling sequential data and capturing long-range dependencies, making them well-suited for sentiment analysis tasks. We first introduce the architecture of recurrent neural networks.

3.1 Recurrent Neural Network (RNN)

Recurrent Neural Networks (RNNs) are a class of neural networks designed to handle sequential data. Unlike feedforward neural networks, which process data in a single pass, RNNs maintain a hidden state that summarizes information from previous time steps. The hidden state is updated at each time step using the input and the previous hidden state, allowing the network to capture temporal dependencies in the data.

The underlying structure of RNNs can be represented using a directed acyclic graph (DAG), where each node represents a time step and the edges represent the connections between the hidden states. An example of RNN is shown in Figure 2. At each time step, the input is passed through a linear layer, which is followed by a non-linear activation function such as the hyperbolic tangent (tanh) or the rectified linear unit (ReLU). The output of this activation function is then combined with the previous hidden state using matrix multiplication and another activation function to produce the new hidden state:

$$h_t = \tanh(W_{hh}h_{t-1} + W_{xh}x_t + b_h)$$

where x_t is the input at time step t, h_t is the hidden state at time step t, W_{hh} is the weight matrix for the hidden state, W_{xh} is the weight matrix for the input, and b_h is the bias vector. The tanh function is used as the activation function to introduce non-linearity to the model.

The key advantage of RNNs is their ability to handle variable-length sequences and capture long-term dependencies in the data. However, they suffer from the problem of vanishing gradients, which makes it difficult to propagate errors back through long sequences. To address this issue, several variants of RNNs have been proposed, such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU), which use more sophisticated gating mechanisms to control the flow of information through the network.

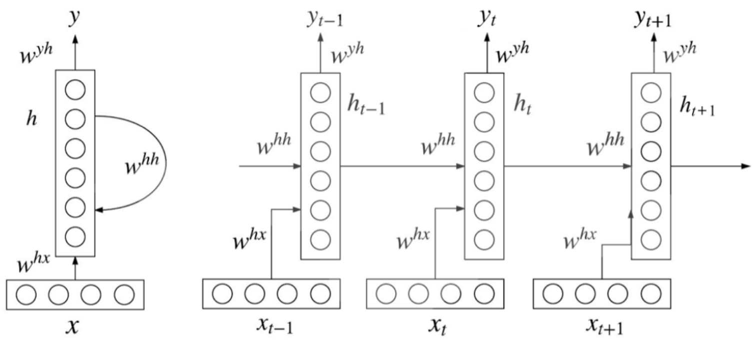

Figure 2: Example of RNN. From Zhang et al. (2018)

3.2 Long Short-Term Memory Network (LSTM)

The Long Short-Term Memory Network (LSTM) is a type of Recurrent Neural Network (RNN) that is capable of learning long-term dependencies and solving the problem of vanishing and exploding gradients, which significantly hindered the learning capabilities of traditional RNNs in long sequences (Hochreiter and Schmidhuber, 1997). The vanishing gradient problem arises when gradients of the loss function with respect to the model parameters become too small as they are propagated back through time, leading to slow or ineffective learning for early layers of the network (Bengio et al., 1994). Conversely, exploding gradients occur when these gradients become too large, causing unstable learning and difficulty in convergence (Pascanu et al., 2013). To address these problems, LSTMs have a chain-like structure composed of repeating modules like all RNNs, but the repeating module of an LSTM is more complex than that of a standard RNN. Instead of a single neural network layer, there are four layers that interact in a special way. Additionally, LSTMs have two states: a hidden state and a cell state. An example of an LSTM structure is shown in Figure 3.

LSTMs have been shown to achieve state-of-the-art performance in various sentiment analysis tasks, outperforming traditional RNNs and other shallow learning methods. For instance, in a study by Tang et al. (2016), LSTM networks were used to analyze sentiment in a hierarchical fashion by first examining sentiment at the word level, followed by the sentence level, and finally aggregating the information to infer document-level sentiment. This approach demonstrated the efficacy of LSTMs in capturing both local and global semantic information, leading to improved sentiment analysis performance. Similarly, Wang et al. (2016) employed LSTMs in combination with attention mechanisms to selectively focus on relevant parts of the input text, further enhancing the performance of sentiment classification tasks. These studies highlight the suitability of LSTM networks for sentiment analysis tasks, as they effectively capture the contextual information necessary for accurate sentiment inference.

The four layers of an LSTM module are the input gate, forget gate, output gate, and the cell state. Each gate has a sigmoid activation function that decides which information to retain and which to discard, while the cell state acts as a memory that stores and updates information over time (Greff et al., 2017). The input gate determines how much new information should be stored in the cell state, the forget gate decides how much information to discard from the cell state, and the output gate determines which information should be outputted based on the current input and modified cell state. By carefully controlling the flow of information through these gates, LSTMs can effectively capture long-term dependencies in sequential data. We discuss the details of the components of the LSTM network below.

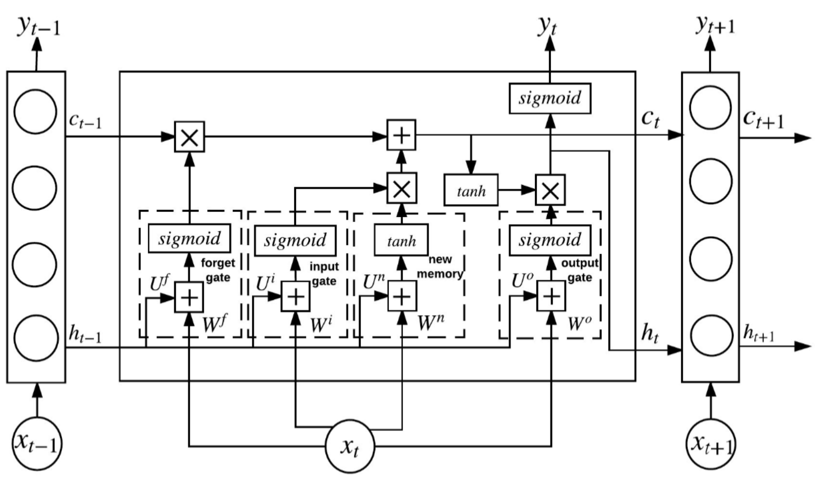

Figure 3: Example of LSTM. From Zhang et al. (2018)

3.2.1 Gates and Cell State Update Formulas

Given an input vector x_t at time step t and the previous hidden state h_{t-1}, the LSTM cell computes the new hidden state h_t and updates the cell state C_t. This process involves the following operations:

1. Input gate (i_t):

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i)$$

2. Forget gate (f_t):

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f)$$

3. Output gate (o_t):

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o)$$

4. Candidate cell state (C'_t):

$$C'_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c)$$

5. Updated cell state (C_t):

$$C_t = f_t \odot C_{t-1} + i_t \odot C'_t$$

6. Updated hidden state (h_t):

$$h_t = o_t \odot \tanh(C_t)$$

Here, σ denotes the sigmoid activation function, $\odot$ represents element-wise multiplication, and W and b are the learnable weight matrices and bias vectors, respectively.

3.2.2 Weight Matrices and Bias Vectors

The weight matrices (W_i, W_f, W_o, W_c) and bias vectors (b_i, b_f, b_o, b_c) are learnable parameters that are updated during the training process using backpropagation through time (BPTT). Each weight matrix has dimensions (n, m), where n is the number of hidden units in the LSTM cell and m is the sum of the input vector size and the hidden state size (i.e., $m = \text{input\_size} + \text{hidden\_size}$).

3.2.3 Activation Functions

The sigmoid activation function (σ) is used for the input, forget, and output gates to generate values between 0 and 1, effectively controlling the flow of information through the gates. The hyperbolic tangent (tanh) activation function is used to produce the candidate cell state and the final output, generating values between -1 and 1, which allows the LSTM cell to capture a wider range of information compared to using the sigmoid function alone.

3.2.4 Adjustment for Classification

In addition to the LSTM configuration outlined, we have incorporated an additional layer of embedding prior to inputting data into the LSTM. This is to ensure the effective representation of our textual input with efficient dimensionality, thereby optimizing the performance of the network.

Furthermore, as we are executing a classification task, we have concluded the network with a fully connected linear layer that produces a scalar output and we use the logit loss function to train the model. Let y be the true label (-1 or 1) of a sample x, and $\hat{y} = \hat{f}(x)$ be the output generated through the LSTM networks described before. The logistic loss (or binary cross-entropy loss) can be calculated as follows:

$$L(y, p) = -[y \log(p) + (1 - y) \log(1 - p)]$$

where the scalar p, the predicted probability of an input belonging to class 1 is given by the fully connected linear layer, for learnable parameter W_p, b_p:

$$p = W_p \hat{y} + b_p$$

Further elaboration on these specifications will be presented in the following section. For a comprehensive overview of the model architecture, refer to Table 3.

4 Data and Model

To conduct the sentiment analysis of the LSTM model described in the preceding section, we first gather the FOMC meeting minutes as text documents from the official FOMC website. Then we construct the labeled dataset by the traditional dictionary method as described in pre-process of the input data, including tokenization, stop-words removal, creating a vocabulary, converting text documents into numerical vectors, splitting the dataset into training and testing sets, creating an LSTM model, and finally, training and evaluating the model.

In order to implement the LSTM neural network as described in Section 3, the `PyTorch` package, which has gained wide popularity in the Python ecosystem, was employed. `PyTorch` is an open-source machine learning framework developed by Facebook's AI research team that provides a seamless path from research prototyping to production deployment. It is known for its dynamic computational graph construction, allowing for flexible and efficient neural network modeling.

4.1 Dataset Construction

The original FOMC meeting minutes are readily available in HTML text format on the official website of BOARD OF GOVERNORS of the FEDERAL RESERVE SYSTEM (2023), as illustrated in Figure 4. We obtained data for the period from January 2012 to December 2022, consisting of 87 meeting minutes and a total of 784,957 words. On average, each document contains approximately 9,022 words. To gain insight into the most common words used throughout the documents, we generated a word cloud shown in Figure 5. This visualization provides a quick overview of the most frequently used terms, with larger words indicating greater frequency of use.

To construct a labeled dataset for our deep learning model to work with, we then employed dictionary-based approaches to label each sentence based on the sentiment expressed. Specifically, we followed the methodology outlined in Tadle (2022) and adopted the Dictionary Method of Content Analysis to analyze information on selected topics such as unemployment, production, inflation, and other economic indicators in FOMC

Figure 4: Meeting minute example from FOMC website

documents and assessed the overall outlook using pre-specified sentiment categories of keywords: hawkish and dovish. The use of hawkish keywords indicated an improving economic outlook and higher inflationary pressures, signaling potential monetary policy tightening, while dovish keywords emphasized deteriorating economic conditions and subdued price changes, suggesting possible monetary policy loosening. Both sets of words are included in Appendix A.

To determine the precise direction of sentiment when invoking keywords, we incorporated lists of positive and negative polarized terms that appeared within the FOMC documents, as shown in Appendix A. Following Tadle (2022), we calculated the sentiment score of each sentence by considering the number of hawkish keywords denoted by h, the number of dovish keywords denoted by d, the number of positive terms denoted by p, and the number of negative terms denoted by n. The sentiment score was then computed using the following metric:

$$\text{score} = \begin{cases} 1, & \text{if } h > d \,\&\, p > n \text{ or } h < d \,\&\, p < n \\ -1, & \text{if } h > d \,\&\, p < n \text{ or } h < d \,\&\, p > n \\ 0, & \text{otherwise} \end{cases}$$

Following the sentiment-score calculation based on this dictionary method, the resulting dataset included a set of 7,815 sentences. To ensure the dataset only contained clear expressions of sentiment, sentences that

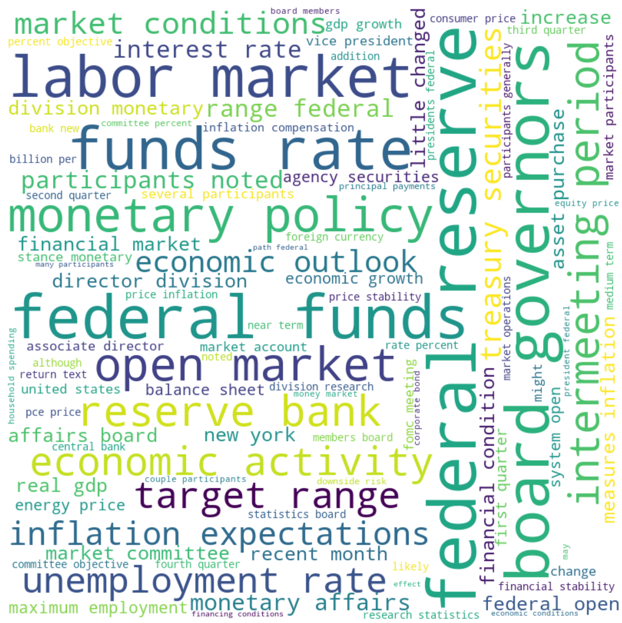

Figure 5: Word cloud from 2012-2022 samples

received a score of 0 were omitted. The remaining dataset consists of 4,640 sentences labeled as Hawkish and 3,175 sentences labeled as Dovish. Table 1 displays a selection of labeled sentences from the dataset. To examine the relationship between the sentiment scores and the Federal Reserve effective fund rate which the document directly impacts, we plot the movement of the sentiment scores of each meeting minute against the effective fund rate in Figure 6. Here the score of each minute document is the average of the scores of all sentences contained in the meeting minute. This visualization allows us to explore potential correlations between the sentiment expressed in the meeting minutes and changes in the interest rate.

4.2 Data Processing

Before training our LSTM model, we need to apply a series of operations to ensure the input data can be correctly recognized by the neural networks.

Text	**Label**
Sales of both new and existing homes advanced in January, and home prices increased further	Hawkish (1)
Many participants regarded the international situation as an important source of downside risks to domestic real activity and employment, particularly if declines in oil prices and the persistence of weak economic growth abroad had a substantial negative effect on global financial markets or if foreign policy responses were insufficient	Dovish (-1)
A few participants, however, mentioned the possibility that economic growth could be more rapid than currently anticipated, particularly if major sources of uncertainty were resolved favorably or if faster-than-expected advances in the housing sector led to improvements in household balance sheets, increased confidence, and easier credit conditions	Hawkish (1)

Table 1: Examples of labeled sentences

4.2.1 Tokenization and Pre-processing

In the first step of our preparation, we tokenize each sentence by splitting it into individual words, followed by cleansing the text data using standard techniques such as lowercasing and removing stopwords. To accomplish this task, we utilize the widely-used Python package, `NLTK`, which provides an array of tools and resources for natural language processing tasks, including tokenization, stemming, tagging, and stopword removal.

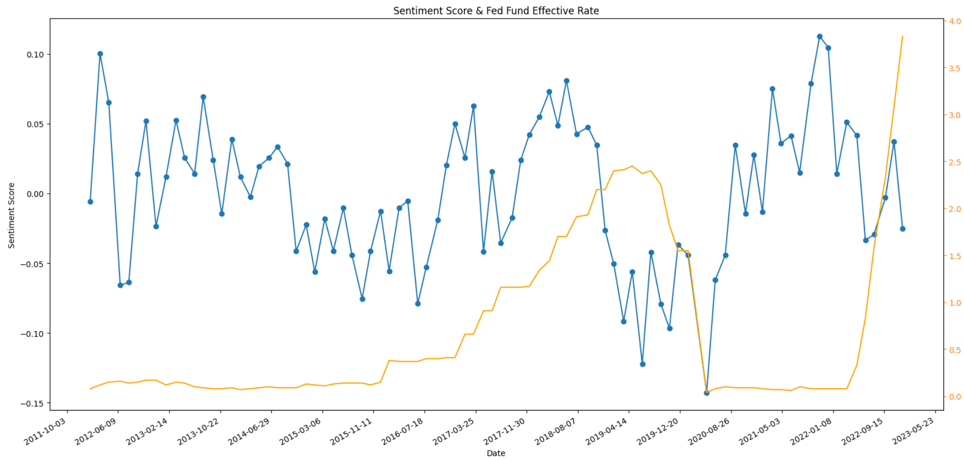

Figure 6: Movement of calculated document sentiment scores

4.2.2 Vocabulary Creation and Embedding

Once the text data has been cleaned and pre-processed, the next step is to create an effective vocabulary, which is a critical step in all NLP tasks. To transform textual data into machine-readable inputs, each word or token is typically represented as a numerical vector, with a unique index associated with each word in the vector using a one-hot encoding scheme. This allows standard neural network techniques to be applied to the data. However, this approach can lead to computational inefficiencies due to the high dimensionality of the resulting representation. For instance, a vocabulary of 10,000 unique words would result in a one-hot encoded representation where each word is represented as a 10,000-dimensional binary vector with a single '1' at the position corresponding to the word and '0's elsewhere, leading to a sparse and high-dimensional representation.

To mitigate the issue of high dimensionality associated with one-hot encoding, we employ word embeddings, a common technique in NLP tasks. Word embeddings utilize dense vector representations to map words from a large, sparse one-hot encoded representation to a continuous, dense vector space with fewer dimensions. In this study, we use the GloVe (Global Vectors for Word Representation) embedding algorithm proposed by Pennington et al. (2014). GloVe embeddings are generated through a co-occurrence matrix and a weighted least-squares model, which enables efficient and effective training. GloVe embeddings have

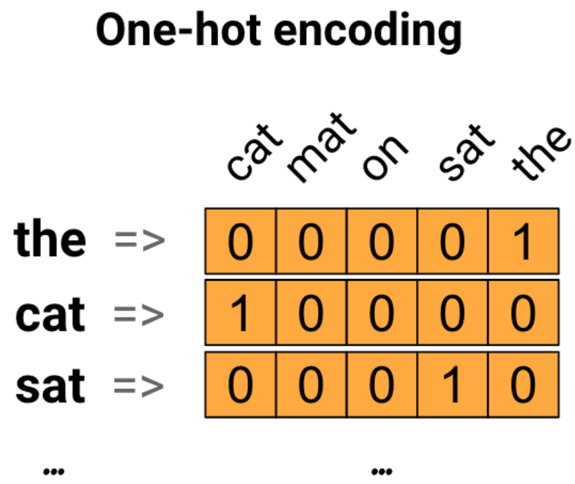

Figure 7: One-hot encoding of vocabulary

emerged as a popular choice for natural language processing tasks due to their ability to capture both local and global context information, resulting in improved performance in tasks such as semantic similarity, analogy detection, and sentiment analysis. More information on GloVe embeddings can be found at https://nlp.stanford.edu/projects/glove/.

4.3 Training and Evaluation

After converting the text documents into dense numerical vectors using the tokenized words and vocabulary, we randomly split the dataset into training (80%) and validation (20%) sets. This partition allows us to train our model on a large portion of the data while reserving a subset for evaluating its performance. We design an LSTM-based classifier with 2 layers and a hidden size of 128. The model comprises an embedding layer, followed by the LSTM layers, and a fully connected layer for producing the final output. We incorporate dropout layers (with a rate of 0.5) before and after the LSTM layers to regularize the model and prevent overfitting. We also choose 64 as the batch size for training. We train the model using the Adam optimizer with a learning rate of 0.001 and a binary cross-entropy with a logit loss function. The hyper-parameters and functions are explained in Table 2.

Hyper-parameter	Value	Description
vocab_size	4091	Size of the vocabulary created from the dataset. It defines the number of unique tokens (words) in the input text data.
embedding_dim	100	Size of the word embeddings or the continuous vector representations for each word.
hidden_dim	128	Size of the hidden state in the LSTM layers.
output_dim	1	Size of outputs which is 1 in the classification task.
n_layers	2	Number of LSTM layers stacked together in the model.
dropout	0.5	Regularization technique used to prevent overfitting. It determines the probability of dropping out (zeroing) the activations of neurons during training.
batch_size	64	Number of training samples used in one iteration (i.e., one update of the model weights) during the training process.

Table 2: Hyper-parameters of the model

We train the model for 10 epochs, monitoring both training and validation losses and accuracies. After training, we evaluate the model on the test set to assess its performance in the classification of sentence sentiment. A summary of the model is shown in Table 3, which has a total number of 791,181 trainable parameters.

5 Results

5.1 Sentiment Analysis on Two Types of FOMC Documents

The final accuracy of our model on the validation set, after 10 epochs, reached approximately 90%. To assess the model's performance, we utilized another set of FOMC document sources, including the FOMC statements. While the meeting minutes provide a more detailed and nuanced analysis of the discussions and

Layer	Output Shape	# of Params
Embedding	[64, N, 100]	409,100
Dropout	[64, N, 100]	–
LSTM	[64, N, 128]	381,952
Dropout	[64, 128]	–
Linear	[64, 1]	129

Table 3: Summary of the model (N is the number of tokens in one sample)

debates that occurred during the meeting, including the individual viewpoints and concerns of the members, the FOMC statement offers a more concise summary of the Committee's decisions and perspectives. It is important to note that the release of the meeting minutes occurs three weeks after the meeting, whereas the FOMC statement is published immediately following the meeting.

To evaluate the model's performance on a broader range of FOMC documents, we utilized a dataset of scraped data from Kaggle (DrLexus and Ozannen, 2023)), consisting of all sentences in FOMC documents and statements from the period 2015 to 2023. The labeling procedure followed the same methodology outlined in Section 4.1, and the results are presented in Table 4.

Sample	Original	FOMC Statements	FOMC Minutes
Period	2012 - 2022	2015 - 2023	2015 - 2023
# of samples (sentences)	7815	412	5963
# of Positive (Hawkish) samples	4640	255	3463
# of Negative (Dovish) samples	3175	157	2500
Train accuracy	92.56%		
Validation accuracy	91.87%	96.39%	94.71%

Table 4: Results of sentiment classification

5.2 Predication on Financial Market

While our model exhibits remarkable precision in binary sentiment classification, it is natural to explore its potential to extend predictability to the complexities found in real-world applications. In particular, we aim to examine whether our model can capitalize on the advantages of deep learning architecture to effectively forecast market movements (LeCun et al., 2015). To this end, we assess our model's performance against two key indicators: (1) the directional change in the single-day return of the S&P 500 index, and (2) the implied volatility (VIX) associated with the Federal Open Market Committee (FOMC) meeting minutes release dates. The S&P 500 index represents a stock market index gauging the performance of 500 prominent companies listed on American stock exchanges, while the VIX, or CBOE Volatility Index, measures the market's expected volatility over the ensuing 30 days based on the prices of options on the S&P 500 index.

The S&P 500 index is a widely acknowledged and closely monitored indicator of the overall performance of the US stock market, with directional changes bearing significant economic and financial implications (Campbell et al., 1997). Conversely, the VIX, often referred to as the "fear index," enables investors to evaluate the extent of market risk and uncertainty (Whaley, 2000). Consequently, both indicators are crucial in determining the health and stability of the US stock market and are scrutinized by investors, analysts, and policymakers alike. Thus, if our model can accurately predict the movements of the S&P 500 index and the VIX, it holds the potential to offer valuable insights and guide investment decisions for individuals, financial institutions, and policymakers. The data we used are obtained from the CRSP financial database. We employ the same experimental setup as in sentiment analysis, with modifications to hyperparameters to account for the reduced number of samples, as each sample constitutes an entire document. We set the dimension of the hidden layer to 64 and the dropout rate to 0. Each sampled document is labeled using the following metric to reflect the direction of movement of indicators:

$$\text{label}(\text{document}_d) = \text{sgn}(S^d_{\text{close}} - S^d_{\text{open}})$$

Here, d represents the release date of the document, S represents the asset price of the two indicators on date d, and `open/close` represent open/close prices, respectively.

Table 5 presents the accuracy of this new experiment. The performance is relatively unsatisfactory, with an accuracy of approximately 50%. We may hypothesize that asset prices follow a random walk process, independent of any effects the FOMC meeting document release might bring onto the financial market. The Random Walk theory is a prominent concept in financial economics proposed by Fama (1965) which posits that asset prices incorporate all accessible information, rendering changes in prices entirely random and unpredictable. Consequently, the Bayes error rate can be no higher than 50% if the efficient market hypothesis is true.

To explore the potential relationship between the sentiments of FOMC meeting minutes and financial markets, we conduct a simple Ordinary Least Squares (OLS) linear regression analysis of the daily returns of the S&P500 index (`vwretd`) on the calculated sentiment scores (`sentiment_score`). The results partially confirm our hypothesis, as presented in Appendix B. The negative R^2 coefficient from the OLS linear regression analysis suggests that `sentiment_score` may not be a good explanatory variable for market movements. This result indicates that the sentiment expressed in the FOMC meeting minutes may not have a significant impact on the daily returns of the S&P500 index and the VIX index.

Indicator	S&P 500	VIX
Train accuracy	54.17%	60.65%
Validation accuracy	62.50%	33.33%

Table 5: Results of market movement prediction

6 Concluding Remarks

We've discussed the results of sentiment analysis on various FOMC documents using the LSTM deep-learning model, which shows a promising accuracy that has been documented widely. However, further research is required to improve the ability to extract information regarding financial market responses to the release of FOMC meeting minutes. While Long Short-Term Memory (LSTM) models have demonstrated notable proficiency in detecting sentiment, there appears to be a weak association between such sentiments and market reactions, incongruent with the extant literature, which has consistently shown that the market does indeed react to such releases and incorporates new information into asset prices such as stock prices (Rosa, 2013) and Treasury yields (Kuttner and Shim, 2012). Thus, it may be necessary to explore alternative models that can provide a more comprehensive understanding of the information conveyed in FOMC meeting minutes.

It is also worth noting that the recent Large Language Models (LLMs), such as OpenAI's GPT series, have increasingly taken up the domain of NLP, providing enhanced capabilities in understanding and evaluating subjective opinions within textual data (Radford et al., 2019; Brown et al., 2020). In particular, LLMs have demonstrated their ability to effectively discern sentiment polarity in product reviews, social media posts, and other text-based data sources (Howard and Ruder, 2018; Zhang et al., 2018). Recent studies, such as BERT-based models, have further expanded these applications, improving the performance of sentiment classification tasks by leveraging pre-trained models fine-tuned on specific sentiment analysis datasets (Devlin et al., 2019; Sun et al., 2019). These advances in LLMs have led to more accurate and reliable sentiment analysis, enabling businesses and researchers to better comprehend consumer opinions, monitor brand reputation, and analyze public sentiment on various issues (Zhang and Liu, 2017; Cambria and White, 2014). Future research may extend beyond the limitations of the current Long Short-Term Memory (LSTM) mechanism and explore the potential of more advanced Language Model architectures (LLMs) to enhance the accuracy and reliability of the analysis of FOMC meeting minutes.

References

Apergis, N. (2015). The role of FOMC minutes for US asset prices before and after the 2008 crisis: Evidence from GARCH volatility modeling. *The Quarterly Review of Economics and Finance*, 55:100–107.

Bengio, Y., Simard, P., and Frasconi, P. (1994). Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166.

BOARD OF GOVERNORS of the FEDERAL RESERVE SYSTEM (2023). Meeting calendars and information. `https://www.federalreserve.gov/monetarypolicy/fomccalendars.htm`.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language Models are Few-Shot Learners. Issue: arXiv:2005.14165 arXiv: 2005.14165 [cs].

Cambria, E. and White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational Intelligence Magazine*, 9(2):48–57.

Campbell, J. Y., Lo, A. W., and MacKinlay, A. (1997). *The Econometrics of Financial Markets*. Princeton University Press.

Cannon, S. (2015). Sentiment of the FOMC: Unscripted.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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)*, pages 4171–4186. Association for Computational Linguistics. Place: Minneapolis, Minnesota.

DrLexus and Ozannen (2023). Federal Reserve FOMC Minutes & Statements Dataset. `https://www.kaggle.com/datasets/drlexus/fed-statements-and-minutes`.

Fama, E. F. (1965). Random Walks in Stock Market Prices. *Financial Analysts Journal*, 21(5):55–59.

Gers, F., Schmidhuber, J., and Cummins, F. (2000). Learning to Forget: Continual Prediction with LSTM. *Neural computation*, 12:2451–71.

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. MIT Press.

Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., and Schmidhuber, J. (2017). LSTM: A Search Space Odyssey. *IEEE Transactions on Neural Networks and Learning Systems*, 28(10):2222–2232. arXiv: 1503.04069 [cs].

Handlan, A. (2020). Text Shocks and Monetary Surprises: Text Analysis of FOMC Statements with Machine Learning.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. Issue: arXiv:1207.0580 arXiv: 1207.0580 [cs].

Hochreiter, S. and Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780.

Howard, J. and Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. Issue: arXiv:1801.06146 arXiv: 1801.06146 [cs, stat].

Huang, Y.-L. and Kuan, C.-M. (2021). Economic prediction with the FOMC minutes: An application of text mining. *International Review of Economics & Finance*, 71:751–761.

Jubinski, D. and Tomljanovich, M. (2013). Do FOMC minutes matter to markets? An intraday analysis of FOMC minutes releases on individual equity volatility and returns. *Review of Financial Economics*, 22(3):86–97.

Ke, S., Olea, J. L. M., and Nesbit, J. (2022). Robust Machine Learning Algorithms for Text Analysis.

Khedr, A. E., S.E.Salama, and Yaseen, N. (2017). Predicting Stock Market Behavior using Data Mining Technique and News Sentiment Analysis. *International Journal of Intelligent Systems and Applications*, 9(7):22–30.

Kuttner, K. N. and Shim, I. (2012). Taming the Real Estate Beast: The Effects of Monetary and Macroprudential Policies on Housing Prices and Credit.

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature*, 521(7553):436–444.

Lucca, D. O. and Trebbi, F. (2009). Measuring Central Bank Communication: An Automated Approach with Application to FOMC Statements. Issue: 15367 Series: Working Paper Series.

Nechio, F. and Wilson, D. J. (2016). How Important Is Information from FOMC Minutes?

Pagolu, V. S., Reddy, K. N., Panda, G., and Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. In *2016 International Conference on Signal Processing, Communication, Power and Embedded System (SCOPES)*, pages 1345–1350.

Pascanu, R., Mikolov, T., and Bengio, Y. (2013). On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, pages 1310–1318. PMLR.

Pennington, J., Socher, R., and Manning, C. D. (2014). GloVe: Global Vectors for Word Representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language Models are Unsupervised Multitask Learners. In *Language Models are Unsupervised Multitask Learners*.

Rosa, C. (2013). The financial market effect of fomc minutes. *Economic Policy Review*, 19(2).

Schmidhuber, J. (2015). Deep Learning in Neural Networks: An Overview. *Neural Networks*, 61:85–117. arXiv: 1404.7828 [cs].

Sun, T., Gaut, A., Tang, S., Huang, Y., ElSherief, M., Zhao, J., Mirza, D., Belding, E., Chang, K.-W., and Wang, W. Y. (2019). Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640. Association for Computational Linguistics. Place: Florence, Italy.

Tadle, R. C. (2022). FOMC minutes sentiments and their impact on financial markets. *Journal of Economics and Business*, 118:106021.

Tang, D., Qin, B., Feng, X., and Liu, T. (2016). Effective LSTMs for Target-Dependent Sentiment Classification. Issue: arXiv:1512.01100 arXiv: 1512.01100 [cs].

Wang, Y., Huang, M., Zhu, X., and Zhao, L. (2016). Attention-based LSTM for Aspect-level Sentiment Classification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 606–615. Association for Computational Linguistics. Place: Austin, Texas.

Whaley, R. (2000). The Investor Fear Gauge. *Journal of Portfolio Management - J PORTFOLIO MANAGE*, 26:12–17.

Yilmaz, B. (2022). Sentiment Analysis Methods in 2023: Overview, Pros & Cons.

Zhang, L. and Liu, B. (2017). Sentiment Analysis and Opinion Mining. In Sammut, C. and Webb, G. I., editors, *Encyclopedia of Machine Learning and Data Mining*, pages 1152–1161. Springer US, Boston, MA.

Zhang, L., Wang, S., and Liu, B. (2018). Deep Learning for Sentiment Analysis : A Survey. Issue: arXiv:1801.07883 arXiv: 1801.07883 [cs, stat].

Appendix A Keywords and Polarized terms

The lists are from Tadle (2022).

Table 6: Keywords by type

Hawkish keywords			
business	businesses	demand	economic
economy	employment	energy	equities
equity	expansion	financial	growth
housing	income	indicators	inflation
inflationary	investment	investments	labor
manufacturing	outlook	output	price
prices	production	recovery	resource
securities	slack	spending	target
toll	wage	wages	
Dovish keywords			
	accommodation	devastation	
	downturn	recession	
	unemployment		

Table 7: Polarized terms.

Positive terms				
abating [a]	accelerated	add	advance	advanced
augmented	balanced	better	bolsters	boom
booming	boost	boosted	eased	elevated
elevating	expand	expanding	expansionary	extend
extended	fast	faster	firmer	gains
growing	heightened	high	higher	improved
improvement	improving	increase	increased	increases
increasing	more	raise	rapid	rebounded
recovering	rise	risen	rising	robust
rose	significant	solid	sooner	spike
spikes	spiking	stable	strength	strengthen
strengthened	strengthens	strong	stronger	supportive
up	upside	upswing	uptick	TF

Negative terms				
adverse	back	below	constrained	contract
contracting	contraction	cooling	correction	dampen
damping	decelerated	decline	declined	declines
declining	decrease	decreases	decreasing	deepening
depressed	deteriorated	deterioration	diminished	disappointing
dislocation	disruptions	down	downbeat	downside
drop	dropping	ebbed	erosion	fade
faded	fading	fall	fallen	falling
fell	insufficient	less	limit	low
lower	moderated	moderating	moderation	reduce
reduced	reduction	reluctant	removed	restrain
restrained	restraining	restraint	resumption	reversed
slack	slow	slowed	slower	slowing
slowly	sluggish	sluggishness	slumped	soft
softened	softening	stimulate	strained	strains
stress	subdued	tragic	turmoil	underutilization
volatile	vulnerable	wary	weak	weakened
weaker	weakness			

Appendix B OLS Results of `vwretd` on `sentiment_score`

	Coefficient	Standard Error
Intercept	0.0014	0.002
sentiment_score	0.0113	0.022
Model Statistics		
R^2	0.003	
Adj. R^2	-0.009	
F-statistic	0.2601	
Observations	87	

Table 8: OLS Regression Results

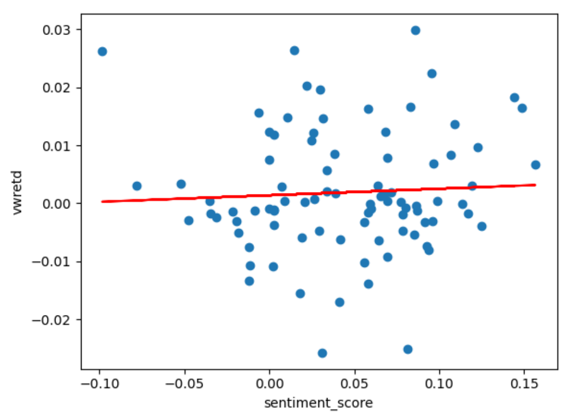

Figure 8: Fitted regression

Author's Contributions

Stella Zhu contributed to the literature review, data acquisition, analysis, interpretation, and manuscript writing.

Ollie Sun contributed to data analysis, interpretation, model building, and manuscript writing.