From Text to Trends: Multi-Dimensional Document Level Insights from FOMC Minutes for Predicting Stock Returns

Anonymous ACL submission

Abstract

The US Federal Open Market Committee (FOMC) releases the monetary policy meeting minutes after every FOMC meeting. This has important information about the trends in monetary policy which has broad implications. In this work, we take a multi-faceted view to the FOMC minute analysis at the document level, and show that this approach has better potential in predicting the stock market returns of S&P 500 companies over 30 years. Further, we release a dataset spanning 31 years from 1993 to 2024 for FOMC minutes at the document level, which rates each minute on four facets: Uncertainty (certain or uncertain), Sentiment (hawkish, dovish, or neutral[1]), Complexity (complex or simple), Consensus (consensus or no-consensus), and show that model finetuned on this data has better stock return prediction capability. To the best of our knowledge, there is no document level analysis of FOMC minutes. We demonstrate the importance of these facets by analysing the impact on stock price returns for the next day of the meeting minute release date and show that relying on various viewpoints is better than purely relying on the sentiment data. Our approach is also consistent in analyzing the impact of the Indian FOMC committee's meeting minutes on Indian stock market (NIFTY50) return. We have released the dataset on google drive[2].

1 Introduction

The US Federal Open Market Committee (FOMC), as the branch of the Federal Reserve, is responsible for overseeing and setting the monetary policy, plays a pivotal role in the U.S. economic and financial systems. This body's decisions on interest rates and monetary policy not only influence macroeconomic conditions but also affect financial markets significantly. The releases of FOMC statements and minutes are therefore highly anticipated events, parsed meticulously by market participants for insights into future policy directions. Amidst this backdrop, the textual content of FOMC communications emerges as a rich dataset ripe for advanced analytical techniques. These documents contain nuanced indicators of the Committee's perspectives on economic conditions, inflation, and future monetary policy actions. The challenge, however, lies in extracting actionable insights from this complex information efficiently and accurately. Recent advancements in large language models (LLMs) offer compelling solutions to this challenge. LLMs have huge context sizes, and this enables document level analysis. Document level analysis enables a better coherent view of the FOMC document enabling more accurate and contextually grounded results compared to sentence-level approaches. The impact of the FOMC minutes on the market goes beyond the broad sentiment (hawkish or dovish) of the monetary policy. The consensus amongst the members, the uncertainty about the future economic scenario and the complexity of the information conveyed through the FOMC minutes has important implications for the broad market. In this work, we will be analyzing the FOMC documents from these multi-faceted viewpoints, and using this information to predict the stock market returns. To the best of our knowledge, there has been no document level LLM analysis on the impact of FOMC minutes on stock return. Further, we release a dataset for the FOMC minutes for all the four facets: Uncertainty, Complexity, Sentiment, Consensus with high inter-annotator agreement, and show that this data helps predict better stock market reactions and unravels new direction in research of FOMC minutes for stock market prediction. Further, we expand our experiments on the Reserve bank of India (RBI) monetary policy minutes data to show the effectiveness of our multi-faceted

[1] https://www.bgifx.com/news/knowledge/hawkish-vs-dovish-central-bank-policy-stance-for-fx-trading

[2] We will release the dataset, code, and the models used upon acceptance.

approach across different geographic region. RBI monetary committee is an Indian equivalent to the US Federal monetary policy committee.

2 Related Work

Textual analysis of financial documents has attracted wide attention recently. There have been language models trained on financial data like BERT (Devlin et al., 2018), RoBERTa (Liu et al., 2019), GPT (Brown et al., 2020); OpenAI, 2023 and LLAMA (Touvron et al., 2023), PIXIU (Xie et al., 2023). There have been datasets in financial sentiment analysis like Financial Phrase Bank (Malo et al., 2014), SemEval-2017 (Cortis et al., 2017), AnalystTone Dataset (Huang et al., 2023), Headline News Dataset (Sinha and Khandait, 2021) and Trillion Dollar Words (Shah et al., 2023). However, these works do not look at FOMC policy documents at the document level. There has been work on building models for sentiment analysis (Zhang et al., 2023), Wonseong Kim et al, (Luo and Gong, 2024)). (Guo et al., 2023) worked on benchmarking ChatGPT on various financial datasets, which includes the Trillion dollar words dataset. To the best of our knowledge, there is no dataset that looks at FOMC minutes at the document level and studies it's impact on the stock market returns. (Luan et al., 2024) analyse the impact of sentiments on Chinese stock markets. (Tang et al., 2023) look at the impact of sentiments on the cryptocurrency markets. (Glodd and Hristova, 2023) look at the impact of financial information present in 10-K annual reports on the stock market. In this work, we go beyond the existing work by analyzing FOMC documents from various viewpoints, and predicting the stock market returns on the next day of release of the meeting minutes.

3 Dataset Construction

We will begin with the information about the dataset, and finetuning. Next, we will present the impact on stock market returns, and how our finetuned models have improved the stock return prediction. We have prepared a dataset of FOMC minutes spanning 31 years, from 1993 to 2023, labeled on four facets: Uncertainty, Complexity, Sentiment, Consensus with a total of 245 samples. We collect all available text from these documents, resulting in a substantial corpus of monetary policy communications.

3.1 Facet definition

Each FOMC document has been tagged on four facets: Uncertainty, Consensus, Sentiment, and Complexity. We define each of these facets as following:

- Uncertainty: Measures the degree of ambiguity or unpredictability in the FOMC policy minute, scored among 1 (Certain) and 2 (Uncertain).
- Consensus: Assesses the level of agreement or unity expressed, with 1 indicating low consensus and 2 indicating high consensus among committee members.
- Sentiment: Evaluates the tone of the minute in terms of monetary policy stance, where 1 denotes dovish (favoring lower interest rates) and 2 denotes hawkish (favoring higher interest rates) and 3 denoting neutral sentiment.
- Complexity: Reflects the intricacy of the language used, among 1 (not-complex) and 2 (complex).[3]

We discard a total of eight documents which had conflicting annotation on at least one of the four labels. On a side note, the previous datasets for the FOMC minutes have focused on sentence level sentiment labeling. Our work differs from the previous works as we look at the minutes at a document level on all four novel facets, and demonstrate the significance of the document level analysis by studying the impact on stock market prices.

3.2 Annotators

The dataset was manually annotated by a team of two annotators, both pursuing graduate degree in Finance. These annotators were selected based on their expertise in finance and familiarity with monetary policy and were paid hourly according to the demographic. Each annotator was provided with detailed guidelines and examples to ensure a consistent understanding of the labeling criteria. Throughout the annotation process, the annotators independently labeled each segment to capture their individual interpretations of the facets. We employ

[3]We had a 1-4 scoring scheme previously which had a range of scores for every facet. The issue with this scheme was that the scoring became highly ambiguous due to the unclear boundaries between scores, reducing the inter-annotator agreement. Hence we have binary labels for each category, having clear demarcation.

the Label Studio platform[4] for the annotation process.

3.3 Annotation Consistency

To ensure high quality annotations, the annotators went through a training and calibration phase where they annotated a small set of documents and discussed discrepancies to align their understanding. For the annotation consistency score, we employ Cohen's Kappa (McHugh, 2012) to evaluate. We obtain a very high agreement score of 0.91 and 0.93 for the sentiment and consensus facets, which falls between 0.8 and 1, demonstrating an almost perfect agreement (Landis and Koch, 1977) among the annotators. For the uncertainty and complexity facets, we obtain satisfactory agreement of 0.72 and 0.79 respectively.

3.4 Dataset Statistics

We have removed eight samples where there was disagreement between the annotators on any of the four facets, to obtain the high quality dataset. The final dataset contains minutes spanning from 1993 to 2024, with 245 meeting of minutes, each of which contains 39K words on an average, calculated using NLTK's (Bird and Loper, 2004) word_tokenize method. 1 shows the distribution of labels within each facet in our dataset.

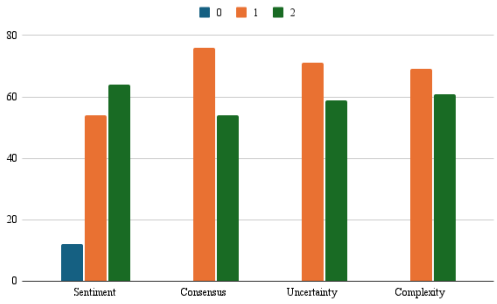

Figure 1: Label distribution of various facets in our dataset. X axis denotes the label and Y axis denotes the count of each label corresponding to that facet.

4 Experiments

To evaluate the effectiveness of our dataset in training language models for policy analysis, we perform zero shot inference, and finetuning of the state-of-the-art large langauge models. We utilize the chronologically first 80% of the data for training and validation (further 80-20 split for train-validation) and the remaining 20% of the dataset for testing purpose. We treat the zero shot version of the above listed open source models along with ChatGPT as the baseline for our work. We have utilized OpenAI's ChatGPT3.5 model programmatically to perform the inference, while the open source models were inferred by accessing the models from huggingface. We perform 16-bit quantization to the open source models to save GPU computation power. Additionally, we perform a rule based post processing step to extract the score from the corresponding full sentence model generation using regular expressions. We fine-tuned the Mistral v0.3 7B model (Jiang et al., 2023) and the Llama 3.2 3B model (Touvron et al., 2023) on each facet separately. The fine-tuning was performed using a consistent training pipeline with the same testing data as the baselines to ensure comparability. We perform parameter efficient finetuning using LoRa adapters (Hu et al., 2021) with the default PEFT configuration. We have quantized the model to 16-bit quantization, to reduce the GPU memory requirement. We have trained the models with AdamW optimizer (Loshchilov and Hutter, 2017), with a batch size of 1, and 4 gradient accumulation steps, an initial learning rate of 1×10^{-5}, and a weight decay of 0.1. We utilize the A100 (40 GB) GPUs for the finetuning process. Table 1 shows the model setting we have used for the finetuning process. We report the macro F1 score and overall Accuracy for each class separately. Table 3 shows the finetuning and baseline mean results obtained on the test split across three runs. We can observe that the finetuned Llama 3.2 3B model outperforms all the other models on the sentiment and consensus facets, and Mistral v0.3 Instruct 7B finetuned model outperforms all other models on complexity and uncertainty facets. Among the zero shot results, we can see that the Mistral v0.3 model outperforms the other models. We can observe that finetuning the open source models outperforms the performance of the closed source GPT model as well. On the contrary to general expectation, higher number of model parameters does not guarantee a better performance on the dataset. We can observe this through the performance of Mistral and Mixtral models, where Mistral performed better on average when compared to Mixtral on the four facets.

[4] https://labelstud.io/

Parameter	Value
Quantization	4-bit
Initial Learning rate	1×10^{-5}
Weight decay	0.1
Batch size	1
Grad accumulation	4
GPU-type	A100 (40 GB)
Optimizer	AdamW
Finetune type	PEFT (LoRa)

Table 1: Finetuning setting for our work

Model	RMSE (SPY on FOMC)	RMSE (Nifty50 on RBI)
Mistral (zero-shot)	0.0095	0.0124
Mixtral (zero-shot)	0.0111	0.0129
GPT3.5 (zero-shot)	0.0109	0.0135
Llama3.1 (zero-shot)	0.0103	0.0112
Llama3.2 (zero-shot)	0.0113	0.0118
Mistral (fine-tuned)	0.0091	-
Llama3.2 (fine-tuned)	0.0094	-

Table 2: RMSE values for SPY prediction on FOMC documents and Nifty50 prediction on RBI documents

Model	Uncertainty		Consensus		Sentiment		Complexity	
	Acc	F1-macro	Acc	F1-macro	Acc	F1-macro	Acc	F1-macro
Mistral (zero-shot)	83.51%	0.7323	83.42%	0.7923	84.46%	0.7913	78.26%	0.7192
Mixtral (zero-shot)	79.41%	0.7171	82.65%	0.7671	81.93%	0.7620	75.06%	0.6862
Llama3.1 (zero-shot)	80.37%	0.7178	84.42%	0.7671	83.72%	0.7713	77.92%	0.6935
Llama3.2 (zero-shot)	78.59%	0.7074	83.82%	0.7671	82.79%	0.7702	76.82%	0.6912
GPT3.5 (zero-shot)	72.23%	0.6647	81.19%	0.7344	81.62%	0.7529	75.81%	0.6738
Mistral (fine-tuned)	86.81%	0.7925	87.92%	0.8225	89.28%	0.8062	81.83%	0.7341
Llama3.2 (fine-tuned)	86.45%	0.7802	88.83%	0.8118	90.06%	0.8102	80.64%	0.7269

Table 3: Accuracy and F1-macro scores across different facets (Uncertainty, Consensus, Sentiment, and Complexity)

4.1 Comparison with Sentence level Model

In this section, we assess the performance of a well-known sentiment model, Trillion Dollar Words (Shah et al., 2023), which was trained on FOMC sentences. This model employs a sentence-level fine-tuned RoBERTa architecture (Liu et al., 2019). To evaluate the impact of document-level sentiment analysis, we applied the RoBERTa model to our dataset by conducting inference on each sentence within a document and assigning the majority label across sentences as the document's sentiment label. We utilize nltk's (Bird and Loper, 2004) sentence tokenizer to perform the tokenization. This approach yielded an accuracy of 22.08%, considerably lower than the performance of both our fine-tuned and zero-shot models.

Upon analyzing the results, we found that the RoBERTa model frequently classified many sentences as neutral, leading to an overall neutral sentiment for entire documents. To address this issue,

we removed samples with neutral ground truth label from the analysis. In this approach, we performed inference on the non-neutral documents using only the hawkish and dovish labels, excluding neutral sentences, and selected the majority label between these two labels. This adjustment improved the RoBERTa model's accuracy to 54.92%, yet it remained significantly lower than our models. These results underscore the limitations of sentence-level models for sentiment analysis in FOMC documents and highlight the importance of document-level analysis for better performance.

5 Application: Stock return prediction

In this work, one of the prominent contributions is to show the effectiveness of the multi-faceted application of LLMs on the FOMC minutes to predict stock market returns, which is the difference in value of stock on the current day with the previous day. We perform an extensive analysis spanning S&P 500 companies, and observe that the multi-faceted view improves the predictability. We rely on the Yahoo Finance dataset to obtain the day wise stock values.

5.1 Task formulation and Feature Extraction

We tried to study the impact of FOMC minutes on the stock returns. For this we consider twelve features: uncertainty, consensus, sentiment, and complexity scores, their mean value across the past K meetings (K is a hyperparameter), and their variance across the past K meetings. Here is a comprehensive definition of each of the parameters:

- Uncertainty, Consensus, Sentiment, Complexity: These labels are defined in 3.1.
- Average <Feature>: Average of the feature value across the past K meetings.
- Variance <Feature>: Variance of the feature value across the past K meetings.

We hypothesize that the mean of the scores and variance determine the trend in the recent past for that facet, and this provides important information to the prediction model. We prepare these twelve values for each of the FOMC minutes, forming the feature values for each FOMC meeting minute document. We perform experiments with two model types: Zero shot, where all the scores have been obtained in a zero shot setting, and finetuned, where we rely on the finetuned models for the uncertainty, complexity, sentiment, and consensus scores.

5.2 Experimental setup and Results

The first experiment we perform is to predict the SPY return value from the 24 features curated. SPY is an exchange-traded fund that tracks the performance of the S&P 500 index. The S&P 500 is a stock market index that measures the performance of 500 large cap publicly traded companies in the United States. Specifically, we predict the SPY value of the day after meeting minute release, as that would capture the impact of the FOMC meeting minute effectively. We can see from 3 that the zero shot evaluation was decent when compared to finetuning. For this experiment, we take the minutes ranging from 1994 to 2024, covering a diverse time span for better reliability of the results. We first perform inference of the model on all the minutes and compile the 12 feature values for each FOMC meeting minute. Next, perform linear regression using the sklearn library (Pedregosa et al., 2011) with the chronologically first 80% of the data utilized for training, and the remaining for prediction. We report the root mean squared error (RMSE) value for the predicted values. Table 2 shows the results obtained. We perform this analysis for a K value of 5. This means that we consider the impact of the past 5 meetings on the current stock return value. In section 2, we have performed a study on the effect of K value on the RMSE score for the Mistral model. We can see that the finetuned models are able to predict the stocks better than the zero shot models. This correlates to the performance on our dataset, as highlighted in table 3.

5.3 Effect of K value

In this section, we explore the effect of K value on the final RMSE value for the regression. The experimental setup remains the same, and we vary the value of K from 0 to 10. K=0 denotes that we do not take any previous values, and K=10 denotes that we take the past 10 values into account for prediction. Figure 2 shows the RMSE values for different K values for various models. We can see that the value of 5 seems optimal for K across models, and hence we have chosen K=5 in the previous section.

6 Feature Importance Analysis

To further understand the contributions of individual features to the stock return predictions, we conducted a statistical analysis of feature importance.

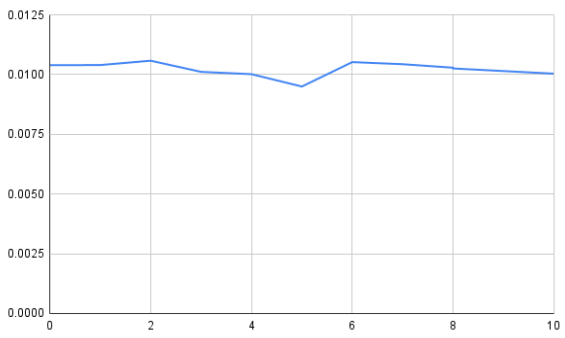

Figure 2: Effect of K value on RMSE

For the study to be robust, we first extract the stock return values for the SPY top 500 companies. We then perform regression for all the models across all the minutes for all the companies. Post that, we obtain two values, highlighted below:

- Coefficient Averaging: We averaged the regression coefficients of each feature across all S&P 500 individual stocks to determine their overall influence.
- Significance Counting: We counted the number of times each feature's coefficient was observed to be significant ($p < 0.05$) across the 500 stocks.

Please refer to 5 for the details on the significance results of each feature for the Mistral model. The average coefficient of each feature across top S&P 500 companies has been reported. A negative average coefficient denotes that a higher value of the feature results in a reduction of stock return, and vice versa. The other two columns in the table denote the number of companies for which the feature value is significant. The important observation with the results is that the features apart from sentiment also proved to be significant, e.g. minute uncertainty score. This shows that we need to go beyond just the sentiment in prediction of stock returns based on FOMC documents.

7 Experiments with RBI documents

In an attempt to further validate our methodology, we also performed our analysis on the monetary policy minutes from a different geographic region; India. The Indian central bank (Reserve Bank of India: RBI) also publishes similar meeting minute documents on its website. Since they belong to a different country and are being written by a different committee, this provides diversity to our experimental setup to validate the usage of four facets on stock price return prediction. We first scrape the RBI meeting of minutes for all the years ranging from 2016 to 2024 and perform zero shot inference on these documents in the same model setting as the FOMC documents. For each RBI document, we obtain 12 feature values as we have done previously. The target prediction value is the Nifty50 index stock return for the next day of the meeting minutes release date as well as all the Nifty50 individual stock return. We fit a linear regression model to predict these returns as before. Table 2 shows the zero shot results on the RBI documents. We can see that the RMSE values are comparable to the FOMC RMSE values and are small enough. Further, we can see the significance results in 4 computed for the Nifty50 stock values. There are features for which number of times it was significant is greater than 25 out of 50 stocks, showing that these feature are playing a big part in the regression prediction.

8 Conclusion

In this paper, we study the impact of FOMC meeting minutes on the stock returns. We show through experiments that going beyond the sentiment of the FOMC minutes at the document level helps improve the predictive power of the FOMC minutes on stock returns. Further, we present the construction of the data for hawkish and dovish sentiments along with other multi-faceted information (complexity, consensus, uncertainty) contained in the FOMC minutes, and benchmark LLMs on our data. We also finetune Mistral and Llama model on this dataset, and show that the stock return prediction has improved. We believe that this work can be a great exploratory base for financial text processing, and it's impact on the stock market returns.

9 Limitations

In this work, we currently experiment with data from 1994 to 2024 to have a representative data. For the meeting minutes prior to that date, we need to experiment to verify the performance. Another limitation is the transferability of the study to other forms of financial text, and languages. This is left to future work.

10 Ethical Statement

The study was done in accordance to the ACL Ethics policy. The annotators were finance graduate students, who were well versed with the FOMC

Feature	Avg Coefficient	No. 10% sig	No. 5% sig
Minute complexity score	0.0022	21	11
Minute consensus score	-0.0004	12	4
Minute sentiment score	0.0004	13	7
Minute uncertainity score	-0.0008	23	15
Minute avg complexity score	-0.0304	31	23
Minute avg consensus score	0.0069	10	3
Minute avg sentiment score	-0.0005	4	2
Minute avg uncertainity score	0.0314	35	29
Minute var complexity score	-0.0162	24	14
Minute var consensus score	-0.0062	16	10
Minute var sentiment score	-0.0054	29	12
Minute var uncertainity score	0.0031	30	19

Table 4: Significance results of each feature for the zero shot Mistral model. The avg coefficient denotes the average regression coefficient across the Nifty 50 companies. The No. 10% sig denotes the number of companies for which this feature was 10% significant. ($p<0.1$), and similarly 5% significant for No. 5% sig.

data. Finally, we also request the users of this work to make sure to apply discretion when using the stock return predictions as the market is highly volatile. This is purely an academic study. The data we release does not contain any harmful content, and does not reveal the identity of the authors or the annotators.

Feature	Avg Coefficient	No. 10% sig	No. 5% sig
Minute complexity score	0.0015	121	68
Minute consensus score	-0.0001	60	28
Minute sentiment score	0.0021	86	55
Minute uncertainity score	-0.0020	135	83
Minute avg complexity score	-0.0051	219	160
Minute avg consensus score	0.0046	183	112
Minute avg sentiment score	-0.0033	49	21
Minute avg uncertainity score	0.0028	64	34
Minute var complexity score	-0.0002	24	11
Minute var consensus score	-0.0008	56	30
Minute var sentiment score	-0.0033	72	44
Minute var uncertainity score	0.0044	87	64

Table 5: Significance results of each feature for the zero shot Mistral model. The avg coefficient denotes the average regression coefficient across the S&P 500 companies. The No. 10% sig denotes the number of companies for which this feature was 10% significant. ($p<0.1$), and similarly 5% significant for No. 5% sig.

Prompt Type	Prompt Description
[FOMC/RBI] Uncertainty prompt	**Instruction:** You are given a statement which is the [FOMC/RBI] Monetary policy meeting minute. Please analyze the statement and give a score about the degree of uncertainty expressed in the discussions regarding economic conditions and policy implications? Please score the statement with a 1 or 2, where 1 denotes an uncertain statement, 2 denotes a certain statement? I want to analyze the impact of this meeting statement on the stock returns. **Statement:** <Document> **Score:**
[FOMC/RBI] Consensus prompt	**Instruction:** Generate a degree of non-consensus score for [FOMC/RBI] Monetary policy meeting minute in terms of economic conditions and policy implications. Score the statement with a 1 or 2, where 1 indicates consensus, 2 indicates non-consensus. This analysis will be used to assess the statement's impact on stock returns. **Statement:** <Document> **Score:**
[FOMC/RBI] Sentiment prompt	**Instruction:** You are given a statement which is the [FOMC/RBI] Monetary policy meeting minute. Please analyze the statement and give a score about the tone in terms of the monetary policy? Please score the statement with a scale of 1 or 2, where 1 denotes dovish tone, 2 denotes hawkish tone. I want to analyze the impact of this meeting statement on the stock prices. **Statement:** <Document> **Score:**
[FOMC/RBI] Complexity prompt	**Instruction:** Assess the complexity of the [FOMC/RBI] monetary policy meeting minute. Score the statement with a 1 or 2, where 1 indicates a complex statement, 2 indicates a simple statement. This evaluation will help in understanding the comprehensibility of the statement and its potential impact on market reactions. **Statement:** <Document> **Score:**

Table 6: Table containing the prompt used for each of the facets for FOMC and RBI zero shot experiments

References

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Alexander Glodd and Diana Hristova. 2023. Extraction of forward-looking financial information for stock price prediction from annual reports using nlp techniques.

Yue Guo, Zian Xu, and Yi Yang. 2023. Is chatgpt a financial expert? evaluating language models on financial natural language processing. *arXiv preprint arXiv:2310.12664*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Allen H Huang, Hui Wang, and Yi Yang. 2023. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*, 40(2):806–841.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Yuchen Luan, Haiyang Zhang, Chenlei Zhang, Yida Mu, and Wei Wang. 2024. Stock price prediction with sentiment analysis for Chinese market. In *Proceedings of the Joint Workshop of the 7th Financial Technology and Natural Language Processing, the 5th Knowledge Discovery from Unstructured Data in Financial Services, and the 4th Workshop on Economics and Natural Language Processing @ LREC-COLING 2024*, pages 167–177, Torino, Italia. ELRA and ICCL.

Wei Luo and Dihong Gong. 2024. Pre-trained large language models for financial sentiment analysis. *arXiv preprint arXiv:2401.05215*.

P. Malo, A. Sinha, P. Korhonen, J. Wallenius, and P. Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. *arXiv preprint arXiv:2305.07972*.

Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 589–601. Springer.

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023. FinEntity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 349–356.