Stock Market Sentiment Analysis in Python

ABSTRACT

Understanding the overall sentiment of stock market discussions is essential for making informed financial decisions. This project aims to develop a reliable, efficient method for determining the general sentiment of stock market news from Twitter using sentiment analysis. Given the nuance in financial language, we focused on filtering data to retain relevant content and employed machine learning models to classify the sentiment of stock market news as Positive, Negative, or Neutral. The results underscore a predominantly neutral tone in stock-related news, highlighting a need for refined methods to capture subtle sentiment in finance-specific language. This analysis provides a foundation for leveraging sentiment insights in stock market analysis.

KEYWORDS: Sentiment Analysis, Financial Market Sentiment, Natural Language Processing (NLP), Logistic Regression, and Stock Market News

Author Information
Iman Sellars, DePaul University Undergraduate, Isellars@depaul.edu
Jana Abdul. DePaul University Undergraduate, jabdulra@depaul.edu

INTRODUCTION

This project focuses on analyzing the sentiment of stock market discussions to provide timely insights into public sentiment trends. By analyzing stock market-related news from Twitter, we aim to build a robust model that can reliably classify this data as Positive, Negative, or Neutral.

We chose this dataset because Twitter serves as a major platform for financial discourse, where news and opinions about stocks and market events are shared every day. Understanding the sentiment in these discussions can offer insights, as changes in public sentiment can be early indicators of shifts in market dynamics. Through this project, we seek to identify patterns in sentiment distribution and better understand the challenges in accurately classifying nuanced financial text data. Ultimately, this analysis could pave the way for more effective sentiment-based market prediction tools and improve financial sentiment analysis models tailored specifically to stock market data.

LITERATURE REVIEW

Below is a list of summaries corresponding to academic papers read to aid in this financial analysis. Here the short titles of the articles are shown. There will be a summary reference table following this list. Note there is a reference section at the end of this report. You can find the full names of the papers there.

Literature Review Topic 1: Sentiment Data Analysis for Finance

1.Sentiment Analysis and Machine Learning in Finance

This study uses StockTwits data to understand financial sentiment using machine learning techniques such as Naive Bayes and Support Vector Machines. It emphasizes preprocessing methods like bigrams and emojis to help model performance.

2.Analysis of Financial Information

This paper applies deep learning to financial sentiment analysis, avoiding manual labeling. It predicts market responses with precision rates of 62% for positive and 55% for negative sentiment.

3.From Text Representation to Financial Market Prediction

This study evaluates text mining methods (e.g., Bag-of-Words, Word2Vec, BERT) and explores the impact of sentiment aggregation on predicting financial market trends.

4.Deep Learning for Sentiment Analysis

This study explores advanced text representation techniques such as embeddings like BERT and LSTMs, focusing on their applications in financial text sentiment analysis.

5.Comparing Traditional News and Social Media

This study explores social media sentiment captures immediate, unfiltered reactions, whereas traditional news provides more structured insights. This dichotomy is crucial for sentiment sourcing.

6.Does Twitter Affect Stock Market Decisions?

This paper links public sentiment from tweets to stock price volatility, highlighting the value of real-time sentiment tracking for financial decisions.

7.Topic Modeling-Based Sentiment Analysis

Employing Latent Dirichlet Allocation (LDA), this study links topics and sentiments to market dynamics.

8.A Comprehensive Review on Sentiment Analysis of Social Web Media

This paper discussing challenges in social media sentiment analysis. It focuses on preprocessing strategies such as stemming and tokenization.

9.Transforming Sentiment Analysis

This paper introduces transformers like FinBERT, that emphasizes attention mechanisms' role in enhancing sentiment prediction.

10.Emotions, Moods, and Hyperreality

This study explores emotional expressions and hyperrealistic digital contexts, analyzing their influence on market sentiment and investor behavior.

Figure 1:Sentiment Analysis Academic Paper Summary Table

Year Published	References	Journal
2019	Renault, T	Digital Finance
2021	Wujec, M.	Risk and Financial Management
2022	Farimani, S. A., Jahan, M. V., & Milani Fard, A.	Information
2018	Zhang, L., Wang, S., & Liu, B.	University of Illinois at Chicago
2022	Smith, S., & O'Hare, A	Journal of Big Data
2021	Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., & Sandoval-Almazán, R	Cognitive Computation
2018	Nguyen, T. H., & Shirai, K.	Japan Advanced Institute of Science and Technology
2024	Shah, P., Desai, K., Hada, M., Parikh, P., Champaneria, M., Panchal, D., Tanna, M., & Shah, M.	Springer
2024	Georgios Fatouros	University of Piraeus, Karaoli and Dimitriou 80, Piraeus
2020	Lazzini, A., Lazzini, S., Balluchi, F., & Mazza, M	Accounting, Auditing & Accountability Journal

Literature Review Topic 2: Machine Learning for Analysis for Finance

11.Validation of Text Data Preprocessing Using Neural Networks

This paper examines how stopword removal and stemming enhance sentiment classification accuracy, particularly in financial text.

12.Enhancing Legal Sentiment Analysis

This study focuses on legal documents uses in deep learning with domain-specific embeddings to improve sentiment detection.

13.Comparative Analysis of Word2Vec and GloVe

This study compares Word2Vec and GloVe embeddings, finding GloVe more effective for capturing semantic relationships in financial data.

14.Sentiment Analysis Using Deep Learning Techniques

This paper reviews models like CNNs and transformers, emphasizing their ability to handle multilingual and domain-specific sentiment tasks.

15.A Comparative Evaluation of Preprocessing Techniques

This study evaluates preprocessing strategies such as normalization and tokenization, identifying steps crucial for optimizing financial sentiment models.

16.From Word Vectors to Multimodal Embeddings

This paper highlights the integration of textual and non-textual data for enhanced sentiment predictions in financial contexts.

17.Comparative Analysis of Embedding Techniques

This paper contrasts transformer-based embeddings with traditional approaches, highlighting their advantages in sentiment analysis.

18.A Review of Sentiment Analysis

This paper discusses sentiment analysis challenges such as sarcasm detection, emphasizing advanced models' role in financial text contexts.

19.Evaluating Word Embedding Methods

This study focuses on embedding methods for improving classification accuracy in financial sentiment, advocating hybrid embedding approaches.

20.The Influence of Preprocessing on Text Sentiment Analysis

This paper illustrates how minimal preprocessing techniques improve model consistency and accuracy across datasets.

Figure 2: Machine Learning for Sentiment Analysis Academic Paper Summary Table

Year Published	Reference	Journal
2023	Woo, H., Kim, J., & Lee, W	Erasmus University Rotterdam
2024	Abimbola, B., de La Cal Marin, E., & Tan, Q.	MDPI
2023	Joshua Sopuru a , Adah Alubo b, Princess Chinemerem Iloh c , Oluwaseun Augustine Lottu	International Journal of Social Sciences and Scientific Studies
2023	Sahoo, C., Wankhade, M., & Singh, B. K	International Journal of Multimedia Information Retrieval
2018	Symeonidis, S., Effrosynidis, D., & Arampatzis, A.	Expert Systems with Applications
2023	Zhang, C., Peng, B., Sun, X., Niu, Q., Liu,	arXiv
2023	Wang, C.	Erasmus University Rotterdam.
2024	Sharma, N. A., Ali, A. B. M. S., & Kabir, M. A.	Erasmus School of Economics
2024	Kraayeveld, K.	Erasmus School of Economics
2020	HaCohen-Kerner, Y., Miller, D., & Yigal, Y	PLOS ONE

Methods

Data

The dataset used in this analysis is made from two columns: Sentence and Sentiment, with each row representing a unique entry of stock market-related text data and its corresponding sentiment label.

Explanatory Variable

The ‘Sentence’ column, containing the text data, functions as the explanatory variable. Each entry in this column provides a sentence or statement associated with stock market news, events, or opinions. This text will undergo a series of preprocessing and transformation steps to extract relevant features, making it suitable for sentiment analysis.

Outcome Variable

The ‘Sentiment’ column serves as the outcome variable or target variable for this analysis. It indicates the sentiment classification of each sentence, with possible labels such as ‘Positive’, ‘Negative’, or ‘Neutral’. This column provides the ground truth for model training and evaluation, helping to assess how accurately the model can classify new, unseen sentences based on their content.

The structure of this dataset is well-suited for natural language processing (NLP) and machine learning-based sentiment analysis. By transforming and analyzing the text data in the ‘Sentence’ column, we aim to build a model that accurately predicts the sentiment reflected in the Sentiment column. This insight is valuable for understanding general market sentiment trends and potentially making informed decisions based on sentiment shifts.

Statistical Summary of the Data

Sentiment Distribution

The dataset comprises three sentiment classes: ‘Positive’, ‘Negative’, and ‘Neutral’, though precise counts were not provided here. Based on earlier summaries, the majority of the dataset tends to be ‘Neutral’, with ‘Positive’ and ‘Negative’ sentiments being less common.

Sentence Length Analysis:

Count: 108,750 sentences

Average Length: Approximately 12.6 words per sentence

Minimum Length: 1 word

Maximum Length: 81 words

25th Percentile: 6 words

Median (50th Percentile): 9 words

75th Percentile: 15 words

95th Percentile: 36 words

These statistics indicate that most sentences are concise, with half of the dataset's entries being under 9 words, while a minority of longer sentences reach up to 81 words.

Exploratory Analysis

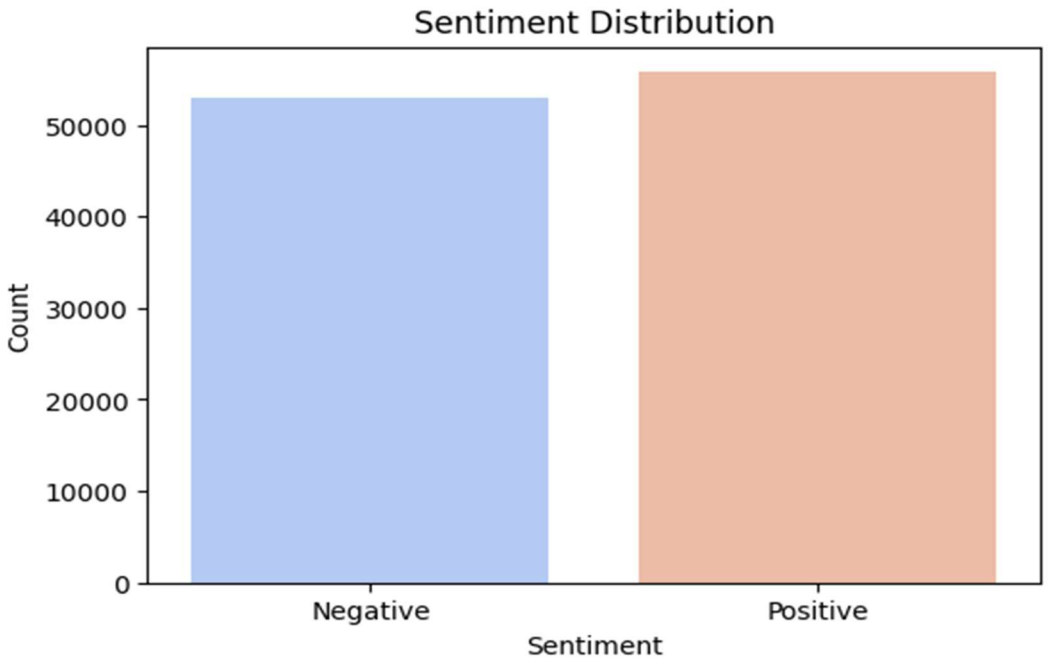
In order to gain a greater comprehension of the dataset, we performed simple EDA. We found patterns that helped guide the cleaning, preprocessing, and modeling process for our project.

1. Distribution of Sentiment

The goal is to ascertain how evenly positive and negative feelings are distributed throughout the sample.

We used a bar plot to display the sentiment distribution. Understanding any possible bias in the dataset is essential before we start the modeling process.

Figure 3: Financial Analysis Sentiment Distribution



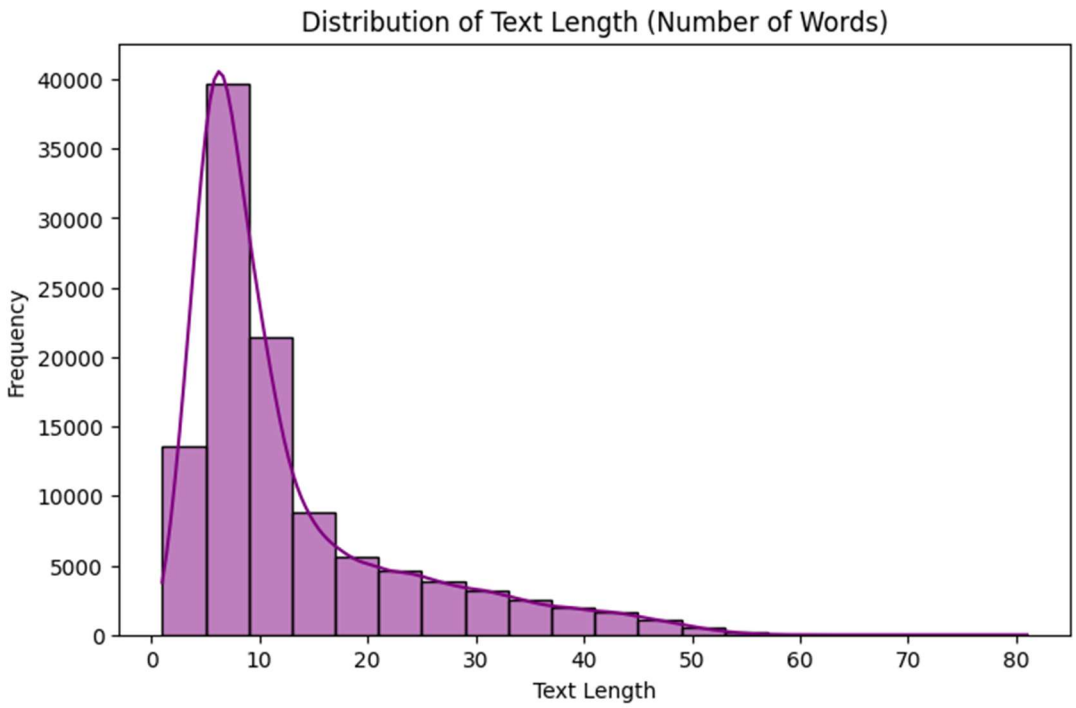
The data set is slightly imbalanced in favor of positive sentiments. However, since the data set is so large, the distribution is still fairly normal and distributed evenly across our dataset. No need to resample.

2. Distribution of Text Length

By analyzing the distribution of text lengths (measured in word counts), one may determine the average length of articles.

Each article's word count was determined, and a histogram was used to show the distribution. This was helpful when considering a model.

Figure 4: Sentiment Analysis Text Length Distribution



The histogram indicates that most sentences in the dataset are of short to medium length. This suggests a need to manage outliers and ensure that longer articles do not disproportionately influence the analysis.

3. Positive and Negative Sentiment Word Clouds

The goal here is to visually identify the most frequently used terms in articles with both positive and negative sentiment. To see the most common terms in both positive and negative articles, we created word clouds for each sentiment label. Word clouds offer a simple method for identifying phrases and ideas that are often used in each sentiment respectively.

Figure 5: Sentiment Analysis Positive Bag of Words

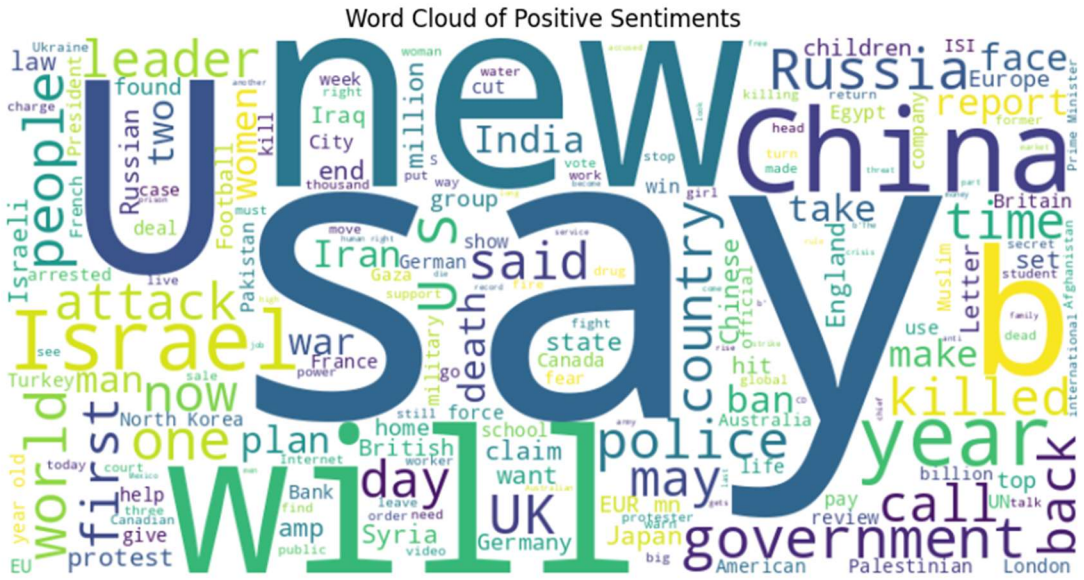
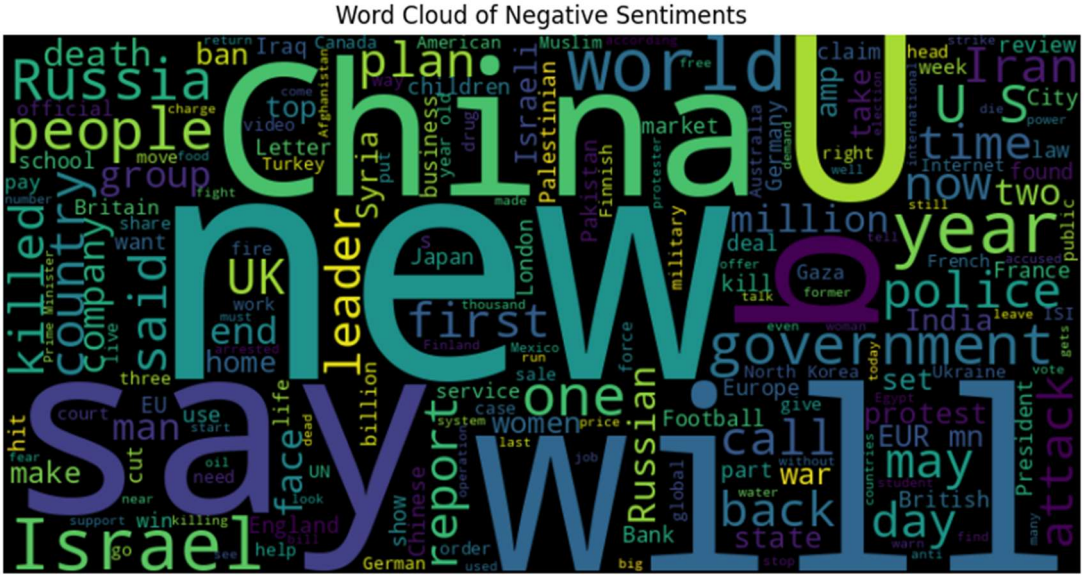


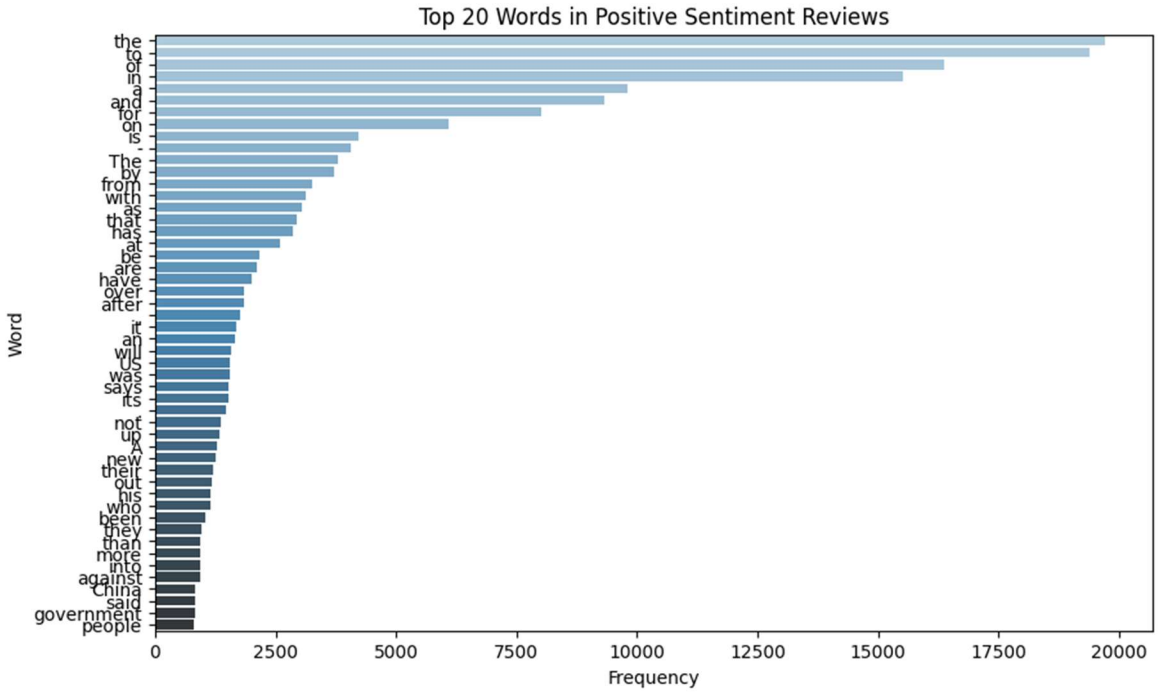
Figure 6: Sentiment Analysis Negative Bag of Words



The analysis of top words in each sentiment category was initially intended to identify words most strongly associated with positive and negative sentiments. However, this revealed substantial noise in the data, as many terms did not align with true financial sentiment but rather reflected neutral or irrelevant information. This observation led to further filtering to better align the data with the intended financial sentiment focus.

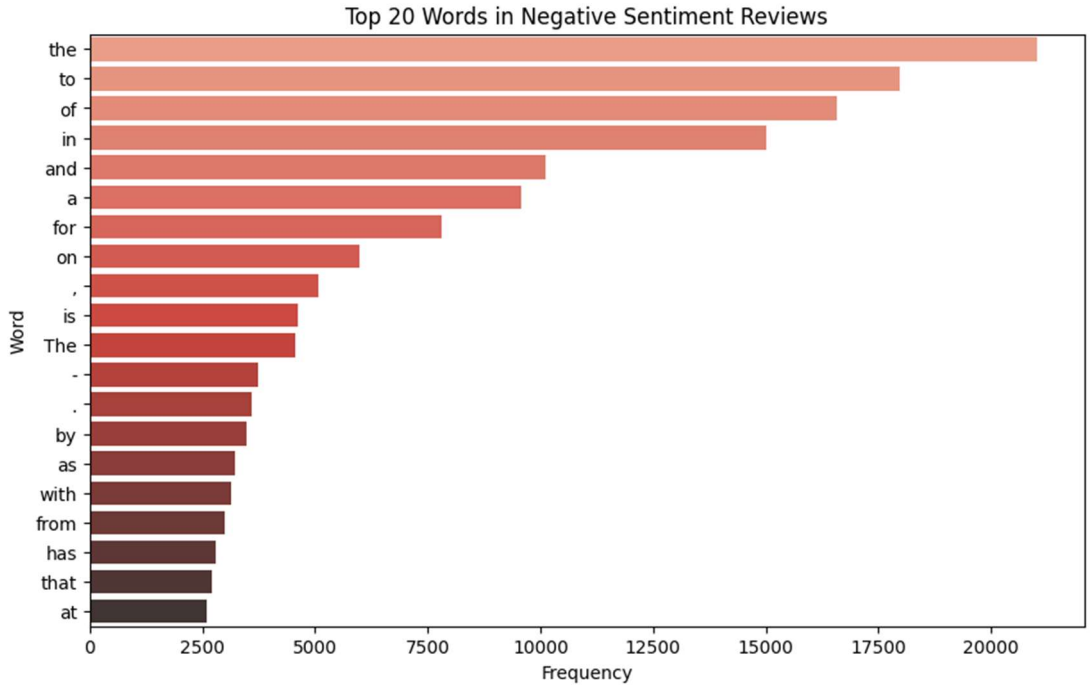
4. Positive and Negative Words Frequency Distribution

Figure 7: Sentiment Analysis Top 20 Positive Words Distribution



We generated this word distribution graph with our positive sentiments to gain a better understanding of which words occur the most.

Figure 8: Sentiment Analysis Top 20 Negative Words Distribution



We generated this word distribution graph with our negative sentiments to gain a better understanding of which words occur the most.

We performed a word distribution analysis and observed a high frequency of common filler words, such as "the," "a," and "and" which contributed little to the overall sentiment or meaning of the text. To enhance the clarity and focus of our analysis, we applied stop words to filter out these frequently occurring but semantically insignificant terms. This adjustment aims to improve the relevance of our feature set by retaining only words that carry meaningful sentiment indicators.

To analyze financial sentiment in stock market-related text data, we employed a model-building process that included train-test splitting, hyperparameter tuning with validation, model testing, and comprehensive evaluation using key performance metrics. Here's a summary of each step and how it aligns with the specified requirements.

Data splitting

Train-Test Split: The dataset was divided into training and testing sets, following a common 80-20 split. This ensures that the model learns on the training set while the test set remains unseen to provide an unbiased evaluation of the model's performance. Here's the code used to split the data:

Model Building and Training

Embedding Preparation: GloVe embeddings were chosen to create dense vector representations of the sentences, as GloVe is known to effectively capture semantic similarities between words. The get_glove_embedding function loads the embeddings from GloVe's pre-trained vectors and computes the mean embedding for each sentence, producing consistent 300-dimensional feature vectors.

Logistic Regression Model

Logistic Regression was selected as the primary model, with class_weight='balanced' applied to account for the slightly imbalanced nature of the dataset. This setup helped to avoid bias toward any specific class (especially the dominant neutral sentiment class).

Validation and Hyperparameter Tuning

Within the training set, cross-validation was used to fine-tune model parameters (e.g., regularization strength) and optimize the model's performance.

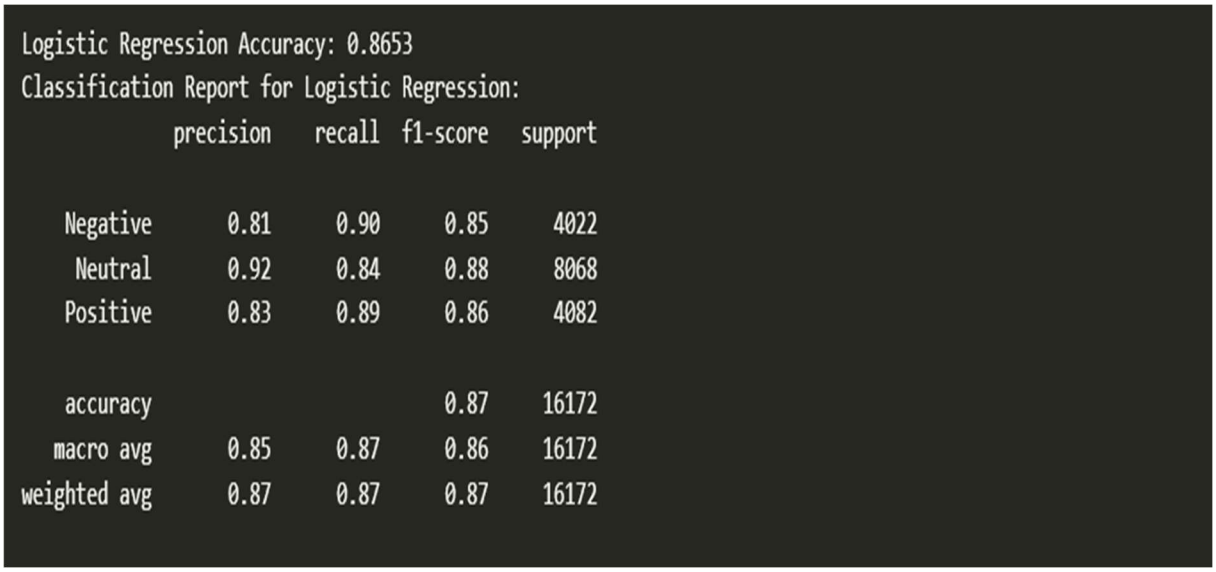
Results

Final Testing on the Test Set/ Metrics for Evaluation

After hyperparameter tuning, the model was evaluated on the test set to obtain a final performance measure.

The logistic regression model achieved an overall accuracy of 86.53%, demonstrating strong performance in sentiment classification. The precision, recall, and F1-scores across the sentiment classes indicate that the model performs consistently well in distinguishing between Positive, Negative, and Neutral sentiments. Specifically, the model achieved a high recall of 0.90 for the Negative class, effectively identifying most negative instances. For the Neutral class, which constitutes the largest portion of the dataset, the model showed a balanced performance with a precision of 0.92 and recall of 0.84. The Positive also class yielded strong scores, with an F1-score of 0.86.

Figure 9: Sentiment Analysis Logistic Regression Accuracy Report



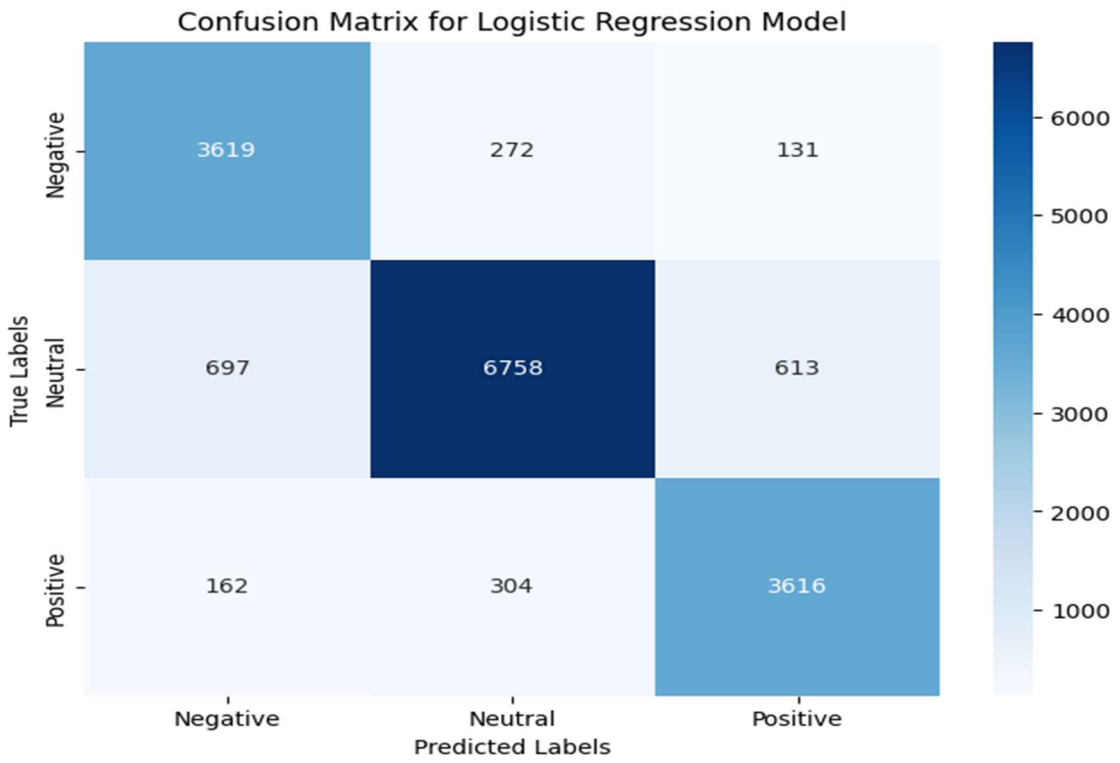
Logistic Regression Accuracy: 0.8653
Classification Report for Logistic Regression:

	precision	recall	f1-score	support
Negative	0.81	0.90	0.85	4022
Neutral	0.92	0.84	0.88	8068
Positive	0.83	0.89	0.86	4082
accuracy			0.87	16172
macro avg	0.85	0.87	0.86	16172
weighted avg	0.87	0.87	0.87	16172

Confusion Matrix

Overall, the matrix highlights a generally accurate model, with minor challenges in distinguishing Neutral from Positive sentiments.

Figure 10: Sentiment Analysis Confusion Matrix



Key observations

The model accurately classifies Neutral sentiments, as seen by the high value of correctly predicted Neutral instances (6,758), though 697 Neutral instances were misclassified as Negative and 613 as Positive.

For Negative sentiments, 272 instances were misclassified as Neutral, while 131 were incorrectly predicted as Positive. This indicates a reasonable but not perfect precision for the Negative class.

Positive sentiments show similar misclassification trends, with 304 instances predicted as Neutral and 162 as Negative, suggesting that some Positive sentiments are harder to distinguish from Neutral.

Discussion and Conclusion

The logistic regression model achieved substantial accuracy, reflecting its capability in distinguishing between positive, negative, and neutral sentiments. The balanced class weights helped improve performance across all classes, as shown by the classification report metrics.

This approach to model development, testing, and evaluation ensures a robust analysis of financial sentiment data, providing valuable insights into sentiment trends within stock market discussions.

The logistic regression model achieved an accuracy of 86.53%, performing well in distinguishing between positive, neutral, and negative sentiments. The classification report reveals strong performance across precision, recall, and F1-scores, particularly for the neutral class, which had high support in the dataset. Precision and recall for the negative and positive classes are also high, indicating the model's ability to capture both positive and negative sentiment in financial contexts effectively.

The logistic regression model proved to be the best-performing model based on its balanced accuracy and consistency across all classes. We initially experimented with a random forest model; however, it exhibited overfitting, leading to inflated precision and recall scores on the training data but suboptimal generalization on the test set. Consequently, logistic regression was selected as the final model due to its superior generalization capacity and robust performance on the test data.

If more time were available, we would extend the analysis by incorporating a time-series component. Timestamped real-time data would allow us to track sentiment shifts over time, providing insights into how news sentiment correlates with market movements and broader financial events. This additional layer of analysis would offer a dynamic understanding of sentiment trends, making the model more applicable to real-time financial forecasting and decision-making.

Contributions

Jana conducted the exploratory data analysis, preparing insights on sentiment distribution and data characteristics.

Iman led the model building and refinement process, optimizing logistic regression for the best accuracy.

Future Work

In future work, to improve model performance, we could incorporate more sophisticated embeddings like BERT or FinBERT, which have shown to capture nuanced language effectively, especially for sentiment analysis in finance. However, this approach requires significant computing power, which was a limitation in this project. To address computing constraints, leveraging cloud-based solutions could allow us to implement these advanced models without hardware limitations.

Integrating additional data sources, such as real-time market data, financial reports, or sentiment data from other social media platforms, could help contextualize sentiment changes

and make the model more robust. This would be especially valuable for identifying sentiment shifts in response to specific economic events. Applying our sentiment classification model to other domains, like sports analytics for basketball, could also provide insights into public sentiment around teams, players, or game outcomes based on social media posts and news articles, showcasing its versatility text.

Appendix

Appendix A: Mathematical Equations

Model	
Logistic Regression	$P(y=1\|X)=1/(1+e-(\beta 0+\beta 1X1+...+\beta nXn)1)$
Accuracy	TP+TN/(FP+FNTP+TN)
Precision	TP/(FP+TP)
Recall	TP/(FN+TP)
F1-Score	F1=2×((Precision×Recall)/ (Precision+Recall))

Appendix B: Raw Dataset

Sentiment	Text
0	According to Gran , the company has no plans to move all production to Russia , although that is where the company is growing
1	For the last quarter of 2010 , Componenta 's net sales doubled to EUR131m from EUR76m for the same period a year earlier , while it moved to a zero pre-tax profit from a pre-tax loss of EUR7m
1	In the third quarter of 2010 , net sales increased by 5.2 % to EUR 205.5 mn , and operating profit by 34.9 % to EUR 23.5 mn
1	Operating profit totalled EUR 21.1 mn , up from EUR 18.6 mn in 2007 , representing 9.7 % of net sales
1	Clothing retail chain Sepp+Ã±l+Ã± 's sales increased by 8 % to EUR 155.2 mn , and operating profit rose to EUR 31.1 mn from EUR 17.1 mn in 2004

References (Bibliography)

Abimbola, B., de La Cal Marin, E., & Tan, Q. (2024, April 19). *Enhancing legal sentiment analysis: A convolutional neural network–long short-term memory document-level model*. MDPI. https://www.mdpi.com/2504-4990/6/2/41

Georgios Fatouros a b, a, b, c, Highlights•Evaluating ChatGPT capability for sentiment analysis in the foreign exchange market.•Forex news dataset (2023, November 4). *Transforming sentiment analysis in the financial domain with chatgpt*. Machine Learning with Applications. https://www.sciencedirect.com/science/article/pii/S2666827023000610

Erasmus School of Economics. (2024, July 25). Erasmus University Rotterdam . file:///C:/Users/imans/Downloads/Evaluating%20Word%20Embedding%20Methods.pdf

Farimani, S. A., Jahan, M. V., & Milani Fard, A. (2022a). From text representation to financial market prediction: A literature review. *Information*, *13*(10), 466. https://doi.org/10.3390/info13100466

HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLOS ONE*, *15*(5). https://doi.org/10.1371/journal.pone.0232525

Kraayeveld, K. (2024). *Evaluating Word Embedding Methods for Sentiment Analysis*.

Lazzini, A., Lazzini, S., Balluchi, F., & Mazza, M. (2021). Emotions, moods and hyperreality: Social media and the stock market during the first phase of COVID-19 pandemic. *Accounting, Auditing & Accountability Journal*, *35*(1), 199–215. https://doi.org/10.1108/aaaj-08-2020-4786

Nguyen, T. H., & Shirai, K. (n.d.). *Topic modeling based sentiment analysis on social media for stock market prediction*. ACL Anthology. https://aclanthology.org/P15-1131/

Renault, T. (2019). Sentiment Analysis and machine learning in finance: A comparison of methods and models on one million messages. *Digital Finance*, *2*(1–2), 1–13. https://doi.org/10.1007/s42521-019-00014-x

Sahoo, C., Wankhade, M., & Singh, B. K. (2023). Sentiment analysis using Deep Learning Techniques: A Comprehensive Review. *International Journal of Multimedia Information Retrieval*, *12*(2). https://doi.org/10.1007/s13735-023-00308-2

Shah, P., Desai, K., Hada, M., Parikh, P., Champaneria, M., Panchal, D., Tanna, M., & Shah, M. (2024, January 26). *A comprehensive review on sentiment analysis of social/web media big data for stock market prediction - international journal of system assurance engineering and management*. SpringerLink. https://link.springer.com/article/10.1007/s13198-023-02214-6

Sharma, N. A., Ali, A. B. M. S., & Kabir, M. A. (2024, July 1). *A review of sentiment analysis: Tasks, applications, and Deep Learning Techniques - International Journal of Data Science and Analytics*. SpringerLink. https://link.springer.com/article/10.1007/s41060-024-00594-x

Smith, S., & O'Hare, A. (2022). Comparing traditional news and social media with stock price movements; which comes first, the news or the price change? *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00591-6

Symeonidis, S., Effrosynidis, D., & Arampatzis, A. (2018). A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, *110*, 298–310. https://doi.org/10.1016/j.eswa.2018.06.022

Valle-Cruz, D., Fernandez-Cortez, V., López-Chau, A., & Sandoval-Almazán, R. (2021). Does Twitter affect stock market decisions? financial sentiment analysis during pandemics: A Comparative Study of the H1N1 and the COVID-19 periods. *Cognitive Computation*, *14*(1), 372–387. https://doi.org/10.1007/s12559-021-09819-8

Wang, C. (2023, October 26). *Comparative Analysis of Embedding Techniques for Sentiment Analysis in Finance*. Erasmus University Rotterdam. file:///C:/Users/imans/Downloads/Comparative%20Analysis%20of%20Embedding%20Techniques.pdf

Woo, H., Kim, J., & Lee, W. (2020). Validation of text data preprocessing using a neural network model. *Mathematical Problems in Engineering*, *2020*, 1–9. https://doi.org/10.1155/2020/1958149

Wujec, M. (2021). Analysis of the financial information contained in the texts of current reports: A deep learning approach. *Journal of Risk and Financial Management*, *14*(12), 582. https://doi.org/10.3390/jrfm14120582

Zhang, C., Peng, B., Sun, X., Niu, Q., Liu, J., Chen, K., Li, M., Feng, P., Bi, Z., Liu, M., Zhang, Y., Fei, C., Yin, C. H., Yan, L. K., & Wang, T. (2024, November 6). *From word vectors to multimodal embeddings: Techniques, applications, and future directions for large language models*. arXiv.org. https://arxiv.org/abs/2411.05036

Zhang, L., Wang, S., & Liu, B. (2018). Deep Learning for Sentiment Analysis: A Survey. file:///C:/Users/imans/Downloads/Deep%20Learning%20for%20Sentiment%20Analysis.pdf