StockSight: Uncovering Market Sentiments

TABLE OF CONTENTS

CHAPTER 1
INTRODUCTION

1.1 Absrtact :

Using cutting-edge natural language processing and sentiment analysis tools, our study seeks to investigate and examine Reddit comments about stocks. To give investors and financial institutions useful insights, the objective is to identify patterns, movements in sentiment, and emerging market sentiments. Our goal is to classify sentiment as positive, negative, or neutral by gathering, preparing, and applying data from Reddit to sentiment analysis algorithms. Brand perception management, enhanced market sentiment monitoring, and investment decision-making can all benefit from these classifications. Ultimately, the goal of this initiative is to provide investors with useful information so they can make better decisions and lower their risk of losing money in the market. We aim to further our understanding of the dynamics of market sentiment in the digital era through this project.

1.2 Introduction :

Using advanced sentiment analysis and natural language processing techniques, our study aims to explore the world of stock market debates on Reddit. From these conversations, we hope to glean important insights and identify prevailing attitudes. Our objective is to provide investors with actionable information by utilizing state-of-the-art technology to identify significant patterns, shifts in sentiment, and emerging market sentiments. In the end, we hope to reduce the danger of possible losses in the market by using our assessments to help investors make better decisions. In the always changing world of finance, our initiative seeks to not only improve knowledge of the dynamics of market sentiment but also to help with better risk management and investment decision-making."

1.3 Motivation:

1.3.1 Help Investors Make Data-Driven Decisions:

- By giving investors access to current market sentiment data, they may make well-informed judgments based on real-time sentiment analysis.
- Investors can make more intelligent decisions by modifying their strategies—whether they involve purchasing, selling, or holding stocks—in response to the current attitude.
- By lowering uncertainty and raising the possibility of success in their investment activities, investors can traverse challenging market conditions with confidence when they have access to timely market mood insights.

1.3.2 Find Market Trends:

- Investors can see developing market trends before the general public does by keeping an eye on trends and sentiment shifts.
- Investors who receive early warnings about shifts in market sentiment have a competitive advantage and can take advantage of opportunities or reduce risks before others do.
- Investors can adjust their investing strategy proactively, optimizing returns and reducing possible losses, by staying ahead of market developments.

1.3.3 Minimize Risk:

- Understanding the sentiment-driven factors influencing market behavior allows investors to assess and manage risk more efficiently. This is made possible by gaining insight into the mood of the market.
- Investors should safeguard their investment assets by using risk-management methods and becoming informed of future market downturns and fluctuations in sentiment.
- Knowing the mood of the market lowers the probability of impulsive or emotionally motivated investing choices, resulting in more careful and logical risk management strategies.

1.3.4 Strengthen Market Surveillance:

- Financial institutions may more efficiently monitor market condition by employing real-time sentiment research, which also lets them modify their trading and risk management techniques in reaction to shifting market conditions.
- Financial institutions can take advantage of opportunities and reduce possible losses for their clients by staying ahead of market sentiment patterns with the use of enhanced market surveillance capabilities.

1.3.5 Give Retail Investors More Power:

- Giving retail investors easy access to market sentiment data helps them make better selections and overcome the competition in the financial markets.
- Retail investors can obtain insights earlier exclusive to institutional investors and level up the competition by expanding access to market sentiment data.
- Increased openness and fairness can be promoted by giving ordinary investors more influence in the market, which eventually results in a more effective and inclusive financial environment.

1.4 Objectives:

1.4.1 Creating prediction models:

- Forecasting future shifts in market sentiment and movements is made possible by predictive analytics models that analyze past sentiment data to find patterns and trends in investor sentiment.
- Investors can profit from opportunities or reduce risk factors before they arise by using predictive analytics to foresee changes in the market and take preventive steps to do so.

1. Making Informed Investment judgments:
    - Sentiment analysis of Reddit discussions offers investors insightful information about the state of the market, enabling them to make wise investment choices.
    - Investors can make more smart investment decisions by assessing the possible risks and opportunities related to various investment options when they have access to insightful information about market sentiment.

2. Brand monitoring:
    - Through the study of sentiment inside Reddit comments, businesses can learn how the public views their brands, goods, or services.
    - Through sentiment analysis, companies may pinpoint areas for development, handle client criticism skillfully, and take active steps to maintain their reputation, all of which boost their image and customer happiness.

1.5 Problem Definition:

Reddit's extensive stock market discussion threads offer a wealth of information about market mood and new trends. However, because of its bulk and complexity, it is difficult to extract relevant and useful information from this unstructured text data. Investors and financial institutions are left without real-time insights on market perceptions due to the lack of effective ways for assessing this data, which can impede informed decision-making and raise the risk of losses. With the use of natural language processing and sentiment analysis, this project seeks to provide a framework for processing, categorizing, and interpreting stock market-related Reddit discussions. The framework will be useful for risk management, market monitoring, and investment decision-making.

1.6 About Dataset:

We are manually extracting the most relevant stock-related posts and comments from Reddit's most popular discussions every week, using a specially designed Reddit bot. Using this method enables us to compile a premium dataset that reflects the views, patterns, and attitudes of users regarding the platform's stock market discussions. We make sure the data is representative and useful by concentrating on the most popular posts and comments, which gives our sentiment analysis a solid starting point. The project's capacity to glean meaningful insights and remain ahead of developing industry trends and sentiment shifts is improved by this focused approach to data collection.

CHAPTER 2
LITERATURE SURVEY

Sr. No.	Title - Year of Publish	Journal	Methodology	Conclusion
1.	Enhancing the Prediction of Stock Market Movement Using Neutrosophic-Logic-Based Sentiment Analysis	Journal of Theoretical and Applied Electronic Commerce Research, 2024	Proposed model demonstrated its advantage by utilizing the StockNet dataset benchmark and comparing it to models that use this dataset. The proposed model feeds the integrated SA scores with historical stock market data into an LSTM model to foresee the stock movement. Notably, our model distinguishes itself as the first to employ NL in the SA process to predict stock market movement.	The proposed model outperformed other models that utilized the same dataset by utilizing NL in the SA process to make the results more compatible with human sentiment and using the integration of historical stock market data with SA results as input factors to our prediction model using LSTM, which resulted in a relatively high accuracy, of around 78.48%, and an MCC score of 0.587.
2.	SOCIAL MEDIA AND THE STOCK MARKETS: AN EMERGING MARKET PERSPECTIVE	Journal of Business Economics and Management ,2021	his paper investigates relationship between the information on the Twitter and the Indian stock markets. This study extracts two different kind of information from the twitter: a) optimistic public sentiments, b) pessimistic public sentiments.he sentiment analysis of the Twitter messages is done using VADER – a parsimonious rule based, human validated sentiment analysis method.	VADER has been found to outperform individual human raters in assessing the sentiments of the tweets and also generalizes more favorably across contexts than any other tools i.e. LIWC, ANEW, the General Inquirer, SentiWordNet and other machine learning oriented approaches such as Naïve Bayes algorithm, SVM, etc.Its sentiment intensity scores, also known as valence scores have been rated by the human raters.

3.	Harmonizing Macro-Financial Factors and Twitter Sentiment Analysis in Forecasting Stock Market Trends	Journal of Computer Science and Technology Studies	In our endeavor to forecast stock market movements, we harnessed the power of social media discourse by leveraging a dataset encompassing ChatGPT-related tweets spanning January to March 2023. Alongside, historical stock price records of major tech entities were amalgamated for training and validation purposes.Utilizing a Gradient Boosting Classifier, we aimed to decipher the correlations between sentiment signals embedded within Twitter conversations about ChatGPT and subsequent shifts in stock prices.	Various classification models, including Random Forest, Decision Tree, Extra Trees Classifier, and Naive Bias Classifier, undergo rigorous evaluation on a dataset. The Random Forest Classifier, employing key hyperparameters like 100 estimators, a depth of 5, and a random state of 0, achieves a balance between complexity and generalization. Models like Random Forest showcase robust predictive abilities, displaying high accuracy, recall, and F1 scores, underscoring their effectiveness in discerning market trends.
4.	Evaluation of Tree-Based Ensemble Machine Learning Models in Predicting Stock Price Direction of Movement	University of Electronic Science and Technology of China (UESTC)	Tis study compared the effectiveness of tree-based ensemble ML models (Random Forest (RF), XGBoost Classifier (XG), Bagging Classifier (BC), AdaBoost Classifier (Ada), Extra Trees Classifier (ET), and Voting Classifier (VC)) in forecasting the direction of stock price movement. Eight different stock data from three stock exchanges (NYSE, NASDAQ, and NSE) are randomly collected and used for the study. Each data set is split into training and test set. Ten-fold cross validation accuracy is used to evaluate the ML models on the training set. In addition, the ML models are evaluated on the test set using accuracy, precision, recall, F1-score, specificity, and area under receiver operating characteristics curve (AUC-ROC).	The experimental results indicated that for the ten-fold cross validation accuracy of the training set, the AdaBoost model outperformed the other models. For the test data, only accuracy, precision, f1-score, and AUC metrics were able to generate results significant to rank the different models using Kendall W test of concordance. The Extra Tree model performed better than the rest of the models on the test data set.

CHAPTER 3
COMPREHENSIVE OVERVIEW

3.1 Data Scrapping - Reddit Bot:

The "IndianStockMarket" subreddit's data is extracted by our Python script through interaction with Reddit's API using the PRAW (Python Reddit API Wrapper) package. First, using the supplied login, password, client ID, and client secret, authenticate the Reddit instance. It next gets the top 25 posts from the previous week by accessing the subreddit. The title and content (self-text) of each post are gathered and kept in a list. In the end, the gathered information is arranged into a Pandas DataFrame and exported to a CSV file called "data.csv". The process of extracting insights and conducting additional analysis is made easier by this script, which efficiently automates the Reddit data scraping process relevant to talks about Indian stocks.

3.2 Setup.py:

The resources and tools needed for tasks involving natural language processing are set up by this script:
Setup for NLTK: Downloads the models for the Punkt tokenizer from NLTK. The Punkt tokenizer is a well-liked tool for tokenization, which is the process of dividing text into discrete words or phrases.
Installing SpaCy: Obtains the "en_core_web_sm" English language model from SpaCy.
Part-of-speech tagging, named entity recognition, dependency parsing, and other NLP tasks are made possible by this model, which comes with pre-trained word vectors and linguistic annotations for English text input.
In order to facilitate the seamless execution of NLP tasks in later code, the "setup" function makes sure that the necessary NLTK and SpaCy resources are downloaded and installed.

3.3 Scrape.py:

Using the given Reddit instance, this script scrapes the most popular posts from the given subreddit (by default, "IndianStockMarket"). Importing "reddit_instance" from "reddit_bot.py" is likely to import the authorized Reddit instance.
"Subreddit" and "time" are two optional arguments for the "Scrape" function. The top posts from the designated subreddit for the designated time frame (by default, "month") are retrieved. It loops over the most popular posts, publishes their titles and contents, compiles the information into a list, and then uses the "Save_csv" function to save it to a CSV file.
The gathered data is transformed into a Pandas DataFrame and saved to a CSV file called "data.csv" using the "Save_csv" function.

All in all, this script makes it easier to extract the most popular posts from Reddit and store them in a CSV file for further analysis.

3.4 App.py:

The "scrap.py" and "setup.py" modules are imported by this script, along with additional necessary libraries like Pandas, NumPy, Spacy, NLTK, and SpacyTextBlob. It defines a function called "extract_stock_names_spacy" that makes use of SpaCy's named entity recognition (NER) capabilities to extract stock names from a given title.

Next, the function "get_sentiment" is defined to examine sentiment in the data that was scraped. Using SpaCy, it first extracts stock names from titles. Next, it uses SpaCyTextBlob to compute sentiment polarity and subjectivity for both titles and contents. Finally, it adds these values to the DataFrame as new columns. After that, a CSV file called "output_file.csv" contains the saved DataFrame.

Lastly, the script runs the "Scrape" function from "scrap.py" to gather data, uses the "get_sentiment" function to do sentiment analysis, and stores the output in a CSV file. The primary function makes sure that the script is only run directly, not when it is imported as a module, and that only then does the sentiment analysis process take place.

CHAPTER 4
THEOROTICAL BACKGROUND

4.1 Methodologies Description :

4.1.1 Input Text Retrieval:

This study uses Reddit posts and user comments as its main source of input data. Utilizing specially designed web scraping bot, these postings and comments are gathered and saved in a structured way as CSV (Comma Separated Values) files. The CSV file is the basis for our sentiment research and next analysis of market sentiments and trends within the Reddit community. Each row in the file represents a single comment or post, complete with essential information like the title and content.

4.1.2 Text Preprocessing:

- Punctuation Removal:

The function remove_punctuation_and_lower() accepts a text input and eliminates any tokens that contain punctuation. After that, it gives back a list of tokens in lower case and without any punctuation.
By using this function on the DataFrame's "text" column, we can create a new column called "content_lower_nopunct" that contains the tokens for each text entry that have had the punctuation removed in lower case.

- Stopwords Removal:

The remove_stopwords() function eliminates any tokens that are stop words after processing a text input with spaCy. After that, a list of cleaned tokens free of stopwords is returned.
By using this method on the DataFrame's "text" column, we can create a new column called "content tokens stop" that holds the text for each entry after stopwords have been eliminated.

- Lemmatization:

A list of lemmatized tokens is returned by the lemmatize() function after it has lemmatized an input text using spaCy.
The lemmatized tokens for each text entry are contained in a new column called "content_lemmatized," which is created by applying this function to the DataFrame's "text" column.

- Sentence Tokenization:

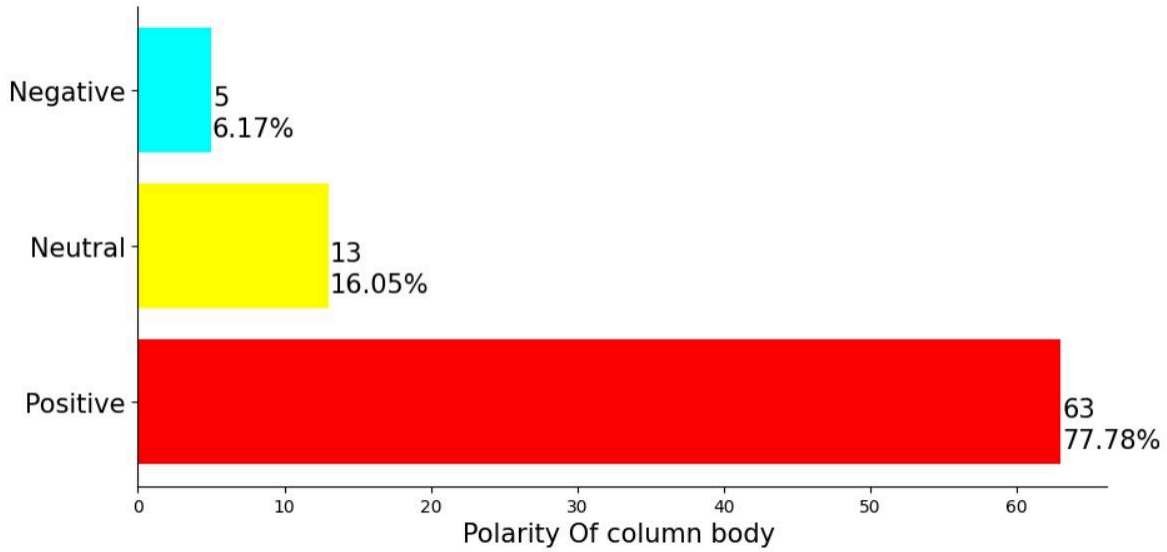
Using spaCy's sentence tokenization, the processed sentence is divided into individual words known as tokens.

4.1.3 Scoring Sentences:

By averaging the polarity and subjectivity scores of each individual token in a sentence, our method calculates sentiment scores for each sentence. To do this, we first tokenize each document using spaCy and then apply TextBlob's sentiment analysis function to each token. In particular, we use TextBlob and spaCy to process each sentence using lambda functions in order to obtain the subjectivity and polarity scores. We can get a comprehensive sentiment score that accurately reflects the overall sentiment indicated in the sentence by averaging these ratings across all the tokens in the sentence. With the help of this method, we can efficiently assess the sentiment of every sentence in the dataframe and extract significant information from the text data.

4.1.4 Displaying Results:

The sentiment of the statement is analyzed based on its scores to determine if it conveys a positive, negative, or neutral sentiment.



4.2 Flow of Work:

4.2.1 Gathering of Data:

Find relevant subreddits for stock conversations, including r/investing, r/stocks, or other subreddits that are specifically about stocks.
Use the Reddit API or other third-party Python tools, such as PRAW, to extract comments and posts that discuss stocks.

4.2.2 Data Preprocessing:

Eliminate stopwords, URLs, special characters, and punctuation from the text data.
Put the text in words or phrases to tokenize it.

4.2.3 Tokenization:

To start, the input text is divided into smaller chunks known as tokens. This can be done by dividing the text according to punctuation marks, other delimiters, or whitespace characters (spaces, tabs, newlines, etc.).

4.2.4 Sentiment Analysis:

Utilizing the preprocessed text data, apply the selected sentiment analysis technique to determine the sentiment polarity (positive, negative, or neutral) of every post or comment.

4.2.5 Aggregate Sentiment Scores:

Aggregate the sentiment scores of all posts or comments mentioning a stock to get an overall sentiment score for each stock.
This might involve applying more advanced aggregating techniques or averaging the sentiment scores of each individual.

- High subjectivity and positive polarity texts can convey strong feelings of positivity together with subjective judgments and emotive emotions.
- Low polarity and high subjectivity texts might imply the presence of subjective thoughts or feelings without explicitly positive or negative sentiments.
- Negative sentiment mixed with subjective opinions or emotional expressions may be indicated in a text with high subjectivity and negative polarity.

4.2.6 Visualization and Interpretation:

Plots or charts can be used to chart the sentiment scores over time and show trends or patterns.
To understand how the market feels about various stocks, interpret the sentiment analysis data.

4.2.7 Output:

Based on threshold levels for polarity and subjectivity scores, the emotion of the postings and comments on stocks is classified as positive, negative, or neutral.

CHAPTER 5
RESULTS AND DISCUSSIONS

5.1 Results:

Overall, our project's findings offer insightful information on the dynamics of market sentiment, enabling financial institutions and investors to make wise choices and successfully tackle the complexity of the stock market.

5.2 Summary:

In brief using natural language processing (NLP) tools, our sentiment research project explores stock market-related Reddit comments for valuable insights. We are able to identify patterns and shifts in market sentiment by examining the subjectivity and polarity of individual words. This allows investors to make well-informed judgments. We can reduce risks, find important aspects, and take advantage of market opportunities by using our analysis. With all factors taken into account, our research advances knowledge of the dynamics of market mood and makes it easier to navigate the world of stocks.

5.3 Conclusion:

Furthermore, our sentiment analysis experiment on stock market-related Reddit discussions has produced insightful information about the dynamics of market sentiment. We have successfully examined the subjectivity and polarity of discussions by applying natural language processing techniques, revealing trends, patterns, and significant factors influencing market sentiment. With the help of this analysis, financial institutions and investors may reduce risk, make well-informed decisions, and seize market opportunities. The project creates opportunities for future study and development, such as improving sentiment analysis algorithms and adding real-time sentiment monitoring features. In the end, our research advances our knowledge about market sentiment and improves our ability to make decisions in the constantly evolving stock market environment.

5.4 Scope for Future Work:

5.4.1 Real-Time Sentiment Analysis: Expand the project to enable investors and organizations to track market sentiment as it changes and responds to current events and market fluctuations by performing sentiment analysis on Reddit comments in real-time.

5.4.2 Adding More Data Resources: To give a more thorough examination of market sentiment, broaden the scope of the research by adding data from other sources like Twitter, websites that cover finance, and communities dedicated to the stock market.

5.4.3 Advanced Sentiment Analysis Techniques: To increase the precision and level of detail in sentiment analysis outcomes, explore sophisticated methods such as deep learning models and sentiment lexicons designed especially for financial discussions.

5.4.4 Integration with Trading Platforms: To give users real-time sentiment information and help them make better-informed trade choices, traders may incorporate sentiment analysis capabilities into investments and platforms for trading.

REFERENCES

1. Abdelfattah, B. A., Darwish, S. M., & Elkaffas, S. M. (2024). Enhancing the Prediction of Stock Market Movement Using Neutrosophic-Logic-Based Sentiment Analysis. *Journal of Theoretical and Applied Electronic Commerce Research*, *19*(1), 116-134.
2. Agarwal, S., Kumar, S., & Goel, U. (2021). Social media and the stock markets: an emerging market perspective. *Journal of Business Economics and Management*, *22*(6), 1614-1632.
3. Amin, M. S., Ayon, E. H., Ghosh, B. P., MD, M. S. C., Bhuiyan, M. S., Jewel, R. M., & Linkon, A. A. (2024). Harmonizing Macro-Financial Factors and Twitter Sentiment Analysis in Forecasting Stock Market Trends. *Journal of Computer Science and Technology Studies*, *6*(1), 58-67.
4. Ampomah, E. K., Qin, Z., & Nyame, G. (2020). Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement. *Information*, *11*(6), 332..
5. Seroyizhko, P., Zhexenova, Z., Shafiq, M. Z., Merizzi, F., Galassi, A., & Ruggeri, F. (2022, December). A sentiment and emotion annotated dataset for bitcoin price forecasting based on reddit posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)* (pp. 203-210).
6. Corbet, S., Hou, G., Hu, Y., & Oxley, L. (2021). We reddit in a forum: The influence of messaging boards on firm stability. *Available at SSRN 3776445*.
7. Loginova, E., Tsang, W. K., van Heijningen, G., Kerkhove, L. P., & Benoit, D. F. (2021). Forecasting directional bitcoin price returns using aspect-based sentiment analysis on online text data. *Machine Learning*, 1-24.
8. Akbik, A., Blythe, D., & Vollgraf, R. (2018, August). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th international conference on computational linguistics* (pp. 1638-1649).
9. Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). TweetEval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421*.
10. Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. " O'Reilly Media, Inc.".
11. Seroyizhko, P., Zhexenova, Z., Shafiq, M. Z., Merizzi, F., Galassi, A., & Ruggeri, F. (2022, December). A sentiment and emotion annotated dataset for bitcoin price forecasting based on reddit posts. In *Proceedings of the Fourth Workshop on Financial Technology and Natural Language Processing (FinNLP)* (pp. 203-210).

12. Galassi, A., Lippi, M., & Torroni, P. (2020). Attention in natural language processing. *IEEE transactions on neural networks and learning systems*, *32*(10), 4291-4308.
13. Gao, Y., Wang, R., & Zhou, E. (2021). Stock prediction based on optimized LSTM and GRU models. *Scientific Programming*, *2021*, 1-8.
14. Kraaijeveld, O., & De Smedt, J. (2020). The predictive power of public Twitter sentiment for forecasting cryptocurrency prices. *Journal of International Financial Markets, Institutions and Money*, *65*, 101188.
15. Kearney, C., & Liu, S. (2014). Textual sentiment in finance: A survey of methods and models. *International Review of Financial Analysis*, *33*, 171-185.