International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 10 Issue: 07 | July 2023 www.irjet.net p-ISSN: 2395-0072

Sentiment Analysis based Stock Forecast Application

Shivani R[1], Devaraju B M[2], Dr.Girijamma H A[3]

[1]*Post Graduate Student, Department of Computer Science and Engineering, RNSIT, Karnataka, India*
[2]*Assistant Professor, Department of Computer Science and Engineering, RNSIT, Karnataka, India*
[3]*Professor, Department of Computer Science and Engineering, RNSIT, Karnataka, India*

Abstract - *An abstract summarizes, in one paragraph (usually), the major aspects of the entire paper in the following prescribed sequence. In the contemporary era, as data continues to grow in volume and significance for businesses, manual data analysis is no longer feasible in the fast-paced world. Therefore, the adoption of artificial intelligence and data mining techniques has become imperative. Among various factors affecting stock price fluctuations, a crucial determinant is the gains or losses incurred by a company. As news is a primary source of information for most traders, it plays a pivotal role in forecasting changes in the stock market. This study focuses on sentiment classification and its influence on stock market prices. Sentiment analysis based stock prediction is a cutting-edge approach that leverages natural language processing and machine learning techniques to predict stock price movements using sentiment data from various sources. This study explores the use of sentiment analysis on financial news, social media, and other textual data to gauge market sentiment, which can serve as a leading indicator for stock price trends. There are several classifiers that could be used in sentiment analysis for stock prediction. The classifiers used in the proposed system include, Decision Tree, Logistic Regression, Naïve Bayes and LSTM.*

Key Words: **Machine Learning, Deep Learning, Sentiment analysis, LSTM, Naive Bayes, Sentiment, score, Logistic Regression, Decision Tree**

1.INTRODUCTION

Stock prediction is an intriguing field that aims to forecast the future movements of stock prices in financial markets. However, predicting stock prices accurately is challenging due to the intricacy and volatility of financial markets. Recent advancements in data science and artificial intelligence have provided new tools and approaches to improve prediction accuracy. These techniques involve the extraction and analysis of vast amounts of financial data, including historical stock prices, company news, economic indicators, and social media sentiment. Stock prediction using sentiment analysis of financial news has emerged as an innovative approach to gain insights into market trends. By harnessing natural language processing (NLP) techniques, this method involves analyzing textual data, such as news articles and social media posts, to gauge the sentiment associated with specific stocks or companies. The underlying premise is that the sentiment expressed in financial news can impact investor behavior, thereby influencing stock prices. Positive sentiment may drive increased buying activity, leading to price appreciation, while negative sentiment could prompt selling and subsequent price decline. Sentiment analysis algorithms employ machine learning and NLP models to classify the sentiment of textual data accurately. These models assess the sentiment polarity (positive, negative, or neutral) and the strength of sentiment expressed. The sentiment analysis results can be combined with other fundamental and technical indicators to generate predictive models that capture market sentiment as an additional input.

1.1 Problem Statement

In the current era, there is a challenge of accurately predicting stock prices in the financial market. The problem lies in the complex and volatile nature of the market, where stock prices are influenced by various factors, including economic indicators, company performance, and investor sentiment. The goal is to develop a reliable prediction model that can forecast future stock prices with a high degree of accuracy based on financial news. This model should leverage historical stock data, market trends, and relevant indicators to make informed predictions and assist investors in making strategic investment decisions. The aim is to overcome the inherent uncertainties and challenges associated with stock prediction, ultimately enabling users to optimize their portfolio management and maximize their investment returns.

2. LITERATURE SURVEY

This study utilizes natural language processing techniques and recurrent neural networks (RNN), specifically long short-term memory (LSTM) networks In the field of stock forecasting. The objective is to predict stock prices using textual data. The study leverages data from Finwiz, a financial information platform, focusing on stocks of companies such as Adidas AG, Continental AG, Deutsche Lufthansa AG, Henkel AG & Co., and Siemens AG. By employing RNN with LSTM, the model aims to capture temporal dependencies and patterns in the textual data, allowing for accurate stock price predictions. The study demonstrates an accuracy rate of 80% for Adidas AG, 70% for Continental AG and Deutsche Lufthansa AG, and 65% for

Henkel AG & CO. and Siemens AG. These findings highlight the potential of natural language processing and recurrent networks in forecasting stock prices based on textual information. The application of natural language processing and RNN in stock forecasting presents promising opportunities for investors and financial analysts. By integrating textual data analysis with traditional quantitative methods, a more comprehensive understanding of stock market dynamics can be achieved. This study paves the way for further advancements in utilizing natural language processing and recurrent networks to enhance stock forecasting accuracy and assist in making informed investment decisions [1].

This study focuses on the sentiment analysis of financial news utilizing the VADER (Valence Aware Dictionary and Sentiment Reasoner) sentiment analysis tool in NLTK (Natural Language Toolkit). The objective is to analyze the sentiment expressed in financial news data obtained from Finwiz. The study reports an accuracy rate of 70% for sentiment analysis using this approach. The VADER sentiment analysis tool, coupled with NLTK, allows for the classification of financial news into positive, negative, or neutral sentiment categories. These findings demonstrate the potential of sentiment analysis techniques in extracting valuable insights from financial news data. By understanding sentiment trends in financial news, investors and financial analysts can gain a deeper understanding of market dynamics and make more informed decisions [2].

In this study, the focus is on predicting stock prices using sentimental analysis through Twitter data. The researchers employed various machine learning algorithms, including Random Forest, LSTM-RNN, CNN, and MLP. The data was obtained using the Twitter API. The accuracy results showed promising outcomes, with MLP achieving an accuracy of 0.82 and 0.84, CNN with two layers achieving an accuracy of 0.83, RNN with two layers achieving an accuracy of 0.77, and Wavelet CNN achieves an accuracy of 0.85. These results underscore the potential of sentiment analysis on Twitter data in predicting stock prices and highlight the efficacy of different machine learning algorithms for this task. This research contributes valuable insights into the realm of stock market prediction and paves the way for further advancements in sentiment analysis and machine learning techniques in this domain [3].

In this study, sentiment analysis is employed to explore the connection between subjective expression in financial reports and company performance. The objective is to analyze how the sentiment conveyed in these reports can potentially impact the success of a company. Regression analysis is utilized as a statistical method to examine the relationship between subjective expression and performance. The study incorporates data from Fortune, a renowned business magazine, and Xueqiu, a prominent financial platform. With an accuracy rate of 78.3%, the sentiment analysis model demonstrates its effectiveness in capturing the subjective nature of financial reports and its potential implications for company outcomes. This research contributes valuable insights into the interplay between sentiment analysis, financial reports, and company performance, providing a foundation for further investigations in this field[4].

This study focuses on conducting sentiment analysis at the entity level in financial texts, utilizing a pre-trained language model to enhance the granularity of financial information. The effectiveness of sentiment analysis for financial texts is demonstrated using bidirectional encoder representations of transformers, which leverage a pre-trained language model. The results of these experiments reveal that a minimum accuracy of 93% can be achieved [5].

3. PROPOSED SYSTEM

In the proposed system two datasets are used in order to train the sentiment analysis based stock prediction model with precision. The first one is obtained from Kaggle i.e fnews_data and the other is a real-time dataset that is obtained from Finviz which is a stock screening, stock research, and stock market financial visualization software and yfinance, Python library that provides a simple interface to access financial data from Yahoo Finance. The dataset consists of two modules.

A) Preprocessing of the data

The preprocessing of data involves several crucial steps to ensure data quality and prepare it for further analysis. These steps include handling missing values through imputation or removal, addressing outliers and noise, standardizing or normalizing numerical features to a common scale, encoding categorical variables, and performing feature selection or dimensionality reduction to reduce complexity and improve model performance. Additionally, data preprocessing may involve handling duplicates, checking for data integrity, and partitioning data into training and testing sets. These preprocessing steps are essential for enhancing the accuracy and reliability of subsequent data analysis and machine learning tasks.

B) Extraction and selection of features

Feature extraction and selection are crucial steps in data preprocessing that aim to identify and retain the most relevant and informative features for subsequent analysis. Feature extraction involves transforming raw data into a reduced set of representative features using techniques such as principal component analysis or feature engineering methods. This process reduces dimensionality and captures the underlying patterns in the data. Feature selection, on the other hand, focuses on identifying and selecting the subset of features that contribute the most to the predictive power of a model, using methods like correlation analysis, recursive feature elimination (RFE), or information gain. These

techniques aid in improving model efficiency, reducing overfitting, and enhancing the interpretability of results.

4. SYSTEM ARCHITECTURE

The dataset is first loaded into the model. The fnews_data dataset obtained from kaggle is used for the experiment. The fnews_data was chosen because it has significant amount of redundant entries in both the datasets. The dataset consists of financial news from multiple companies. It has over 80794 records, of which the model was trained with 64635 records and tested with 16158 records.

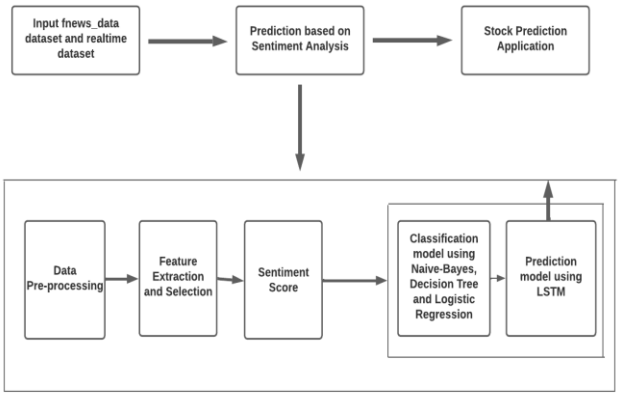

Fig – 1: System Architecture for Stock Forecast Application

Once after the dataset is loaded into the model the next step is data pre-processing where the model is been processed to check if there are any outliers. Outliers are the points that are different from the other data present in the dataset. If the outliers are present they should be eliminated. The next step in stock prediction is feature selection. There are 41 features present in the finviz dataset, of which 5 features are chosen. Once the features are chosen, the next step is to categorize the financial news as positive, negative or neutral using machine learning and deep learning algorithms and check for its accuracy, precision, recall and F-measure. The algorithm that has the highest accuracy is chosen and is applied on the realtime finviz dataset to obtain the sentiment score. The sentiment score obtained is combined with yfinance dataset to predict the stock prices using LSTM.

5. METHODOLOGY

The algorithms used includes Naive Bayes, Logistic regression and decision tree to obtain the sentiment score for the repository dataset from kaggle. The steps involved in these algorithms are stated.

Logistic regression

Step 1: Gather the dataset with input features and labels.

Step 2: Divide the dataset into a training set and a test set.

Step 3: Set initial weights and bias for the logistic regression model.

Step 4: Use the sigmoid function to compute predicted probabilities.

Step 5: Determine the cost between predicted probabilities and actual labels.

Step 6: Adjust weights and bias to minimize the cost using gradient descent.

Step 7: Iterate through the training set to update parameters and reduce the cost.

Step 8: Use the test set to assess the model's performance.

Step 9: Optimize model settings for better results.

Step 10: Make predictions on new data using the trained model.

Decision Tree

Step 1: Collect the dataset containing features (attributes) and corresponding target labels.

Step 2: Divide the dataset into a training set and a test set.

Step 3: Start building the decision tree by selecting the best attribute to split the data.

Step 4: Create nodes representing decisions based on the selected attributes.

Step 5: Split the data into subsets at each node based on the chosen attribute.

Step 6: Repeat the splitting process for each subset to grow the decision tree.

Step 7: Set criteria for stopping the tree growth, such as reaching a maximum depth or a minimum number of samples per node.

Step 8: Assign target labels to the leaf nodes based on the majority class in each leaf.

Step 9: Perform pruning to reduce the size of the tree and prevent overfitting.

Step 10: Use the test set to evaluate the performance of the decision tree.

Step 11: Fine-tune parameters like maximum depth or splitting criteria to optimize the decision tree's performance.

Step 12: Apply the trained decision tree to make predictions on new data.

Naïve bayes

Step 1: Gather the dataset containing features and corresponding class labels.

Step 2: Divide the dataset into a training set and a test set.

Step 3: Choose the appropriate Naive Bayes variant based on the data type:

Step 4: Calculate the prior probability of each class in the training set.

Step 5: Estimate the likelihood of each feature for each class:

Step 6: Assume independence among features given the class.

Step 7: Apply Bayes' theorem to compute the posterior probability of each class given the input features.

Step 8: For classification, choose the class with the highest posterior probability as the predicted class.

Step 9: Use the training set to calculate the probabilities and parameters required for classification.

Step 10: Assess the Naive Bayes classifier's performance using the test set.

Step 11: Apply Laplace smoothing to handle zero probabilities and improve generalization.

Step 12: Fine-tune any relevant hyperparameters, such as smoothing factor or feature selection, for better performance.

Step 13: Use the trained Naive Bayes classifier to make predictions on new data.

Table -1: Accuracy table for Kaggle dataset

Algorithm	Accuracy	Recall	F-Score	Precision
Naïve Bayes	73 %	73 %	73 %	75%.
Decision Tree	61 %	61 %	61%	61%
Logistic Regression	66 %	66 %	66%	67%

As seen in the Table -1 above, Naïve Bayes gives Accuracy of 73.23%, Recall of 73.23%, F-score of 73.03%, and Precision of 75.93%. Decision Tree gives Accuracy of 61.08%, Recall of 61.08%, F-score of 61.10% and Precision of 61.64%. Logistic Regression gives Accuracy of 66.46% , Recall of 66.46%, F-score of 66.25%, and Precision of 67.18%

Upon comparing the performances of the three algorithms, Naïve Bayes achieved the highest accuracy, recall, F-score, and precision among the three. Decision Tree exhibited moderate performance, while Logistic Regression fell in between Naïve Bayes and Decision Tree in terms of the evaluation metrics

The pie chart as shown in Chart -1 visually represents the distribution of classes in FNews_dataset on using Naïve Bayes. Each slice of the pie represents a class, and the percentage inside each slice indicates the proportion of that class in the dataset. The different classes in the dataset based on sentiment analysis includes positive, negative and neutral.

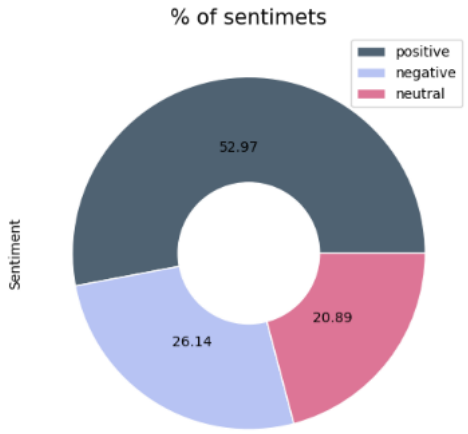

Chart -1: Sentiment distribution in Fnews_dataset

The Naïve Bayes algorithm has the highest accuracy thus it is chosen and is applied on the realtime finviz dataset to obtain the sentiment score. The sentiment score obtained is combined with yfinance dataset to predict the stock prices using LSTM.

6. CONCLUSION AND FUTURE WORK

Sentiment analysis-based stock prediction is a state-of-the-art method that utilizes natural language processing and machine learning algorithms to forecast stock price movements based on sentiment data. The sentiment analysis models are trained on datasets obtained from two primary sources: yfinance, which provides historical stock price data and real-time market information from Yahoo Finance, and Finviz, a financial visualization platform offering market insights and visualizations for informed decision-making. By integrating these datasets and algorithms, investors can gain

valuable insights into market sentiment, correlate it with stock price data, and make more informed investment decisions. The visualization capabilities of Finviz further enhance the understanding of market trends, technical indicators, and fundamental metrics, providing a comprehensive approach to sentiment-driven stock prediction. In future work, the exploration of different datasets from various financial markets and economic conditions can provide valuable insights into the model's adaptability and generalization capabilities. Investigating a wide range of algorithms, including support vector machines, deep reinforcement learning, or hybrid models, will enable a comprehensive evaluation of their strengths and weaknesses for sentiment analysis based stock prediction. An opportunity to assess the model's sustainability and performance in long-term investment scenarios can be presented by extending the prediction horizon to longer time ranges, such as weekly, monthly, or yearly intervals. This analysis could uncover potential challenges and trends that might not be evident in shorter time frames.

REFERENCES

[1] Dileep Kumar, "Stock Forecasting Using Natural Language and Recurrent Network", International Conference on Emerging Technologies in Computer Engineering: Machine Learning and Internet of Things, February 2020, 07-08.

[2] Arul Agarwal,"Sentiment Analysis of Financial News", International Conference on Computational Intelligence and Communication Networks, June 2021.

[3] Niveditha N Reddy, Naresh E, Vijaya Kumar B P, "Predicting Stock Price using Sentimental analysis through Twitter data", Institute of Electrical and Electronics Engineers, September 2020.

[4] Ni Zhong, JunBao Ren, "Using sentiment analysis to study the relationship between subjective expression in financial reports and company performance", International Conference on Frontiers in Psychology, July 2022.

[5] Zhihong Huang, Zhijian Fang, "An Entity-Level Sentiment Analysis of Financial Text Based on Pre-Trained Language Model", IEEE 18th International Conference on Industrial Informatics, 2020.

[6] Rubi Gupta, Min Chen, "Sentiment Analysis for Stock Price Prediction", IEEE Conference on Multimedia Information Processing and Retrieval, 2020.

[7] Ethan Seals, Steven R. Price, " Preliminary Investigation in the use of Sentiment Analysis in Prediction of Stock Forecasting using Machine Learning", IEEE Southeast Conference, 2020.

[8] He Wang, Zhiqiang Guo, "Financial Forecasting based on LSTM and Text Emotional Features", IEEE 8th Joint International Information Technology and Artificial Intelligence Conference, 2019.

[9] Saloni Mohan, Sahitya Mullapudi, Sudheer Sammeta, Parag Vijayvergia, David C. Anastasiu, "*Stock Price Prediction Using News Sentiment Analysis*", IEEE Fifth International Conference on Big Data Computing Service and Applications, 2019.