Algorithmic Trading and Sentiment Analysis in Indian Stock Market

Smita Satish Patil[1], Pramod Kubsad[2], Savitha Kulkarni[1]

[1]KLS Institute of Management Education and Research, Belagavi
[2]M S Ramaiah University of Applied Sciences, Bengaluru

Abstract. The rapid growth of social networks has produced an unprecedented amount of user-generated data, which provides an excellent opportunity for text mining. Sentiment analysis, an important part of text mining, DWWHPSWV WR RHWHUPLQH WKH RSLQLRQ RU HPRWLRQ H[SUHVVHG WKURXJK LWV FRQWHQW DQG VWUXFWXUH. Such information is particularly valuable for determining the overall opinion of a large number of people. Examples of its usefulness are predicting box office sales or stock prices. One of the most accessible sources of user-generated data is Twitter, which makes the majority of its user data freely available through its data access API. This study, will predict a sentiment value for stock-related tweets on Twitter, and demonstrate a correlation between this VHQWLPHQW DQG WKH PRYHPHQW RI WKH VWRFN PDUNHW. We Rich Data Streaming Environment. This study data UDQJHV from the period 2013 to 2024. The study reveals that the percentage of error which is less than 5% on almost all companies except one. Where it tells that if the percentage of Error is less than 5 then the accuracy is high and the predicted prices are more accu-

1. Introduction

Financial Markets are important indicators for the global economy as their stability is paramount for a country's economic wizardry. These are markets that stock prices, based on supply- and demand-based forces, drive them asinine. Characterizing market behaviour and predicting stock performance is a fundamental practice for an investor but a complex task as it involves many influencing factors. Stock prices are influenced by a range of factors beyond just the performance of the company, including market conditions, legislation, and media coverage which makes forecasting both incredibly difficult but also highly desired.

Sentiment analysis, a specialization in text mining that seeks to detect subjective opinions and attitudes as written by users within a large dataset, offers the potential resolution of this complexity. The emergence of social media platforms, such as Twitter in recent years has changed the whole game and provided a new platform to get instant feelings of the masses, which one cannot even possibly obtain by conducting public surveys. With this, unlike traditional ways that use where one can take a long time to gather public opinion of the people and effort, social media allows easy access to large amounts of user-generated content in real-time. It is a treasure trove of information to figure out what the general public thinks: about stock prices, choices as consumers, or changes in political trends.

This functionality of sentiment analysis has now matured considerably from its early form. While the idea of surveying public opinion goes back to the 1930s, it wasn't until more recent uses of new-age computational tech such as Natural Language Processing (NLP) and machine learning that we were able to perform sentiment analysis and make use of this data. In the days before getting real-time anything, prices of stocks would move and get interpreted very slowly – often times leaving companies out in the open with more rumors and misinformation than not. In the world today, thanks to digital communication and the internet, maintaining a constant flow of information has never been so easy, providing opportunities for companies to speedily react to any development in its market as well as allowing investors an instantaneous tool for assessing trends or simply tracking currencies. Twitter is really helpful for keeping count of real-time indicators of how the public feels. Twitter's API can be used by researchers to collect big data sets containing tweets of over 500 million that further may help them in analyzing the sentiment about specific companies or industries. It helps to capture the real-time micro-expressions of opinion enabling it to understand how public sentiment may drive stock prices. Sentiment combined with social media data increases the potential of sentiment analysis by giving traders and investors a tool that is not set in stone, but pliable to adapt to new market conditions.

Sentiment Analysis in Algorithmic Trading High-speed algorithmic trading is the use of computer programs to buy or sell orders at very speeds determined by a large number of inputs from historical prices, trend forecasts and technical volatility information. Now, with sentiment analysis integrated into these algorithms, traders can optimize their strategies to now include real time market sentiment. This extra piece of information can then allow better decision making which can lead to a more profitable trading result.

The increasing significance of sentiment analysis in stock market predictions is being affirmed by the latest research. Several studies have established that there is often a close tie between the sentiment of news articles and stock price movements: when the news is good, stocks go up; when the news is bad, stocks go down. Extending this analysis to Twitter, researchers have found that the social media site's user base, at least the portion of it that tweets in English, makes a pretty darn good stock market predicting machine. They use it to tweet a sentiment that is at times overwhelmingly positive, and at other times crushingly negative, and they store those sentiments in a database (hopefully the database's servers have enough RAM to handle all those good and bad feelings). By now, most of us know that participants in real-time public sentiment have a strong Twitter presence. What we don't know, for sure, is whether or not that presence is actionable for investors and financial institutions.

What is Sentiment Analysis?

Sentiment Analysis is a type of data mining that PHDVXUHV WKH LQFOLQDWLRQ RI SHRSOH¶V RSLQLRQV WKURXJK Natural Language Processing (NLP) and machine Learning. Further, WKH SHRSOH¶V RSLQLRQV DUH FRQWHUWHG into polarity, which consists of Negative, Neutral, and Positive polarity and can be used for analysis purposes like predicting price, movie ratings, etc.

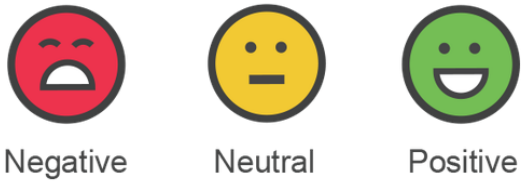


Figure 1. Sentiment Analysis (Polarity)

2. Literature Survey

A prediction regarding the trend of a stock price is considered to be one of the most relevant and exciting areas of research. Zhang, J. et al. (2018) novel data driven stock price trend prediction system. Expert Systems with Applications, 97, 60-69. This field remains of special interest for mainly the reason that the rewards in terms of financial feasibility are kept alive while maintaining rewards through predictive models. Investors, traders, and financial institutions rely on predictive models to maximize returns while managing risk. The recent development in computing techniques, namely Machine Learning (ML), Natural Language Processing (NLP), and sentiment analysis, has made it easy to develop advanced models to forecast future trends not only in the case of individual stock prices but also in the entire market indices. What stands out in this field is that not only do we use numerical data, like historical price and technical indicators but also textual data from news articles, and social media, among others precisely the sources of public sentiment, which have been found to have significant effects on stock movements. Review of several leading studies and approaches that enabled the development of sentiment analysis-based models for stock prediction and some common classification methods used in this field.

Nagar and Hahsler (2012) proposed an automated text mining approach to aggregate news stories from various sources and then construct a comprehensive news corpus. The technique filtered relevant sentences and analyzed those with the help of NLP techniques (Balahur, D., et al. (2010, September)). One of the important contributions from their work was the introduction of the notion of a sentiment metric called News Sentiment, where they were able to quantify the sentiment polarity of news articles as positive or negative. The measure obtained was based on the count of positive and negative words in the corpus. The authors concluded in that the variation in News Sentiment correlated well with factual changes in stock prices over time; in other words, the extracted sentiment from news can be a good indicator of the future behavior of the stock price. The open-source news aggregation tools that collect, aggregate, and evaluate the sentiment of news point toward the potential availability of scalable systems that can predict stocks in real time.

Building on this foundation, several studies have probed into the relationship that exists between sentiment derived from text data and stock price movements. (Pagolu, V. S., et al. (2016, October)). For example, Bollen et al. shared how the public mood is directly related to a large-scale social media platform like Twitter-impacted stock prices. Using machine learning to classify mood states such as calm, alert, happy, or anxious, they were able to show changes in public mood that "mapped closely with the direction of movement of the Dow Jones Industrial Average. It is quite an interesting work where, the predictability power of social media sentiment can provide very timely insights into investor psychology as well as market trends, especially in financial markets where real-time updates are necessitated.
Technical indicators are a very important component of stock forecasting research (Oriani, F. B., and Coelho, G. P. (2016, December)). Technical are statistical calculations that are based on the history of price, volume, and open interest and which attempt to project the near future price movements. Goh et al. integrated technical indicators with sentiment analysis to enhance

the predictive capability of their model. Their model would integrate technical indicators and NLP-derived sentiment metrics that account for the dynamics in the market and the investor sentiment simultaneously.

In addition to the related news and media, other fields of research interest are financial reports, earnings announcements, and any other form of corporate disclosure. To classify the transcripts of earnings calls into positive, negative, or neutral content, Li et al. (2014) (Chen, Y., et al. (2023)). adopted a sentiment analysis approach. From the study, sentiment scores from the earnings calls may be used in predicting post-announcement stock price movements. Thus, all such studies refer to a huge requirement of combining structured financial data with unstructured textual data so as to get more accurate results of stock predictions. The classification techniques formed the cornerstone of the stock prediction models, especially in sentiment analysis. Some of the simple traditional methods used included Naïve Bayes and Support Vector Machines (SVM) to classify the sentiments as positive, negative, or neutral. However, with the advent of deep learning, more sophisticated models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory networks have come into view to model sequential text data. It has really worked well in capturing temporal dependencies in news articles and tweets among others. For example, Chen et al. (2018) used LSTM networks to predict the stock prices of a company based on sentiment derived from financial news. Through that, the model captured better the temporal dependencies that had been established between consecutively appearing news articles and their effects on the stock prices and resulted in better performance predictions than traditional classification methods.

These include the pre-trained model of Transformer recently, such as BERT (Bidirectional Encoder Representations from Transformers), GPT (Generative Pre-trained Transformers), and others (Toprak, M. O., Bas, A., and van Heerden, I. (2021)). They have been widely applied to sentiment analysis tasks. Their capability for fine-grained sentiment nuances makes better performance available on tasks such as financial text classification. For example, Yang et al. (2020) used BERT for sentiment analysis of financial news. According to the results, the contextual understanding capability of the model greatly improved the accuracy of the stock price prediction.

Another stream of research is the ensemble methods (Opitz, D., and Maclin, R. (1999)) which make an effort to combine several classifiers to enhance the performance of the prediction. Ensemble methods, including Random Forests and Gradient Boosting Machines (GBMs), composite outputs of individual models into a stronger or more reliable predictive model. Kogan et al. developed an ensemble system that uses a multitude of sentiment classifiers for predicting stock movements based on financial blogs, achieving above-present single models.

In short words, the integration of sentiment analysis with the machine learning approach is a promising one for forecasting stock prices. Researchers have shown that extracting the sentiment from news articles, social media, and financial reports improves the aptness for predicting stock prices. Enhanced classification techniques using LSTM models, Transformer-based architectures, improve the performance of the stock forecasting system based on sentiments. As these techniques continue to improve, much promises lie ahead for stock price prediction in the future. Of course, this partly depends on the increased availability of data sources and continually improved algorithms.

3. Methodology

3.1 System Design

Following System design is proposed in this project to classify Twitter Tweets for Predicting historical price

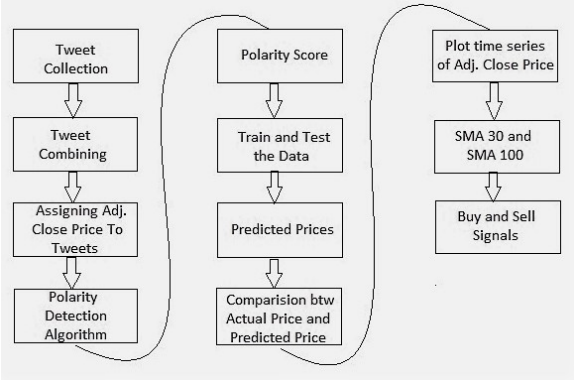


Figure 2. Research Process

3.1.1 Data Collection

I have taken 8 companies for research purposes they are Asian Paint, Cipla, HDFC Bank, Indian Oil Corporation (IOC), Maruti Corporation, Oil and Natural Gas Corporation (ONGC), State Bank of India (SBI) and Tata Motors Corporation. Collected the historical data from 05-06-2015 to 05-06-2020. The data contains Open, High, Low, Close, Adj. Close, and Volume. We considered Adj. Close Price as stock price. We used Twitter API to collect the Tweets of specified companies.

3.1.2 Pre-Processing of Datasets

The tweets that are collected from Twitter will be in the form of raw data, which is obtained using Python libraries (Packages for Twitter API). The Twitter API allows to collect the tweets in two ways: Sample stream and filter stream. SampleStream provides only a small number of random samples of tweets that are in real time and FilterStream provides tweets that match the same criteria.

Tweets are searched using three criteria:

- Specific keywords to search the tweets.
- User Twitter account User ID is used for tweet extraction.
- Tweets searched related to the keyword specified.

3.1.3 Human Labelling

For human labelling purposes, we labelled the tweets and categorized them into three classes according to the sentiment of the Tweets. The three classes are:

- Positive: If the user expresses his opinion with the attitude of positive/happy in the tweet is considered a Positive Tweet. Also, if the tweet contains more than two sentiments and has more positive sentiments in it.
- Negative: If the user expresses his opinion with the attitude of sad/negative in the tweet is considered as a Negative Tweet. Also, if the tweet contains more than two sentiments and has more negative sentiment in it.
- Neutral: The tweeter user expresses the tweet with no sentiment in it but transmits information about an advertisement of the product.

3.1.4 Feature Extraction

Previously processed dataset contain different properties in it. In this Feature Extraction method, we extract different aspects from the dataset. Then the aspect is used to calculate the polarity in sentences that are useful for determining individual opinions on particular asset, companies etc.

To identify key features from text or document we need techniques like machine learning techniques. The machine learning technique is used to classify the opinion of the user. Some of the identifying opinion techniques are:

- Tokenization: In this process, we divide the text into words, ³ D Q G ´ ³ W K H ´ ³ K H ´ ³ V K H ´ ([They are separated by the characters like space and punctuation. This is done to see tokens as individual components that make up tweets.
- Stop-Word Removal: Stop words are the common word that does not have any additional information used in the text formation and it is said to be useless or unimportant.
- Punctuation, numbers, and tagged people will be removed.
- Tweet is converted into lowercase which is more convenient for comparison.
- URLs will be removed (eg. WWW.apple.com).

3.1.5 Train and Test the Data

In sentiment analysis train and test both are needful methods. Here the train of the data makes the data recognize its pattern and cross validates the data to get more accuracy using machine learning techniques. Testing the data is useful to see how well the machine learning technique can predict new answers based on training.

3.1.6 Classification

Classification is a type of process in which the data is divided into several types of classes, which correspond to common patterns found in different classes. The main goal of this process is to create a classifier that classifies the tweets into the following sentiments, they are Positive, Negative, and Neutral classes.

Generally, in this classification process, we have followed two step approach. Objectivity. The first classifications are made in relation to the classification of tweets as objective or subjective. After the first step, we conduct the polarity to check whether the tweet is Positive, Negative, and Neutral.

3.2 Conceptual Model

Problem Statement: To demonstrate how social media content can be used to predict real-world outcomes. In particular, we use Tweets from Tweeter API's to predict the Stock Price using Sentiment analysis and to generate the buy and sell value of stock using Algorithmic Trading.

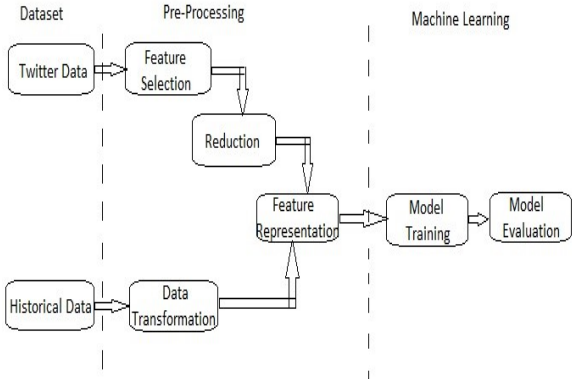


Figure 3.Conceptual Model framework

4 Evaluation

4.1 Graphical Representation of Actual Price Vs Predicted Price

- Asian Paint

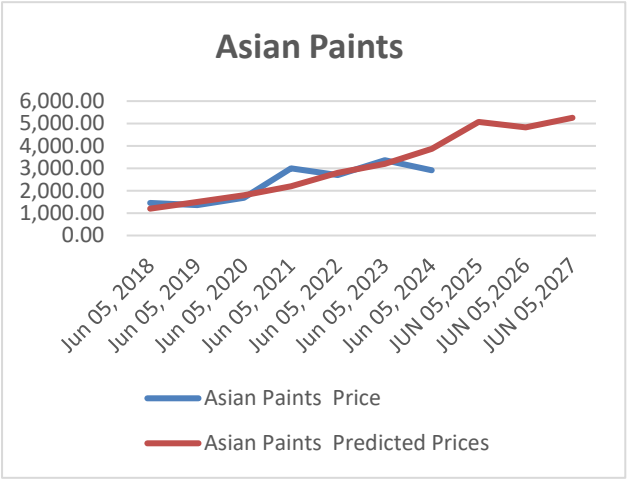


Figure 4. Actual Price Vs Predicted Price (Asian Paint)

- Cipla

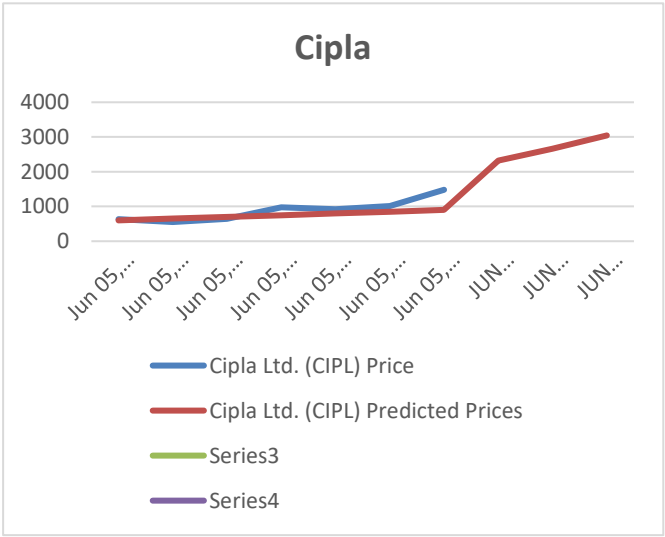
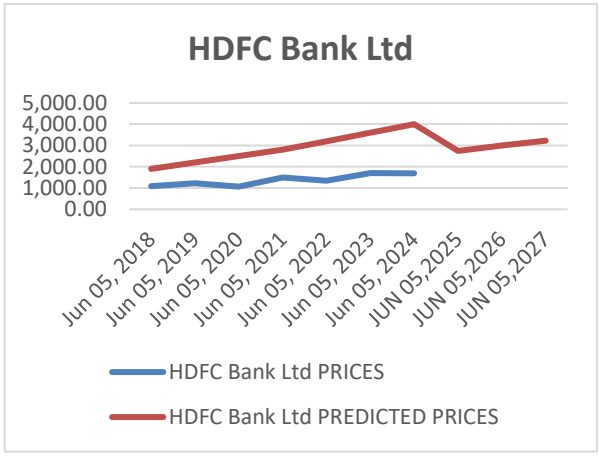


Figure 5. Actual Price Vs Predicted Price (Cipla)

- HDFC Bank

Figure 6. Actual Price Vs Predicted Price (HDFC Bank)



- Indian Oil Corporation

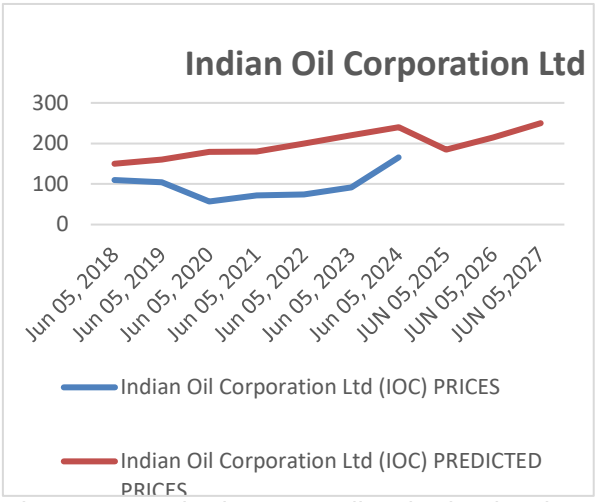


Figure 7. Actual Price Vs Predicted Price (IOC)

- Maruti

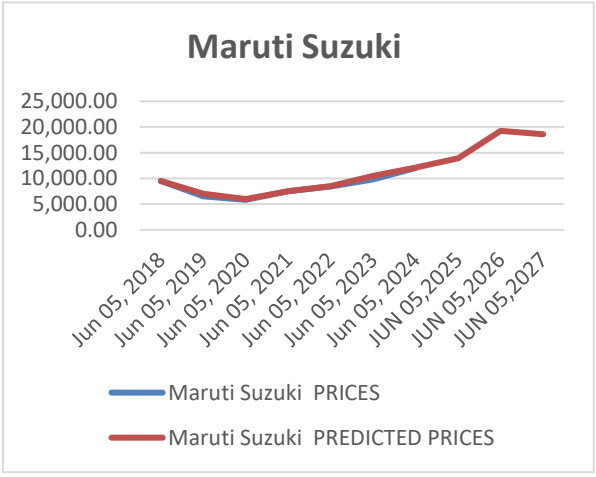


Figure 8. Actual Price Vs Predicted Price (Maruti)

- Oil and Natural Gas Corporation

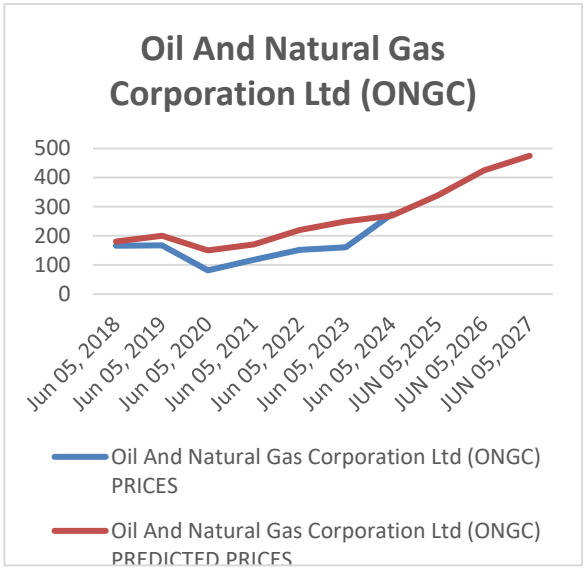


Figure 9. Actual Price Vs Predicted Price (ONGC)

- State Bank of India

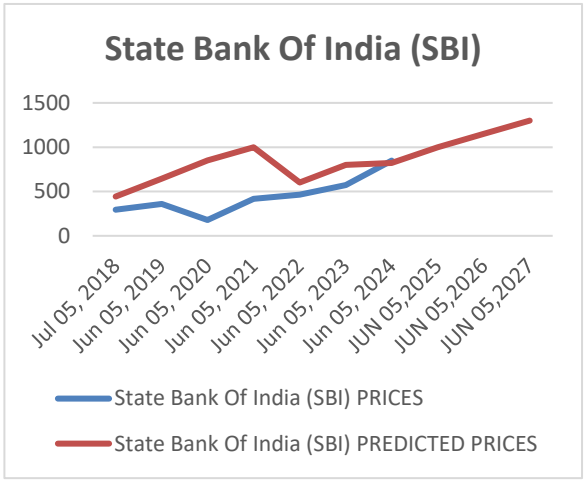


Figure 10.Actual Price Vs Predicted Price (SBI)

- Tata motors

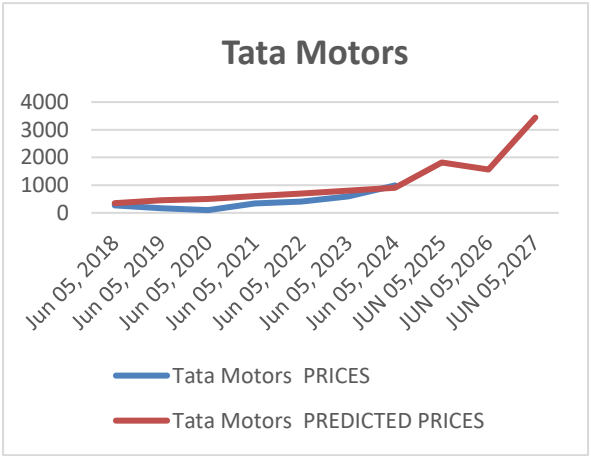


Figure 11. Actual Price Vs Predicted Price (Tata Motors)

4.2 Analysis

% of Error can be Positive or Negative.

% of Error < than |5%|= High Accuracy.

_ _" RI (UURU " _ _ 0RGHUDWH $F

% of Error > than |10%|= Low Accuracy.

The formula to calculate % of Error is given below:

$$\%\ Error = \frac{P.Price - A.Price}{Price} * 100$$

1) Asian Paint

$$\%\ Error = \frac{1677.1 - 1747}{1747} * 100$$

% Error =-4.00114

2) Cipla

$$\%\ Error = \frac{638.34 - 655}{655} * 100$$

% Error =-4.69295

3) HDFC Bank

$$\%\ Error = \frac{1030 - 1042}{1042} * 100$$

% Error =-1.15163

4) IOC

$$\%\ Error = \frac{85.82 - 91}{91} * 100$$

% Error =-5.58241

5) Maruti

$$\%\ Error = \frac{5795.93 - 5897}{5897} * 100$$

% Error =-1.713922

6) ONGC

$$\%\ Error = \frac{81.85 - 85}{85} * 100$$

% Error = 3.70588

7) SBI

$$\%\ Error = \frac{184.76 - 192}{192} * 100$$

% Error =-3.770833

8) Tata Motors

$$\%\ Error = \frac{102.57 - 104}{104} * 100$$

% Error =-1.375

collected. And we get the polarity as output, which contains Positive, Negative, and Neutral Polarity.
The task of analyzing feelings, especially in micro-blogging/Tweets, is still under development and far from complete. So, we propose several ideas which we believe are worth exploring in the future and can produce even greater improvements the performance.

Future Work

- The study can be extended by considering more variables I have considered only Twitter data for the analysis part, where they can make use of news articles and Twitter data at the same time.
- The study can be extended by taking such companies where it is challenging to get the news or data of that company.

Table 4.1 : Hypothesis Testing

Hypothesis Test	% of Error	Accuracy Level
Asian Paint	-4.00114	High Accuracy
Cipla	-4.69295	High Accuracy
HDFC Bank	-1.15163	High Accuracy
IOC	-5.58241	Moderate Accuracy
Maruti	-1.713922	High Accuracy
ONGC	-3.70588	High Accuracy
SBI	-3.770833	High Accuracy
Tata Motors	-1.375	High Accuracy

5. Conclusion

Finding the stock market trend is not easy and a complicated process. Because the stock price depends on many factors like Economics, Politics, Disasters Etc. It is assumed that Stock Price and Twitter Data are related to each other and it has the power to fluctuate the Price of the Stock Market. So a study has been conducted on Sentiment Analysis in the Stock Market using Twitter Data.
Twitter Tweet captures the Sentiment given by the people. We use search terms to search the tweets of a specific company, where we collect the tweets of that company and execute some algorithms on the data

References

1. J. Bean, R by example Mining Twitter for consumer attitudes towards airlines, In Boston Predictive Analytics Meetup Presentation, 2011
2. Yauheniya Shynkevich, T.M. McGinnity, Sonya Coleman, Ammar Belatreche, Predicting Stock Price Movements Based on Different Categories of News Articles, 2015 IEEE Symposium Series on Computational Intelligence
3. Yeole, Ashwini V, P.V.Chavan and M.C.Nikose. *"Opinion mining for emotions determination"* 2015 International Conference on Innovations in Information Embedded and Communications Systems (ICIIECS),2015.

4. Berger, David W., Alain P. Chaboud, Sergey V. Chernenko, Edward Howorka, and Jonathan H. Wright, 2008
5. *Order flow and exchange rate dynamics in* electronic brokerage system data, Journal of International Economics 75, 93-109
6. Biais, Bruno, Thierry Foucault, and Sophie Moinas, 2011, Equilibrium fast trading, Working paper, Toulouse School of Economics. Biais, Bruno, and Paul Woolley, 2011, High frequency trading, Working paper, Toulouse School of Economics, IDEI.
7. Zhang, J., Cui, S., Xu, Y., Li, Q., & Li, T. (2018). A novel data-driven stock price trend prediction system. Expert Systems with Applications, 97, 60-69.
8. Falessi, D., Cantone, G., & Canfora, G. (2010, September). A comprehensive characterization of NLP techniques for identifying equivalent requirements. In Proceedings of the 2010 ACM-IEEE international symposium on empirical software engineering and measurement (pp. 1-10).
9. Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350). IEEE.
10. Chen, Y., Han, D., & Zhou, X. (2023). Mining the emotional information in the audio of earnings conference calls: A deep learning approach for sentiment analysis of securities analysts' follow-up behavior. International Review of Financial Analysis, 88, 102704.
11. Topal, M. O., Bas, A., & van Heerden, I. (2021). Exploring transformers in natural language generation: Gpt, bert, and xlnet. arXiv preprint arXiv:2102.08036
12. Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. Journal of artificial intelligence research, 11, 169-198.