DA4900 - R & D Project

FINAL RESEARCH REPORT

Impact of Twitter posts on stock prices at the Colombo Stock Exchange (CSE)

Supervisors: Mr. Supun Gothama

Mr.Oshadha Senivirathne

Thesis by: 186013N - De Silva S.H.K.H.

186033B - P.A.S.S.Jayawardhana

186041X - Lakshan K.A.N.

186057C - Perera P.P.G.S.

In Partial Fulfillment of the Requirements for the Degree of Bachelor of Business Science (BBSc.)

UNIVERSITY OF MORATUWA

FACULTY OF BUSINESS

DEPARTMENT OF DECISION SCIENCES

ACKNOWLEDGEMENT

We would like to use this opportunity to extend our deepest appreciation to the many people and organizations who have assisted us over the duration of our final year research. To begin, we would like to offer our most sincere gratitude to our mentors, Mr. Supun Gothama and Mr.Oshadha Senivirathne. We are grateful to them for their patience with us, their enthusiasm for the work that we do, and the insightful comments, suggestions, helpful information, and practical advice that they provided to us during all the meetings that were held. Additionally, we are appreciative of the never-ending ideas that they have provided for the development of our models, which have always been of.

And also, during this study time, we would want to convey our thanks and obligation to the other lecturers at University of Moratuwa, all of our relatives, family, and friends for their assistance, compassion, and patience.

STATEMENT OF ORIGINAL AUTHORSHIP

This is to certify that, to the best of our knowledge, the content of this thesis is our own work. This thesis has not been submitted for any degree or other purpose. We certify that the intellectual content of this thesis is the product of our own work and that all the assistance received in preparing this thesis and sources have been acknowledged.

Signature: _____

Name: 186013N - De Silva S.H.K.H.

Signature: _____

Name: 186033B - Jayawardhana P.A.S.S.

Signature: _____

Name: 186041X - Lakshan K.A.N.

Signature: _____

Name: 186057C - Perera P.P.G.S.

ABSTRACT

social media is a rich platform for websites and applications that focus on communication, community-based input, interaction, content sharing, and collaboration. Most people use social media to stay in touch. The Colombo Stock Exchange (CSE) is the main stock exchange in Sri Lanka. It utilizes an electronic trading platform. In the modern world, social media users use posts to share their thoughts with others. Those posts can have a huge impact on stock prices. It will be crashing stock prices or increased hype for increased stock prices. Stock price movements are determined by supply and demand, and emotions play a key role in determining supply and demand. Emotions, on the other hand, are influenced by social media. The purpose of this study is to analyze the Twitter posts on stock prices at the CSE. Twitter is a popular social media platform with a large user base. This paper examines sentiment analysis, which is an approach to natural language processing (NLP) that detects the emotion of Twitter posts. According to the sentiment of those Twitter posts, we find out how the price of a share changes and provides signals for treceivetors. We describe the method of data collection, methods of analysis, and machine learning technique that we performed, and also discuss the impact of Twitter posts on stock prices at the CSE.

Keywords – Sentiment, Sentiment Analysis, Colombo Stock Exchange (CSE), Stock price, Natural Language Processing (NLP), Bidirectional Encoder Representations from Transformers (BERT),

Machine Learning (ML).

Table of Contents

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

CSE	Colombo Stock Exchange
NLP	Natural Language Processing
BERT	Bidirectional Encoder Representations from Transformers
ML	Machine Learning
EST	Eastern Standard
SVM	Support Vector Machine
API	Application Programming Interface
LSEG	London Stock Exchange Group
NB	Naive-Byes
VAR	Vector Autoregression
S & P	Standard and Poor
TF-IDF	Term Frequency – Inverse Document Frequency
NLTK	Natural Language Toolkit
MSE	Mean Square Error
RMSE	Root Mean Square Error

C h a p t e r 1

Introduction

1.1 Background

Social media is one of the most revolutionary tools in the world, and in the ever-evolving business world, social media has gained a very high place in today's business world, primarily because it facilitates communication. In the developing world, effective and efficient communication is essential to running a business efficiently and effectively and to opening the doors of success in the face of business competition. Although social media started about two decades ago, it started to become popular in Sri Lanka about a decade ago, and then businesses in Sri Lanka started using this social media to do their promotion. Communication between businesses and business customers is an extremely important task. Twitter and other social media platforms may have an impact on the stock prices of companies listed on the Sri Lanka Stock Exchange due to communication. We can analyze the different ways in which the stock prices of companies registered in Sri Lanka's stock market go up and down. Social media is a rich platform to study the behavior of users.

The Colombo Stock Exchange is Sri Lanka's sole stock exchange for trading equity securities on the secondary market. The term "stock market" refers to open markets where firms and stock brokers can trade equities. Investing in the stock market involves risk because of the volatility of stock prices and the complexity of the stock market. Studying how tweets on Twitter (social media) affect company share prices on the Colombo Stock Exchange has become a necessity with the growth of social media, so it is necessary to get real information from social media and past information about the Colombo Stock Exchange. The Colombo Stock Exchange (CSE) has 297 companies representing 20 GICS industry groups as of June 30, 2022, with a market capitalization of RS 3,184016 Bn. According to the CSE's official website as of March 16, 2022, the all-share price index is 10222.55. CSE is a company established under the act and is licensed by the Securities Exchange Commission of Sri Lanka. There are 297 public companies registered under the Companies Act in that stock, and it is a big challenge to get the right data from social media politely while researching the share prices in the stock market and the influence of those prices on

social media. Twitter has been included as a social media platform in this, and Twitter is a popular social media platform among investors and CEOs. This is a rich platform to learn about people's opinions and sentiments regarding investments.

We extract data from tweets using sentiment analysis and create Twitter advance accounts to extract tweets, as well as use a consumer key, a consumer secret, an access token, an access token secret, and an authentication code. For extracting tweets, use the tweepy library. Tweepy is an open-source Python package that gives a very convenient way to access the Twitter Application Programming Interface (API) with Python. Text Blob is a Python library for processing textual data. It provides a simple API for diving into common natural language processing (NLP) tasks such as part-of-speech tagging, noun phrase extraction, sentiment analysis, classification, and translation. Tweets were in an unstructured format. Therefore, we had to use text processing methods. For that, we used tokenization, stop word removal, lemmatization, punctuation, and lowercase methods.

Sentiment analysis, or "opinion mining," is a Natural Language Processing (NLP) technique. It also used machine learning and statistics to determine the sentiment orientation of text. The roots of sentiment analysis go back to the beginning of the 20th century. Modern sentiment analysis techniques are classified into three categories.

knowledge-based, statistical, and hybrid. Using sentiment analysis, we can determine whether a given text contains negative, positive, or neutral emotions.

According to the published reports, Twitter had around 296.7 thousand users in Sri Lanka in the middle of 2022. There are several communities on Twitter that actively distribute information about CSE.

We were able to identify specific stock behaviors by analyzing data from Twitter posts and stock prices. Twitter sentiment analysis allows you to keep track of what's being said about the Colombo stock exchange on Twitter posts, and it can help you detect positive, negative, or neutral posts.

The goal of this study is to determine the impact of Twitter posts on CSE stock prices, analyze tweets, and provide signals to external parties about stock price movements. Python is the

programming language that we use in this project. Using the Twitter API, we scrape the data from Twitter. We will be using NLP and machine learning techniques to build the model.

When moving towards with this research, we had to faced following challenges and limitation such as; when extracting many tweets per day, the Twitter account was blocked and a new one was created, which correctly filled the requirement that they expected. Conflicts that occur because of some companies' cultural bias, some company's' behavioral issues and other personal issues, Limited access to data.

1.2 Research Problem and Rationale

Identify the social media impact on stock prices at the CSE. Identify how Twitter posts impact stock prices at the CSE. This research was conducted by the London Stock Exchange Group, also known as London Stock Exchange Group (LSEG). They observed and identified for a long time that there are some accounts that post some details about CSE, and those posts have an impact on stock prices in CSE. And they are trying to create some demand for stock prices. Then, we want to figure out how to predict stock price movements based on the relationship between Twitter posts and stock price movements.

1.3 Research Objectives

Objective 1: To identify whether there is an impact on, stock prices and twitter posts.

Stocks prices are changed in working time of CSE. Due to one reason of twitter posts to change stock prices, let's expect to identify how impact twitter posts on stock prices and expect to identify is there positive or negative impact on share prices.

Objective 2: develop a model according to relationship between Social media post (twitter: tweets) and Share prices oscillating.

According to both data sets expect to create a one model considering tweets sentiment and share prices changes. That model support to find what kind of impact has and support to predict. We used the following objectives to achieve our goal: collect sentiment using the Twitter API; collect stock price data and preprocessing; identify the correlation; and build a prediction model.

- Collect sentiment using the Twitter API.
- Collect stock price data and perform preprocessing.
- Identify the correlation.
- Build a prediction model.

1.4 Scope

Twitter is one of the key social media platforms in Sri Lanka that has an impact on stock prices on the CSE. Then, in this work, we assess how social media posts (tweets) affect stock price oscillating, identify the sentiment of those posts using sentiment analysis, and create models utilizing time series data (CSE data and tweets data).

Chapter 2

LITERATURE REVIEW

2.0 INTRODUCTION

A substantial amount of finance research is devoted to demonstrating how news about a publicly traded company affects its stock price performance and how there is a clear relationship between the direction and magnitude of stock price reaction and the sort of Twitter postings that are disclosed. And several research' authors have found that the effect of news is connected to how corporations are covered in the media.

Furthermore, authors of the other researches have claimed which it is irrelevant whether the information is private or public; what matters is that traders have access to it. Part documented events illustrate how unfavorable false claims about corporations had a substantial impact on stock prices in a short amount of time, as well as how stock price recovered a portion of its value once twitter news was proven to be incorrect (Chowdury, Jansen, Sobel & Zhang, 2009). For instance, on June 14, 2021, the famous football player Cristiano Ronaldo said at a press conference, "Argue, no Coca"; after this comment, he was heavily criticized because the price of Coca-Cola stock dropped by 1.06% that day in comparison to the previous closing price. In addition, the performance of the company's shares over time demonstrated that the declaration had no impact whatsoever on the performance of the company. The announcement was made by Cristiano Ronaldo at 9:43 Eastern Standard (EST), and three minutes earlier, at 9:40 EST, the stock price had dropped to $56.26. In addition, by the end of the trading day, the price of the stock had increased by $0.30, reaching $56.26. The news outlet CNN used the headline "Cristiano Ronaldo throws away coke bottles, causing Coca-Cola stock prices to drop." On the sixteenth of the month of June, Spain had a bigger influence on the price on both the eighteenth and the twenty-third of the same month (Fuehres, Gloor & Zhang, 2011; Brown, 2012).

Similar to other financial markets, the Colombo stock market is extremely volatile, making it possible for any event to have an impact on it and, consequently, the share values of different stocks. However, the announcement of events has a greater impact on the stock market & share returns than the actual events do (Rao & Srivastava, 2012). The effects of a specific announcement of an event on share prices have long been a focal point of finance research, commonly known as

"event studies." It is impossible to exaggerate the role of social media in influencing capital markets. It facilitates the quick transmission of information. Arguments could be made that it levels the playing field. Even investors living on the outskirts of Colombo will be able to keep up with market trends. These developments do not have to be confined to the market's performance (Liu, Mao, Wang & Wei, 2012). It might include substantial information on high-net-worth investors' investment decisions, company performance, managerial decisions made by the companies/stockbroker companies, professional networking, and so on. This allows investors to keep a close eye on information and make changes as needed to maintain a lucrative portfolio.

According to the publications analyzed and past investigations studied, event announcements do effect share prices. And to best of the knowledge, and most previous studies have the only taken one sort of incident into account when researching reaction of market. Few studies have evaluated the effect of all types of financial event announcements on the price of a company's stock. Even this earlier research was restricted to official and publicized customary procedures, those are not publicized via the social media (Twitter). Internet penetration in Sri Lanka has been quickly expanding in recent years. According to the Central Bank of Sri Lanka, overall Internet connections increased by 68.4% in 2014 (Kraus & Nann, 2013; Bollen, Counts & Mao, 2015). As a result, Twitter - social media has evolved into an effective and convincing means of communication. It has affected our life in every way conceivable, including financial decisions. The connection between social media (twitter) as well as the local capital market (Colombo Stock market) will be the topic of this research. This is a strong factor that must not be overlooked if markets are to remain efficient (Bar-Haim, Dinur, Feldman, Fresko & Goldstein, 2011).

2.1. Sentiment Analysis for Collecting Twitter Data

Twitter has experienced a tremendous global expansion due to the rise in popularity of social networks over the past ten years. It is not surprising that many academics started looking for ways to leverage a platform similar to microblogging to find new applications in a range of areas. Online sentiment analysis has become more psychologically and socially popular, but its range of applications is actually quite wide. Some authors were particularly interested in the possibilities of forecasting financial markets (Lee, Pang & Vaithyanathan, 2002). To discern the mood of tweets, text categorization algorithms must be used. Because of the proliferation of papers available to businesses, this topic has grown significantly in recent years. Their requirement for client

understanding has been decupled in order to supply them with services and products that fulfil and generate demand. To assign classes to text documents, many machine learning approaches may be applied. The most well-known are the Support Vector Machine (SVM) & the Naive-Byes (NB) classification. These methods, which differ from unsupervised learning clustering, allow categorization by teaching a computer on annotated datasets so that texts can be automatically tagged. (Forman, 2003).

Text mining relies on a corpus of documents. In this context, academics focus mostly on tweets due to their characteristics: first, they are brief and more accurately describe actual events. Additionally, they are simpler to crawl than papers from other social networks. Any developer who wants to use Twitter data may now use an open Application Programming Interface (API). Furthermore, tweets frequently include hash tags and symbols that aid in the search for related articles. After collecting a sufficient sample, writers use several machine learning approaches, that they may validate using specific methods such as validation of cross and then test using accuracy metrics like the F1-measure (Rao & Srivastava, 2012).

Stock data, which is widely available online, is another source of interest. Authors, on the other hand, nuance their assessments by using multiple levels in their technique. The most common is to use an index made up of multiple firm stocks that can be forecasted separately. Furthermore, the market may be divided into sectors such as agriculture or telecommunications because the tweets connected to these industries are more focused than just investigating tweets about the whole index (Rajamohan and Muthukamu, 2014). Depending on the regression performed in the following stage, the type of data might vary; it's sometimes preferable to the try to anticipate prices are closing, and returns, or simply whether market moves down or up (Chalothorn & Ellman, 2014). The objective of tweet categorization is to utilize them as a time series expected to be related with another time series obtained from market data. The length of time varies, although the majority of authors consider data from one to six months. Function that connects both time series is the unknown, however it may be approached using regression techniques. The Granger causality analysis is a popular approach for determining if series of one time provides predictive the information about other. Vector Autoregression (VAR) & neural networks are two more approaches. After estimating the parameters, the model may be validated and tested using previously annotated data (Bollen, Mao & Zeng, 2011).

The research on these topics is still evolving, although numerous studies have already tackled the issue of employing sentiment analysis to anticipate financial markets. These, however, have significant limits.

First, they focus on the notoriously turbulent Colombo stock markets, whereas this is less true in other regions of the world. Due to the fact that the United States is the most represented nation on Twitter, this platform may never be as effective on other platforms. Second, because to constraints in Twitter's API, the time span evaluated is frequently relatively small. Historical data gathering for more than the few months might be difficult, which raises the dilemma of generalizing conclusions over a longer time period. Third, writers make several assumptions about time series that are rarely checked. They go into sophisticated regression & machine learning techniques without doing systematic formal hypothesis tests (Kivinen & Warmuth, 1995).

2.1.1. Variables

First quantitative research study of Twitter & its subsequent delayed information spread targeted at the stock prediction was published in 2010 by Kwak, Lee, Moon, and Park. Since then, the number of studies has increased, and some writers have synthesized and organized disparate findings. Many businesses have discovered uses for global social media analysis in recent years, although the research remains focused on specific areas and situations. Using sentiment analysis for stock forecasting or correlation analysis (Bar-Haim et al., 2011). There are various interesting values to study in stock data. The variables used in prior research indicate the breadth of potential applications. The stock's closing price or the adjusted closing price, which is the closing price changed to account previously paid dividends & distribution of others, are two of the most common dependent variables that are observed. Additionally, some authors tried to employ different market levels while concentrating on particular businesses or industries. Adjusted price is previous price less dividends & distributions specified above. (Sandner et al.,2010), Consider, for instance, the likelihood of an increase in traded volumes, which is positively correlated with the number of tweets that reference a stock. Additionally, some authors make an effort to forecast both return and volatility (Challothorn & Ellman, 2014).

The returns are calculated using the natural of the logarithm returns R of the S(t), which are the stock prices over the course of one day. This additional step comes with a number of benefits, including the normalizing of variation. Even while many authors use other internet sites as the source of their text data, Twitter's success may be partially explained by the fact that it makes

research easier. However, combining microblogs with mainstream finance news raises concerns about the disparity in prediction power between the two sources. More stock info is accessible through the texts. Although (Sandner et al., 2014) also tried a two-day delay, several authors believed that a one-day delay between the twitter and stock data was appropriate. When it comes to technique, studies using Twitter data frequently start by counting the number of tweets without classifying them based on opinion or mood. (Liu et al.,2012) examined daily tweets number on S& P500 stocks. Benefit of this type a choice is that it simplifies operations of the eliminating a machine learning and the manual annotation phase of many tweets at the hand. Volume of the message is calculated by calculating natural of the logarithm of number of the tweets. The number of people who tweeted is a modest version of message volume which is frequently used by Gionis et al., (2012).

Nonetheless, these type factors are very simplistic and are unlikely to produce meaningful findings in case of stock movement prediction. Furthermore, because they don't supply any of opinionated information, they aren't suitable for the initiating of the relationship with the bearish or bullish market. Because of that, writers frequently use distinct public the mood states in correlation analyses. Technique might vary in many ways, although the fundamental premise stays same. Aim is to pick an indication that accurately represents reality (Chalothorn & Ellman, 2014).

Overall mood of the tweets is most traditional time series in this sense. It's made up of the 2 as well as 3 classes of point data to that the tweets are attributed. For instance, the tweet is about the Colombo stock under consideration might be favorable, neutral or negative. (Bollen et al., 2011) integrate extra mood components in order to support this structure's basic approach to lowering human mood. (Bollen et al., 2011) experiment with different states of mood in another paper. However, other writers may find it excessively basic to derive bullishness / bearishness from overall emotion.

2.1.2. Classification Techniques

In general, classification is assigning a text label or the value of discrete using a function of discriminating based on previously categorized data or prior experiences. After that, the class can be forecast based on this approach and the new data. It also works when we try to use it to help forecast the discrete value for each data point that we have collected. The banking business is a

good illustration of the binary classification because it can predict not only whether or not a certain person is qualified for a loan, but it can also predict various classes to solve other kinds of concerns (Bollen, Mao& Zeng, 2011).

Probabilistic classifier Naive Bayes is special prominent text classification approach. It functions on the basis of two very strong assumptions: the first is the independence of the conditional, which states that terms are independent of each other given class, and the second is the positional independence, which is the same as considering the bag-of-words model. Both of these assumptions state that terms are independent of each other. And despite the fact that these fundamental assumptions are made, And Naive Bayes has been shown to be highly effective in both the theoretical and also the empirical study, in addition to being quite easy to understand. (Rao & Srivastava, 2012).

2.2. Stock Price Forecasting via Sentiment of the Twitter Post

Financial engineering's burgeoning and crucial issue of stock price forecasting is made all the more significant by the new methods and strategies that are consistently gaining ground in this field.

The consistent use of social media in the modern day has reached levels that have never been seen before, which has led to the concept that the behaviour of stock prices can be integrated with the emotions of the general public. The goal is to characterize and identify their patterns in such a way that this association can be validated, and then to make use of those patterns in order to predict the behaviour of future stock prices. In addition, there is no doubt that when aggregated, tweets can provide an accurate reflection of the mood of the general public, even if individual tweets are boring. Opinion mining and analysis of sentiment began to increase alongside the growth in the number of blogs and other forms of social media. Both of these practices became increasingly popular. In, a complete summary of previous work was provided (Wysoki', 2001). The research endeavour aimed to investigate stock market message boards for more than 3,000 equities in order to determine whether or not there was a correlation between the quantity of discussion forum message volume and the quality of message, and whether or not this had any impact on the volume or price of a stock. This study provided a significant contribution by demonstrating a large positive connection between the volume and stock returns on the following trading day and the number of posts placed on the discussion boards during non-market hours. (Gr čar et al., 2013).

Researchers claim that a tenfold increase in the number of message posts made during the overnight hours led to a 15.8% gain in the next day's volume of stock as well as a 0.8% improvement in the next day's stock returns. In a similar manner, we took into consideration the comments that were posted on company-related message boards and calculated the capacity of those comments to affect changes in the stock price. Text categorization and sentiment analysis were used by the researchers to examine over 1.5 million posts that had been made on two different message boards for 45 different companies. This allowed them to establish the tone of each individual comment. Through the use of their work, they were able to demonstrate the existence of a positive correlation between the postings on the message board and the price changes in the stock market the following day (Gao et al., 2007).

Both of the initiatives mentioned above used standard message boards, attempts were made to include new media. Many people have utilized Twitter, Facebook, and the other platforms of the social media to forecast movement of the stock prices also market in general. For instance, linked to tweet collecting and processing methods and expressly suggested utilizing Twitter as the corpus for sentiment analysis (Deng et al., 2013).

The researchers employed unique emoticons to create a set training for emotion classification, eliminating the need to manually tag tweets. Based on happy & the sad emoticons, training their set was separated into both positive and also negative samples. The work of is, without a shadow of a doubt, widely recognized as being some of the best in the industry. In their work, the researchers discussed the application of sentiment analysis to a large corpus of Twitter discussions in order to predict the general disposition of Twitter users on a given day. After that, the findings were input into the neural network's prediction engine, which then anticipated movement of the Colombo stock prices the following day with a stated accuracy of 88.6 percent of the prediction of the Dow Jones Industrial Average. (Mao et al., 2013).

Furthermore, the article accomplishes a 75 percent accurate forecast method on Twitter & DJIA feeds using the Fuzzy Neural Networks of Self-Organizing. In the course of their inquiry, they devised a one-of-a-kind questionnaire consisting of phrases to analyze tweets for their emotion. Not least but certainly not least is a noteworthy publication. They decided to take a less complicated method by concentrating on the top 100 businesses included in the "Standard and Poor's" (S&P) index and collecting only connected tweets (Sharma, 2011). Following this, an analysis was done to see whether or not the sentiment expressed by a corporation on Twitter had

any bearing on the changes in price or volume. Because there is such a large amount of "noise" on Twitter, the decision was made to put the dollar sign before stock market symbols. This practice was promoted by the website stocktwitstook.com and also by the users of that website. Using this terminology, they were able to collect just tweets generated and exchanged by financial market enthusiasts (Rao and Srivastava, 2012). To prevent purchasing dangerous stocks, investors evaluate a company's performance and stock before deciding whether to buy the company's shares. An examination of the business's performance on social media networks is part of this evaluation. In the financial and stock market industries, Twitter is a well-liked social networking tool. Every day, around five hundred million tweets are updated by one hundred million active Twitter users. These tweets allow people to share their ideas, opinions, feelings, and prophecies, which may then be transformed into useful information (Vardavaki and Mylonakis, 2013). However, investors cannot analyze such vast amounts of the social media data on their own. It is almost impossible for anyone to finish on their own. As a consequence of this, investors require a computerized analysis system that is capable of automatically evaluating stock movements by making use of the enormous volumes of data contained within the data sets. (Grčar, Lavrač et al., 2013).

Previous stock prediction research has used a substantial amount of expertise to historical or the tweeter data. When conducting research, the use of historical data denotes the application of a method known as technical analysis, which employs the application of mathematics to analyze data in an effort to predict future stock market movements and prices. Researchers used a variety of machine learning techniques, such as deep learning and regression analysis, on historical stock price data (Ramesh and Nimalathasan, 2011). Exogenous elements like social media were not taken into consideration in these studies due to methodological limitations. It is essential to make advantage of the data that Twitter provides since the events that are reported on social media may have a big impact on stock prices and trends. This is due to the fact that stock values change according to human behavior, which can be reflected in social media. Analysis of mood on social media platforms provides a wealth of information that may represent optimistic or pessimistic perspectives on equities and trends. In recent years, a substantial amount of study has been conducted on the topic of sentiment analysis on a wide variety of topics, such as Twitter feeds and movie reviews. (Agarwal et al., 2011).

The constraint is the selection of the time duration. When equities are subjected to large fluctuations, the association between stock data and the Twitter data becomes apparent. When

market is not responsive to fluctuations, using projections becomes monotonous (Dragota and Oprea, 2014). To utilize tweet as the input for the stock movement prediction pretty securely, the investor should be confident that stocks would move significantly enough. And this begs question of the relevance of such forecasts.

Chapter 3

Methodology

3.1 Data & Data Extraction

To create a model for this analysis, two data sets are collected, the tweets with sentiment analysis and a data set which has included prices of the Colombo Stock Exchange (Opening price and closing price) with relevant dates.

The data which are collected from Twitter social media platform is by using the Twitter advance account and search APIs, and when doing the tweet extraction process, mainly using the keywords. Such as “CSE”, “Colombo Stock Exchange”, “ABANS ELECTRICALS PLC”, “ACL CABLES PLC”, “GOOD HOPE PLC”, “LB FINANCE PLC”, “WINDFORCE PLC” etc. likewise using company names also. In tweets, data is set to have “Negative, Neutral, and Positive” variables as “Sentiment Analysis.” This variable takes as an independent variable. In CSE data set have the company name, share volume, trade volume, previous close price, and open price, high, and low, last trade as variables. The last trade variable (share price) takes as the dependent variable. When gathering data for this research, considered Data set I and Data set II.

1) *Data set I*

   The data extracted from Twitter platform (Tweets/ messages) are included Data set I.

2) *Data set II*

   The data collected externally (from Colombo Stock Exchange) are included in Data set II. The research applies quantitative research approach because it deals with data such as statistics and numbers and facilitates data collection and analysis using machine learning techniques.

3) *Data set III*

   The dataset, which is created using by combining CSE data set (dataset I) and Tweeter scraped data set (dataset II).

3.1.1 Variables

Regarding data set, which is gain from the CSE, they are mainly consisted with numerical variables with existing variables and newly added variables.

The following figure 1 shows that sample date set of one company in CSE.

	A	B	C	D	E	F	G
1	Trade Date	Open (Rs.)	High (Rs.)	Low (Rs.)	Close (Rs.)	Difference (Rs.)	Impact (-1/0/+1)
2	08/08/2022	21.3	21.3	21	21	-0.3	-1
3	08/05/2022	21	21.5	21	21.4	0.4	1
4	08/04/2022	21	21	21	21	0	0
5	08/03/2022	21.3	21.3	21	21.1	-0.2	-1
6	08/02/2022	21.3	21.3	20.3	21	-0.3	-1

Figure 1:CSE Data set

The following table 1 shows that selected variables and their descriptions.

Name of the Variable	Description
1. Date (Old Variable)	Considering date of the stock price
2. Open – Rs (Old Variable)	Opening price of the share with relevant day
3. High – Rs (Old Variable)	Highest price, that take for the share to relevant day
4. Low – Rs (Old Variable)	Lowest price, that take for the share to relevant day
5. Close – Rs (Old Variable)	Close price of the share with relevant day
6. Difference (New Variable)	The difference between the highest price and the lowest price of a share.
7. Impact: 0/1 (New Variable)	Considering the difference value, marked that value as negative (0), positive (+) one.

Table 1: Description of CSE dataset variables

Tweeter data set, which is scraped from Tweeter, mainly consist with Date, Company, and Related Tweet details.

The following figure 2 shows that sample date set of Tweeter data set.

	B	C	D
1	Date	Company Name	Tweets
2	11/29/2022 17:30	ACL Cables	_Island _Mindset _amila A very valid point.
3	10/16/2022 10:32	ACL Cables	So far , LK haven't had a person who is fitting to be a pet dog of such
4	10/15/2022 15:28	Amana Bank	This is not for investing experts like you..this is just for poor investors
5	10/15/2022 8:04	ACL Cables	Fyi.. https//t.co/YrOqn39wPl
6	10/15/2022 4:41	Arpico Insurance	In the current context, there is no better investment than FDs. Justify

Figure 2:Tweeter Data set

The following table 2 shows that variables and their descriptions of Tweeter data set.

Name of the Variable	Description
1. Date	Considering date of the tweeter post published.
2. Company Name	The company name which is related to the scraped tweeter post.
3. Tweet	Scraped tweet.

Table 2: Description of Tweeter dataset variables

The combine data set, which is combine Tweeter dataset and CSE dataset which is mainly consist with Trade_Date, Open, Close, Change, Label, Tweets, Neutral, Positive, Compound, Subjectivity, Subjectivity and Sentiment_Analysis.

The following figure 3 shows that sample date set of Combine data set (Scraped Tweeter and CSE data sets).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	Trade_Date	Open	High	Low	Close	Change	Lable	Tweets	Negativ	Neutral	Positive	Compo	Subject	Polarity	Sentime	_Analysis
2	6/28/2022	50.5	52.1	50.5	51	0.5	1	LIOC top 1	0	0.795	0.205	0.2023	0.5	0.5	Positive	
3	6/27/2022	53.6	53.6	51.2	53.8	0.2	1	If you have	0.194	0.806	0	-0.8555	0.363333	-0.1075	Negative	
4	6/27/2022	53.6	53.6	51.2	53.8	0.2	1	_UNP will	0.304	0.696	0	-0.8625	0.8	-0.8	Negative	
5	6/22/2022	51	51	50	50.5	-0.5	0	Entire Sri I	0	0.647	0.353	0.9729	0.491667	0.166667	Positive	
6	6/22/2022	51	51	50	50.5	-0.5	0	CSE https/	0	1	0	0	0	0	Neutral	

Figure 3:Combine Data set (CSE & Tweeter)

The following table 3 shows that variables and their descriptions of Tweeter data set.

Name of the Variable	Description
1. Trade_Date	Considering date of the stock price.
2. Open	Opening price of the share with relevant day.
3. High - Rs	Highest price, that take for the share to relevant day.
4. Low - Rs	Lowest price, that take for the share to relevant day.
5. Close - Rs	Close price of the share with relevant day.
6. Change	The difference between the highest price and the lowest price of a share.
7. Label	Considering the difference value, marked that value as negative (-0), positive (+) one.
8. Tweets	Scraped tweet which are after sentiment analysis.
9. Negative	Negative Sentiment Score.
10. Neutral	Neutral Sentiment Score.
11. Positive	Positive Sentiment Score.
12. Compound	Sum of the valence score of each word in the lexicon.
13. Subjectivity	Amount of personal opinion, if a sentence has high subjectivity.
14. Polarity	Polarity refers to the overall sentiment conveyed by a particular tweet.
15. Sentiment Analysis	Emotion that showing by a particular tweet.

Table 3: Description of Combine dataset variables

3.2 Theoretical Approach

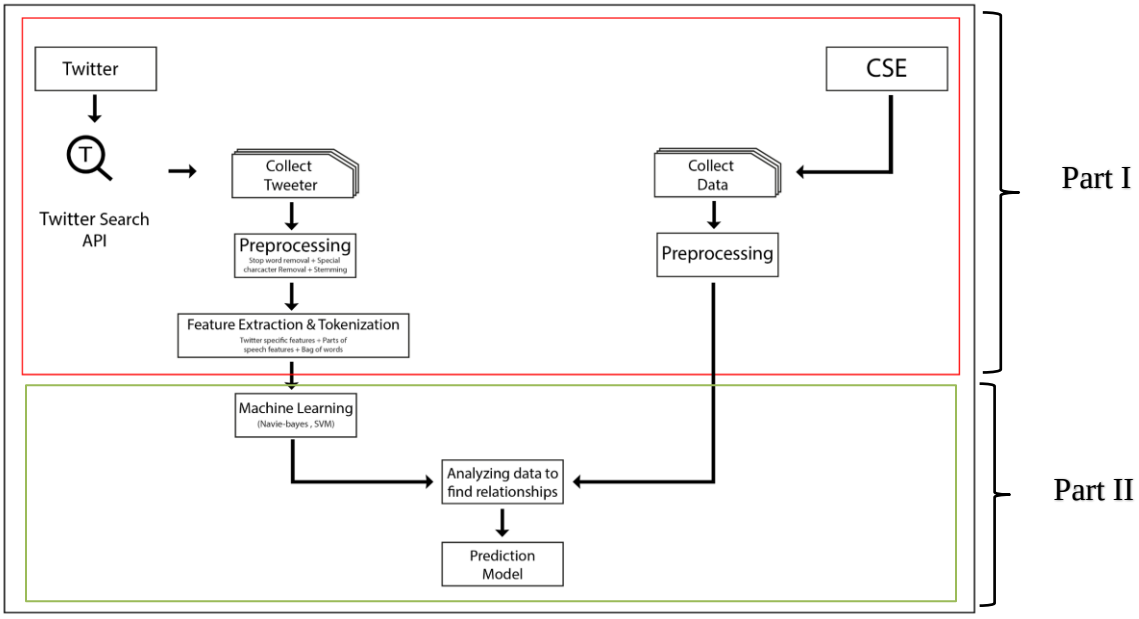

Figure 4:Conceptual Framework of the study

According to this framework, In the first part, mainly we preprocess (stop word removing, special character removing, stemming etc.) and clean the scraped tweet and use of NLP, World Cloud, IF-IDF techniques and approaches to do the sentiment analysis of Twitter posts. Then in 2nd part has model prediction and the results analyze.

3.2.1 Natural Language Processing (NLP)

Natural language refers to how humans communicate with one another, such as through speech and text. It is also a component of artificial intelligence that assists computers in comprehending, interpreting, and manipulating natural language. NLP benefits from recent advances in machine learning and deep learning, and its algorithms are typically Machine Learning (ML) based. It can significantly help with tasks like speech recognition, translation, and sentiment analysis. (Lutkevich & Burns, 2021) So we can use the techniques under NLP for sentiment analysis.

Tokenization

This is the first step in natural language processing, and it's useful for breaking up larger units of text like paragraphs, sentences, and phrases into more manageable chunks. Because of this, Natural Language Toolkit (NLTK) provides each word with its own unique token. When it comes to sentiment analysis, tokenization is more crucial than in any other part of natural language processing. By using this technique, the Tweeter data set, and we can separate words from one sentence etc. (Singh, 2019).

	Date	Company Name	Tweets	Negative	Neutral	Positive	Compound	Subjectivity	Polarity	Sentiment_Analysis	text_len	text_word_count	tokenized
0	11/29/2022 17:30	ACL Cables	_Island _Mindset _amila A very valid point.	0.000	1.000	0.000	0.0000	0.300000	0.200000	Positive	46	7	[island, mindset, amila, a, very, valid, point]
1	10/16/2022 10:32	ACL Cables	So far , LK haven't had a person who is fitt...	0.000	1.000	0.000	0.0000	0.666667	0.200000	Positive	94	21	[, so, far, lk, havent, had, a, person, who, i...
2	10/15/2022 15:28	Amana Bank	This is not for investing experts like you..t...	0.155	0.700	0.146	-0.0772	0.525000	-0.150000	Negative	159	28	[, this, is, not, for, investing, experts, lik...

Figure 5:Data output for Tokenization

Stop word Removal

Stop words are typically English words that do not affect the meaning of the sentence, and they are not required for the categorization of sentiment polarity. Stop words can be found in both spoken and written English. This method may be safely ignored, and there will be no adverse effects on the sentence.

	Date	Company Name	Tweets	Negative	Neutral	Positive	Compound	Subjectivity	Polarity	Sentiment_Analysis	text_len	text_word_count	nonstop
0	11/29/2022 17:30	ACL Cables	_Island _Mindset _amila A very valid point.	0.000	1.000	0.000	0.0000	0.300000	0.200000	Positive	46	7	[island, mindset, amila, valid, point]
1	10/16/2022 10:32	ACL Cables	So far , LK haven't had a person who is fitt...	0.000	1.000	0.000	0.0000	0.666667	0.200000	Positive	94	21	[, far, lk, havent, person, fitting, pet, dog,...
2	10/15/2022 15:28	Amana Bank	This is not for investing experts like you..t...	0.155	0.700	0.146	-0.0772	0.525000	-0.150000	Negative	159	28	[, investing, experts, like, youthis, poor, in...

Figure 6:Data output for Stop word Removal

Punctuations

The use of punctuation can assist in removing unneeded components of the data. By using this, we can remove the unwanted punctuations used in extracted tweets.

	Date	Company Name	Tweets	Negative	Neutral	Positive	Compound	Subjectivity	Polarity	Sentiment_Analysis	text_len	text_word_count	stemmed
0	11/29/2022 17:30	ACL Cables	_Island _Mindset _amila A very valid point.	0.000	1.000	0.000	0.0000	0.300000	0.200000	Positive	46	7	[island, mindset, amila, valid, point]
1	10/16/2022 10:32	ACL Cables	So far , LK haven't had a person who is fitt...	0.000	1.000	0.000	0.0000	0.666667	0.200000	Positive	94	21	[, far, lk, havent, person, fit, pet, dog, ima...
2	10/15/2022 15:28	Amana Bank	This is not for investing experts like you..t...	0.155	0.700	0.146	-0.0772	0.525000	-0.150000	Negative	159	28	[, invest, expert, like, youthi, poor, investo...

Figure 7:Data output for Punctuations

Stemming

The process of stemming involves reducing a word to its root form, which can then be attached to various suffixes and prefixes, and this technique helps us find the root word.

	Date	Company Name	Tweets	Negative	Neutral	Positive	Compound	Subjectivity	Polarity	Sentiment_Analysis	text_len	text_word_count	punct
0	11/29/2022 17:30	ACL Cables	_Island _Mindset _amila A very valid point.	0.000	1.000	0.000	0.0000	0.300000	0.200000	Positive	46	7	Island Mindset amila A very valid point
1	10/16/2022 10:32	ACL Cables	So far , LK haven't had a person who is fitt...	0.000	1.000	0.000	0.0000	0.666667	0.200000	Positive	94	21	So far LK havent had a person who is fittin...
2	10/15/2022 15:28	Amana Bank	This is not for investing experts like you..t...	0.155	0.700	0.146	-0.0772	0.525000	-0.150000	Negative	159	28	This is not for investing experts like youthi...

Figure 8:Data output for stemming

BERT Model

The Bidirectional Encoder Representation Transformer, or BERT for short, is a model of the language that was developed by Devlin and colleagues (2018). The name of the method suggests that the process of picking up the language has been done in both directions. Both the entire English Wikipedia (2,500 words) and the Brown Corpus were used during BERT's training (800M words). In this analysis, BERT model for keyword extraction from Twitter posts.

3.2.2 TF-IDF

The term frequency-inverse document frequency, or TF-IDF, is a statistical measure that determines how pertinent a given word is to a given document within a collection of documents. Multiplying the term frequency metric by the inverse document frequency metric is required to arrive at the TF-IDF.

Formula for calculate TF -IDF value as follows.

$$tf\ idf(t,d,D) = tf(t,d).idf(t,D)$$

$$tf(t,d) = log\ (1 + freq(t,d))$$

$$idf(t,D) = log\ log\ \frac{N}{count(d \in D: t \in d)}\ )$$

Figure 9:IF – IDF formulae

3.2.3 Linear regression

A prominent technique for predictive analysis and modeling is linear regression. Regularization parameters are used to increase prediction accuracy while lowering the error of the regression model because when linear regression is used to predict stock prices, the model has a tendency to overfit the data.

Reason for use Linear regression

In CSE and tweets data behave according to date and time. Therefore, this data has time series method. Because of that reason try to run our data using linear regression model.

3.2.4 Random Forest

Because there are large number of data just want to get best patten using machine learning algorithms. After combine and preprocessing CSE & tweets data it is show complex data set and want to check accuracy of the sentiment. Random forest is also the best to use classification and regression things. Random forest algorithm is able to use develop and merge multiple decision trees to design a "forest" machine learning algorithm. For a more precise forecast, Random Forest produces numerous decision trees that are then combined. To prevent overfitting, the algorithm takes a collection of variety of trees into account. The algorithm inserts random data samples into each tree and produces results based on the input variables; the most popular result is then deemed the outcome and the best result.

3.2.5 Discriminant Analysis

discriminant analysis, is a widely utilized method for supervised classification issues. A dimensionality-reduction method called linear discriminant analysis is applied as a preprocessing step in pattern classification and machine learning applications. The purpose of linear discriminant analysis is to features higher-dimensional features onto a lower-dimensional space. In this research we use discriminant analysis for get accuracy. However, Discriminant analysis accuracy also lower than Random Forest analysis.

Chapter 4

Result and Discussion

4.1 Sentiment Analysis

Sentiment analysis is an approach to Natural Language Processing technique. which is used to identify the emotional tone behind the body of the text. When using the sentiment analysis tools, its normally process using the unit of text, those units can be sentences, paragraphs, books etc. And the output of the sentiment analysis is, quantitative score and classification to show whether the texts that process under the sentiment algorithm considers that those text units are fetch positive, negative, or neutral emotion.

There are four types of Sentiment analysis, Fine-grained Sentiment Analysis, Emotion Detection Sentiment Analysis, Aspect-based Sentiment Analysis, and Intent Analysis. In this study we used an Aspect-based Sentiment Analysis type.

In this study as we explained in the methodology, after scraping the tweets from the Twitter using the tweepy library we calculate the sentiment of those tweets by using NLP (Natural Language Processing) and ML (Machine Learning) algorithms.

No.	Date	Company Name	Tweets	Negative	Neutral	Positive	Compound	Subjectivity	Polarity	Sentiment_Analysis
0	11/29/2022 17:30	ACL Cables	_Island _Mindset _amila A very valid point.	0	1	0	0	0.3	0.2	Positive
1	10/16/2022 10:32	ACL Cables	So far , LK haven't had a person who is fitting to be a pet dog of such an imaginary leader.	0	1	0	0	0.666666667	0.2	Positive
2	10/15/2022 15:28	Amana Bank	This is not for investing experts like you..this is just for poor investors who have been considerably mislead by SM. No advice is required for pros like you	0.155	0.7	0.146	-0.0772	0.525	-0.15	Negative
3	10/15/2022 8:04	ACL Cables	Fyi.. https//t.co/YrOqn39wPl	0	1	0	0	0	0	Neutral
4	10/15/2022 4:41	Arpico Insurance	In the current context, there is no better investment than FDs. Justifying stocks against Fixed income for new/inexperienced investors are beyond stupidity IMO. Risk management is not known to many inexperienced investors.	0.189	0.735	0.076	-0.5106	0.533333333	-0.0583	Negative
5	10/15/2022 3:38	ACL Cables	Very true. If someone started investing in January, this frustration would have been much bigger.	0.184	0.659	0.157	-0.1361	0.6725	0.2275	Positive
6	10/15/2022 2:51	Amana Bank	Never liquidate your fixed deposit and buy stocks, if you haven't been in CSE at least 1 year. Invest smaller amounts and experience 1 market cycle completely. Understand the complexity. Investing is never easy. FDs are the safest and guaranteed source of income. No comparables	0.097	0.846	0.057	-0.2278	0.466666667	-0.0633	Negative

Figure 10:Data set after sentiment analysis

As explained in the methodology section above table (Figure 10) shows output data set after running the sentiment analysis on Tweets that we scrape from the twitter. Negative, Neutral, Positive, Compound, Subjectivity, Polarity and Sentiment analysis columns are generated automatically from the sentiment analysis algorithm.

Under those columns most important columns id Polarity. Also, we can define it as sentiment score. Polarity referred to the overall sentiment that fetch by the particular text, phrase, or word. Polarity classification of the tweets means calculate the positive, negative, or neutral sentiment on entire tweet. Normally polarity lies between the range of 1.0 and –1.0. 1.0 refers to as positive sentiment and –1.0 refers negative sentiment. If the polarity score shows 0 it defines as neutral sentiment. After assigning the scores to all words by individually, final polarity calculated by taking and average of all the sentiments. According to the polarity score each tweets has its own sentiment as shown above figure 10.

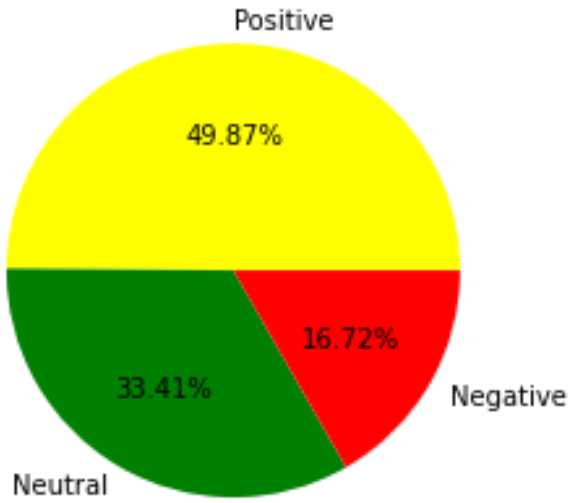

Figure 11:Percentage of each sentiment.

Normally pie charts are used to represent the percentages of whole, and pie charts can use to represent the percentage in a set point in time. Above pie chart (Figure 11) shows the percentage each sentiment as a percentage of total tweets that we were scrape. According to the above pie chart 49.87 % tweets have positive sentiment, emotion that given to the society by those positive tweets are confident, hopeful, and good aspects. 16.72% tweets have negative sentiment, emotion that given to the society by those negative tweets are unpleasant, depressing, or harmful. 33.41%

tweets have neutral sentiment, those neutral sentiments tweets are neither good nor bad, neutral posts are not helping to the reputation, but they are not hurting it either.

According to the above analysis 83% of the tweets are helping to the reputation of the company or neither helped nor not. Only 17% tweets are harmful for the reputation of the company.

Figure 12:Word cloud visualization of all tweets

Basically, word cloud is a graphical representation of word frequency. Most used words appear within the text analyzed by the word cloud. Collection of words are sketch in varied sizes in the word cloud. The biggest and bolder words are the most commonly used words within the selected texts. Figure 3 visualize word cloud of all tweets that we scrape from the tweets. Those are the mostly used words in tweets. According to the Figure 12, “https” is the most used words, but we have to ignore it because it is a word with no specific meaning in common use. Other than that, we can see “market, share, stock, investor, daily, CSE” are the most used words in all tweets.

Below word cloud (Figure 13) visualizes the commonly used words only in positive sentiment tweets. In here also we ignore word “https.” In positive sentiment tweets, “good, stock, good, market, share, new, best, investor” are the some of most frequent words.

Figure 13:Word cloud visualization of positive sentiment tweets

Figure 14:Word cloud visualization of negative sentiment tweets

Above word cloud (Figure 14) visualizes the commonly used words only in negative sentiment tweets. In here also we ignore word “https.” “Price, will, company, due, risk” are the most frequent words in negative sentiment tweets.

Figure 15:Word cloud visualization of neutral sentiment tweets

Above word cloud (Figure 15) visualizes the commonly used words only in neutral sentiment tweets. In here also we ignore word “https.” “correct, media, daily, market, share” are the most frequent words in negative sentiment tweets.

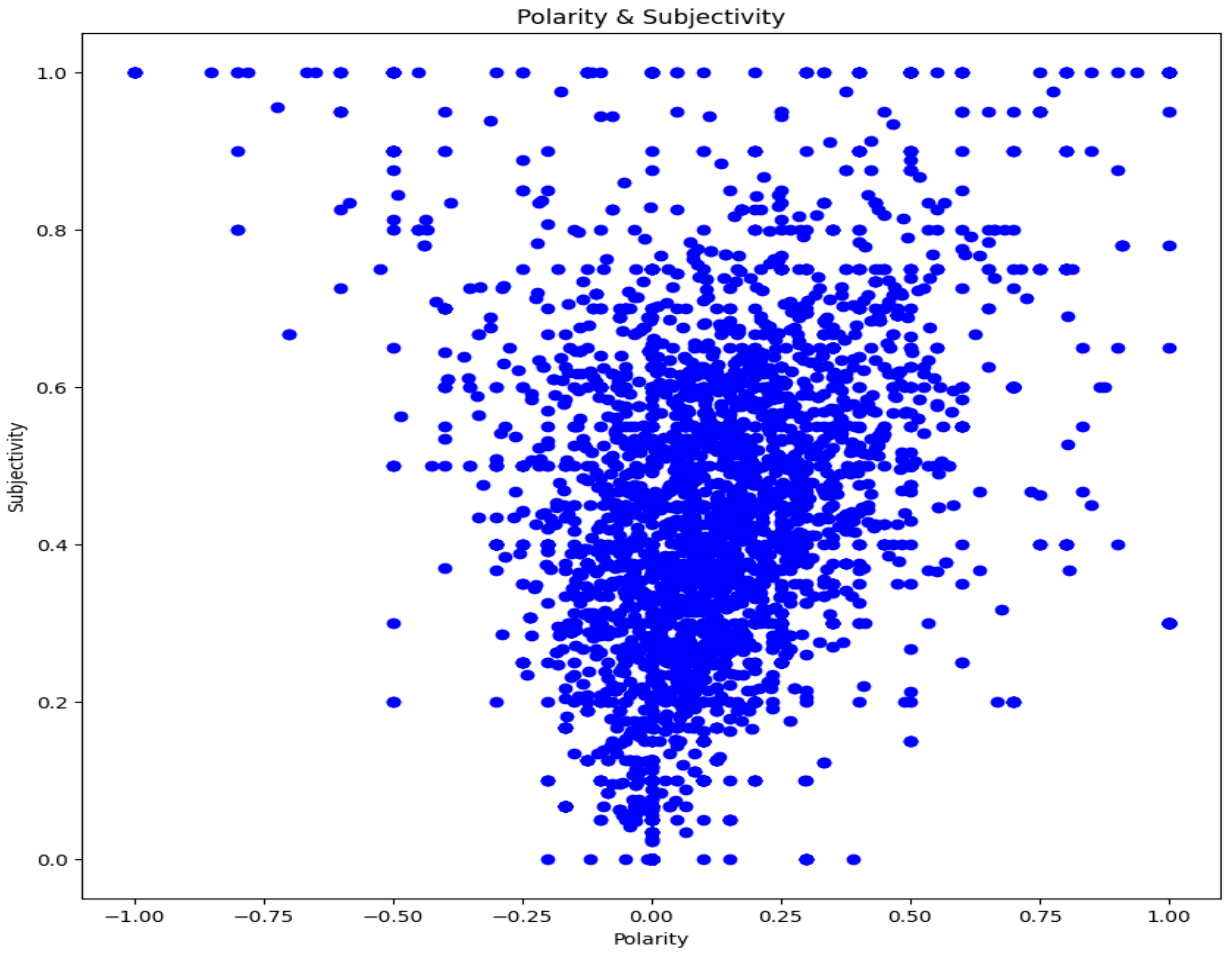

Figure 16:Scatter plot of Polarity & Subjectivity

Figure 16 represents the scatter plot of polarity and subjectivity for all tweets. Polarity is spread mostly in center of the graph while the subjectivity is spread across the graph. It indicates that our tweets are in wide range of subjectivity. When most tweets are fall in polarity score between –0.50 to +o.75. That shows extremely negative or extremely positive sentiment tweets are low. Users have shared their opinion about the Colombo Stock Exchange, most of the tweets mid-range of negative sentiment tweets and with more weight on the positive sentiment tweets. When analyzing the graph, we can see tweets that are with the low subjectivity lies on the center of the polarity. Mostly polarity is approximately 0. Low subjectivity tweets are more likely to be neutral sentiment. When the subjectivity is high in the tweets, sentiment is more likely to have a diverse range of negative to positive sentiments

TF-IDF very popular NLP topic when it comes dealing with human languages. In text processing after cleaning the data those data need to needs to be converted into a numerical format where each

word is represented by a matrix. In below Figure 16 show the bigram of all tweets. Bigram means two consecutive words in a sentence. And Figure 17 represent the trigram of all tweets. Trigram means three consecutive words in a sentence. Sometimes using unigrams, we cannot present details correctly. Unigrams means only one word. For further processing we develop the bigram and trigram of tweets.

```
[('daily ft', 332),
 ('sri lanka', 269),
 ('correct answer', 99),
 ('stock market', 72),
 ('ft https', 71),
 ('answer option', 70),
 ('long term', 66),
 ('share price', 48),
 ('thanks sharing', 43),
 ('port city', 42),
 ('sri lankan', 38),
 ('dividend yield', 35),
 ('cse https', 30),
 ('covid 19', 30),
 ('palm oil', 28),
 ('short term', 27),
 ('2021 https', 27),
 ('like expo', 25),
 ('dipd hayc', 25),
 ('bull market', 24)]
```

Figure 17:n_2 bigram

Above bigram (Figure 17) represents the most frequent consecutive two words in all tweets. “daily ft” is the most frequent to words. In this bigram we can see “stock market, long term, share price, port city, bull market” kind of words. Below trigram (Figure 18) represents the most frequent consecutive three words in all tweets. Using this bigram and trigram we can get understand of the words that are using in tweets when it's related to the Colombo Stock Exchange.

```
[('correct answer option', 70),
 ('daily ft https', 70),
 ('shared tweet end', 15),
 ('tweet end poll', 15),
 ('colombo stock market', 14),
 ('news daily mirror', 14),
 ('stock like expo', 12),
 ('business news daily', 12),
 ('daily ft says', 12),
 ('finding great investment', 11),
 ('dividend paying stocks', 11),
 ('10 successful investing', 10),
 ('successful investing finding', 10),
 ('investing finding great', 10),
 ('high dividend paying', 10),
 ('good times ahead', 9),
 ('ordinary voting shares', 8),
 ('valuation metrics doesn', 8),
 ('metrics doesn mean', 8),
 ('growth stock like', 8)]
```

Figure 18:n_3 trigram

Figure 19 represent the stock price changes in days and Figure 20 represent the sentiment of the tweets in days those tweets are posted. Dates are considered in 2006- 01- 01 to 2022- 08- 31. Stock price changes range from –15 to +25. Range of the sentiment polarity is between the –0.6 to +1.0. By analyzing those graphs, it is difficult to identify the relationship between sentiment and stock price change according to the date. Some time periods we can see there is no deviation in the stock price changes as well as same time period we can see sentiment was neutral. In some time periods these two graphs show the negative relationship between stock price changes and sentiment. but some time periods it indicates the positive relationship.

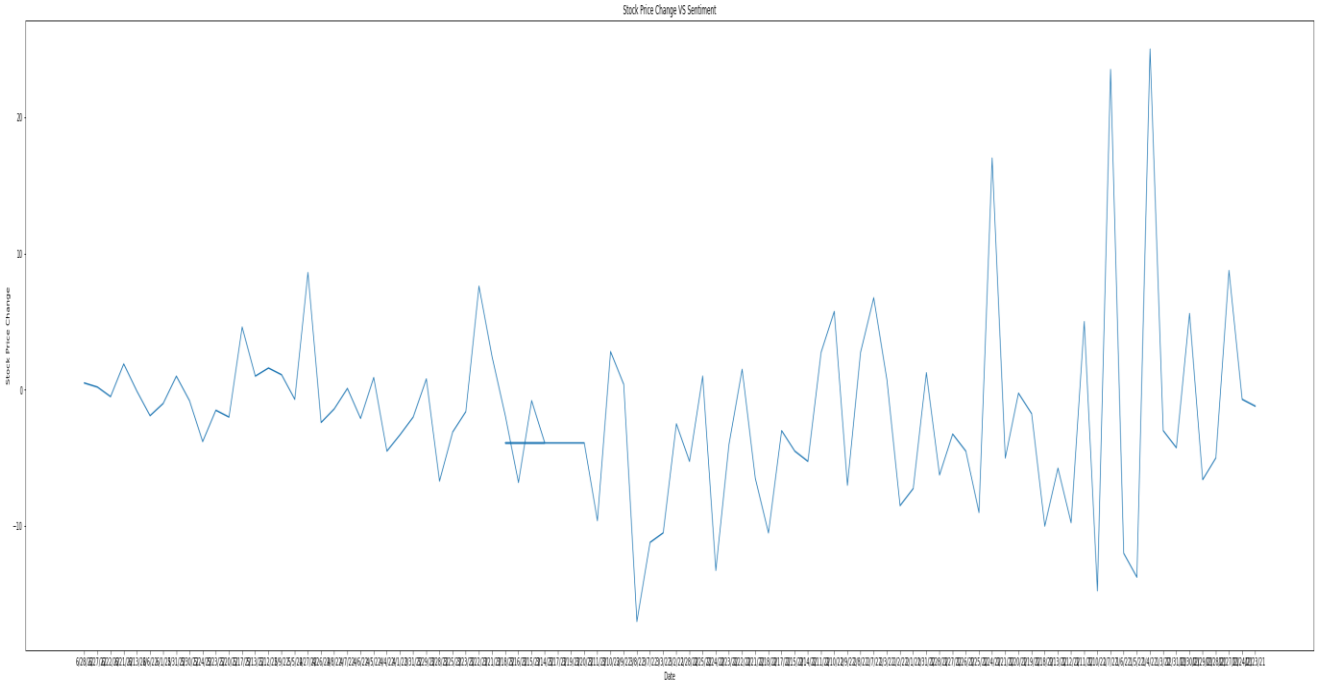

Figure 19:Stock price change vs Date

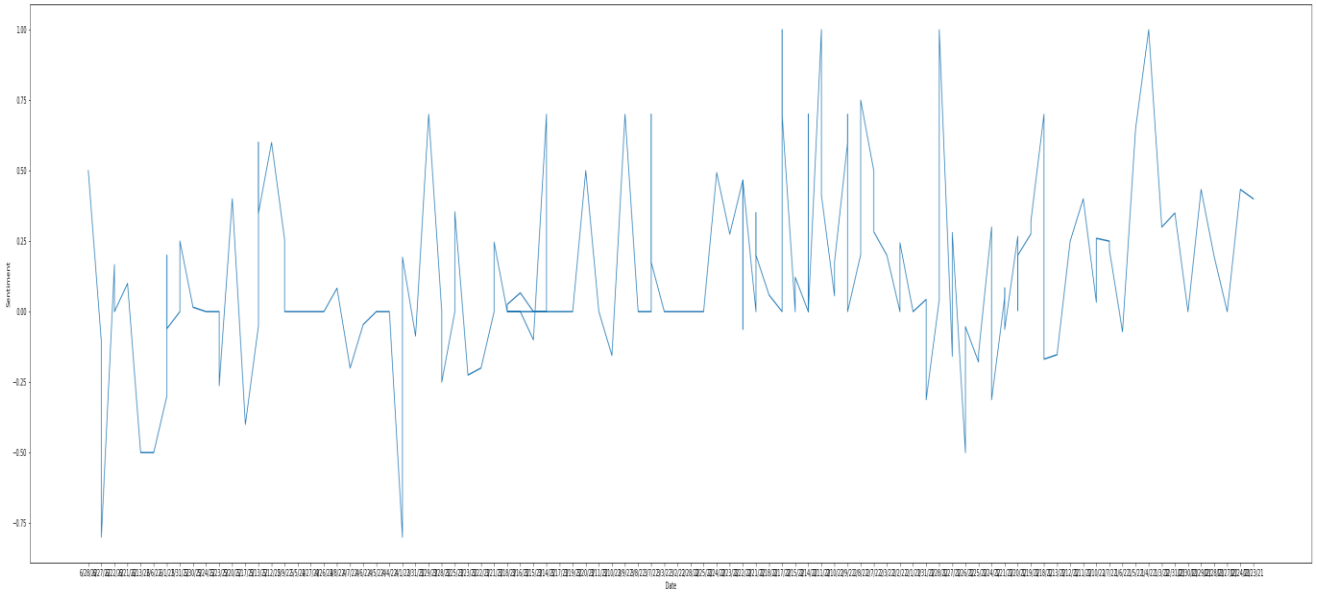

Figure 20:Sentiment vs Date

4.2 Result & Interpretations

Mode / Facto	Discriminant Analysis	Random Forest	Linear Regression
Accuracy Score	0.8529411764705882	0.941176795882353	-
Intercept	-571.34900844	-	-65.83999639986897
Model Score	94.07407407407408	100.0	51.02162154340297
MSE	0.14705882352941177	0.058823529411764705	0.14782943472765568
RMS	0.3834824944236852	0.24253562503633297	0.3844859356694919

Table 4: Results obtained for both RF, Discriminant Analysis and Linear Regression models

Accuracy Score

According to the above table, it's filled with the results obtained by the second part of the conceptual framework figure (). It's mainly based on a combined date set (CSE dataset & Tweet dataset) and the result obtained by the model, which is made for prediction of oscillating price of stocks in CSE.

By considering this result obtained using discriminant analysis, random forest and linear regression that result shows as order 85.29%, 94.11% and 0%. According to that result, we are able to see, there is higher accuracy of random forest models than discriminant analysis and linear regression models.

Intercept

In the linear regression model, the coefficient describes the change of the response triggered by one unit increase of the independent variables. When focusing on discriminant analysis and linear regression model's intercept, it shows negative value when independent variables are zero.

Mean Square Error (MSE)

When focusing on Mean Square Error of Random Forest, it has 0.058823529411764705 (5.88%) error. Random forest has lowest Mean Square Error more than discriminant and linear regression model.

Root Mean Square Error (RMSE)

When we consider the value of RMSE, it's a good measure to say how accurately the model predicts the response. So according to the above result, the random forest model has the lowest RMSE value than the other two methods. Therefore, we can say the random forest model is the best fitted model for this prediction.

Chapter 5

Challenges and Limitations

There may be some possible limitations in this study. When we are conducting this study, we had to face several challenges. Our research topic is performed sentiment analysis on twitter posts to identify the impact of twitter posts on stock prices at the Colombo Stock Exchange (CSE). As we explained previously in the methodology section, we mainly focus on the two data sets.

Data set 1:
Changes in share prices in daily basis.

Trade Date	Open (Rs.)	High (Rs.)	Low (Rs.)	Close (Rs.)	Difference (Rs.)	Impact (-1/0/+1)
08/08/2022	160	160	151	146.75	-13.25	-1
08/05/2022	144	160	144	146.75	2.75	1
08/04/2022	155	164.5	155	162	7	1
08/03/2022	144.25	145	144.25	145	0.75	1
08/02/2022	147.75	155	144	132.25	-15.5	-1
08/01/2022	140.25	140.25	140.25	132.25	-8	-1
7/29/22	145	145	140	132.25	-12.75	-1
7/28/22	145.25	145.25	145.25	132.25	-13	-1
7/27/22	132	145.75	132	132.25	0.25	1
7/26/22	147	147	146.25	130.5	-16.5	-1
7/25/22	149	149	147.75	130.5	-18.5	-1
7/20/22	135.5	142	129	130.5	-5	-1
7/19/22	135	135	135	135.25	0.25	1
7/18/22	149	149	130	135.25	-13.75	-1
7/15/22	130	138	130	135.25	5.25	1
07/11/2022	131	131	131	132.5	1.5	1
07/07/2022	139.75	139.75	132	132.5	-7.25	-1
07/06/2022	147	147	133.5	127.75	-19.25	-1
07/05/2022	139	139	139	127.75	-11.25	-1
07/01/2022	137	140	137	127.75	-9.25	-1
6/30/22	140	140	127	127.75	-12.25	-1

Figure 21:CSE data set

Data set 2:

Twitter posts that are relevant to the Colombo Stock Exchange with the sentiment analysis.

	Date	Company	Tweets	Negative	Neutral	Positive	Compoun	Subjectivi	Polarity	Sentiment_Analysis
2	7/16/2022 9:21	C M Holdi	A battle that you win cancels all your mistakes.- Nic	0.417	0.357	0.226	-0.296	0.4	0.8	Positive
3	7/15/2022 13:05	Balangoda	_g _sl While big banks are thinking about debt r	0.083	0.831	0.086	0.0258	0.1	0	Neutral
4	7/9/2022 12:20	Amana Ba	Finally, people have realised true power and comm	0	0.574	0.426	0.9118	0.74	0.32	Positive
5	6/27/2022 5:45	Alliance F	LIOC top 100 SH comparison 30/05 to 20/06 ...	0	0.795	0.205	0.2023	0.5	0.5	Positive
6	6/25/2022 14:08	Alliance F	If you have a little bit of commercial sense, you wo	0.194	0.806	0	-0.8555	0.363333	-0.1075	Negative
7	6/25/2022 13:39	Alliance F	_UNP will have a separate hell to go for rescuing a d	0.304	0.696	0	-0.8625	0.8	-0.8	Negative
8	6/21/2022 18:50	Alliance F	Entire Sri Lanka is grateful to team Australia for visi	0	0.647	0.353	0.9729	0.491667	0.166666667	Positive
9	6/21/2022 13:45	Alliance F	CSE https//t.co/OREuFsx5Kl	0	1	0	0	0	0	Neutral
10	6/21/2022 10:46	Alliance F	LIOC in action https//t.co/kT1MoYZczh	0	1	0	0	0.1	0.1	Positive
11	6/21/2022 10:45	Alliance F	"The LIOC has 10,000 MT of petrol in its stores in Trin	0	0.871	0.129	0.7579	1	0	Neutral
12	6/11/2022 7:39	Alliance F	Should "I was wrong" be accepted with NO punishm	0.378	0.493	0.129	-0.6166	0.9	-0.5	Negative
13	6/5/2022 10:47	Alliance F	Should "I was wrong" be accepted with NO punishm	0.403	0.46	0.138	-0.6166	0.9	-0.5	Negative
14	5/31/2022 17:15	Alliance F	But that's not a part of operating profit. Just a one o	0.254	0.553	0.193	-0.296	0.4	-0.3	Negative
15	5/31/2022 15:45	Alliance F	What about one off negative good will ??????	0.344	0.442	0.214	-0.4137	0.5	0.2	Positive
16	5/31/2022 15:36	Alliance F	I haven't reviewed financials. Please have a look at	0.116	0.884	0	-0.2411	0	0	Neutral
17	5/31/2022 15:06	Alliance F	To very small- non finance investors Don't be Foolec	0.039	0.745	0.216	0.8198	0.411429	-0.06071429	Negative
18	5/30/2022 18:23	Alliance F	_UNP Nothing will happen till Rajapakshas exit.	0	1	0	0	0	0	Neutral
19	5/30/2022 16:02	Alliance F	_UNP ????	0	1	0	0	0	0	Neutral

Figure 22:Tweeter data set

Among these two data sets data set 1 was created by using the data gathering from the Colombo Stock Exchange. We can access stock prices in the Colombo Stock Exchange official web site. As we explained in the methodology section, we gathered those data from the Colombo Stock Exchange web site and then we pre-process and cleaned those data.

Data set 2 was expected to be created from the data which scraped from the twitter posts. Using python programming language, we scrape the data from the twitter and compute the sentiment of those posts using one of the Natural Language Processing (NLP) technique called sentiment analysis.

When scraping twitter posts and after that analyzing the sentiment of those posts, we had to face some critical challenges and limitations. We can analyze them under two grant points.

i. *Conflicts that occur because of the cultural bias and other personal issues.*

   Conflicts arise from cultural situations and other personal issues affect for the research study. Because of the cultural situation we had some limitations when conducting this study.

ii. *Limited access to data.*

Because this research involved in surveying in specific data from the certain people, or organizations we faced the limitation of access to the required data. One objective of this research was building a statistical model to identify the relationship between twitter posts and fluctuation of the share prices. The last objective was developing a prediction model to predict then share prices. For that a considerable amount of data was required.

iii. *Conflicts that occur because of the cultural bias and other personal issues.*

This study conducted based on the Colombo Stock Exchange. It means grant focus on the Sri Lankan context. This research was based on the twitter posts. Twitter builds brands personality and awareness. Using twitter business can share information quickly and start conversations with your target audience. audience would find tweets and content valuable and ideally, even share with their followers. This method variously used by developed countries.

Sri Lanka is still developing country. There are some cultural limitations in Sri Lanka than the developed countries. In developed countries most of the times majority of the people used their mobile phones, laptops computers or any other electronic device as an education and business purposes. But in the other hand developing countries like Sri Lanka, we barely see people use their laptops, mobile phones, computers any other electronic device as an education and business purposes.

According to the survey done by Department of Census and Statistics Sri Lanka in January 2020, digital literacy rate of the Sri Lanka is 50.1%. Its means one out of two people between the aged group 5-69 digital literate. Digital literacy means a person is considered as a digital literate person if he/she could use computer, laptop, tablet, or smartphone on his/her own. According to the survey computer literacy of the Sri Lank is 32.3%. Its means Almost one out of three-person between aged group 5-69 computer literate. Computer literacy means a person is considered as a computer literate person if he/she could use computer on his/her own. As an example, even if a 5-year-old child can play a computer game then he/she is considered as a computer literate person.

According to the reports 22.2% households owned laptop or desktop computers. It means one out of every five households owned at least one computer or laptop

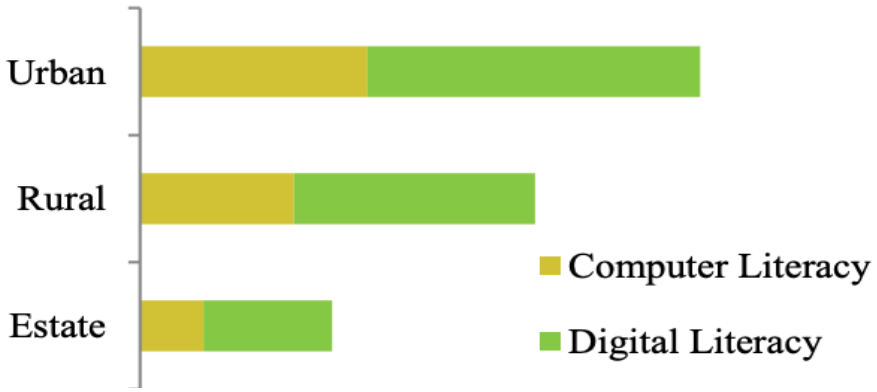

Figure 23:Computer literacy and Digital literacy by Sector

Above graph (Figure 23) show the computer literacy and digital literacy sector wise according to the survey done by the Department of Census and Statistics Sri Lanka in January 2020. And Department of Census and Statistics Sri Lanka done this survey among the computer aware population by occupation.

Occupation group	Computer literacy rate (%)		
	2018	2019	2020
Sri Lanka	63.2	65.1	65.2
Managers, Senior Officials and Legislators	70.2	76.1	78.3
Professionals	87.4	87.9	90.3
Technicians and Associate Professionals	83.8	85.1	87.7
Clerks and Clerical support workers	89.9	90.9	94.2
Services and Sales workers	60.5	55.6	55.2
Skilled Agricultura1, Forestry and Fishery workers	21.0	22.6	21.9
Craft and Related Trades workers	41.6	41.7	38.6
Plant and Machine operators and Assemblers	42.4	43.1	39.7
Elementary occupations	30.5	27.8	34.0
Armed Forces Occupations & unidentified occupations	80.1	88.1	80.2

Table 5: Computer literacy among computer aware employed population (aged 15 – 69 years) by Occupation group – 2018, 2019 & 2020

Above table (Table 5) shows the computer literacy among computer aware employed population aged group between 15 – 69 years by Occupation group in 2018, 2019 & 2020 years. Computer literacy among the employed population employees who are aware of computer in Sri Lanka is around 65.2 % in 2020. Managers and senior officials and legislators have 78.3 computer literacy in 2020, other professionals have 90.3 computer literacy in 2020, technical and associate professionals have 87.7% computer literacy in 2020, clerks and clerical support workers have 94.2% computer literacy in 2020.

All the above stats are commonly show the computer and digital literacy of Sri Lanka in 2020. Now we look at the social media usage of the Sri Lanka. There were 8.20 million social media users in Sri Lanka in January 2022. According to the data published by Meta, it indicates that Facebook had nearly 7.15 million users, Instagram had 1.55 million users in Sri Lanka in early 2022. According to the Google there are 6.68 million YouTube users in Sri Lanka in early 2022. LinkedIn had 1.50 million users in Sri Lanka. But Twitter had only 296.7 thousand users in Sri Lanka in early 2022.

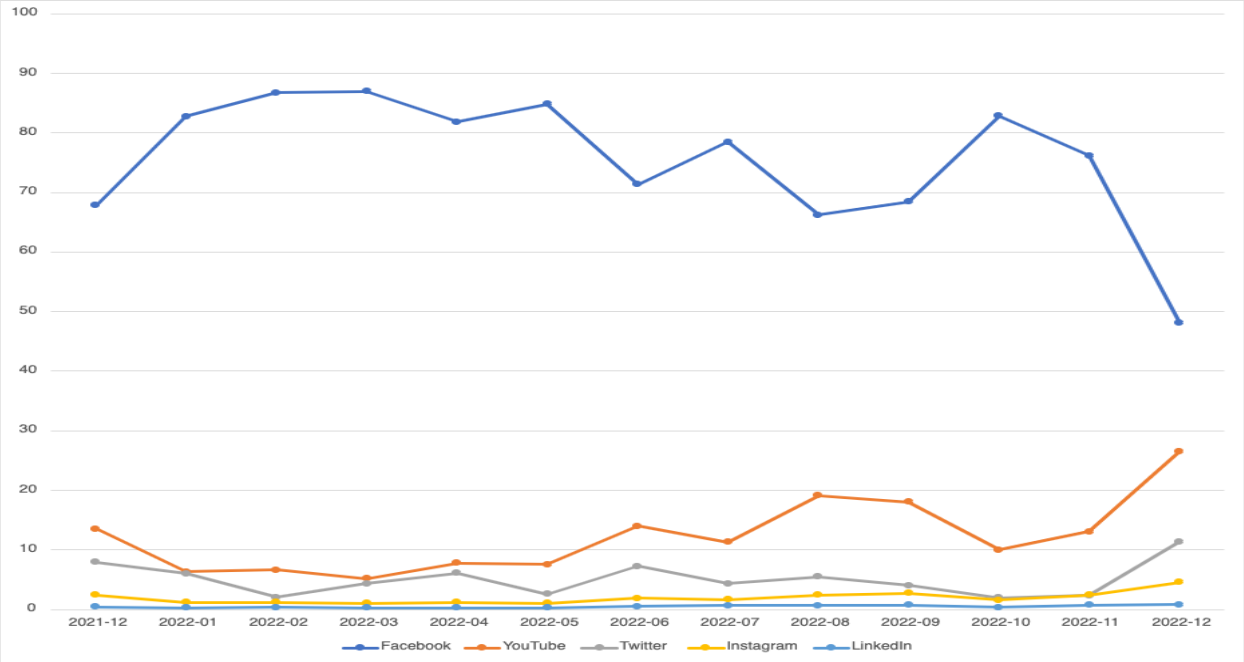

Figure 24:Social media Stats in Sri Lanka (Dec 2021 - Dec 2022)

Above graph show the comparison of social media usage in Sri Lanka in December 2021 to December 2022. Highest used social media platform is Facebook, second highest platform is YouTube. Comparing to YouTube and Facebook, usage of Twitter platform is bit lower than the other two main platforms.

Above mentioned stats show Sri Lanka has some good computer literacy and digital literacy as well as considerable usage in social media platforms. Those stats only show the common statistics. But when consider the business usage of the social media platforms in Sri Lanka, there is no lot of surveys or research papers to find the statistics or evidence. Most probably reason for that is in Sri Lankan culture doing investments or business based on the social media platforms is not popular due to various reasons.

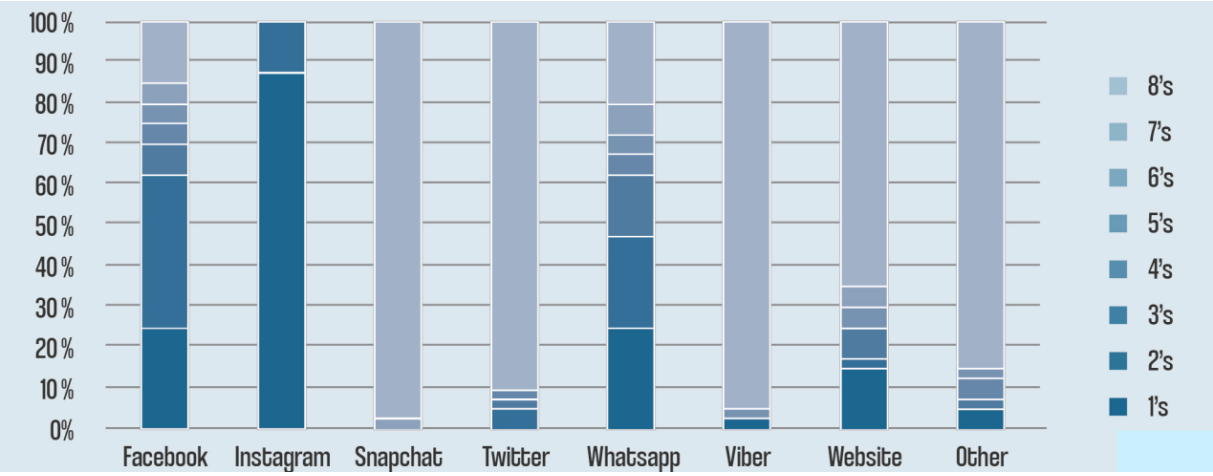

Figure 25:Social media usage for market sector in Sri Lanka (Dec 2021 - Dec 2022)

Above chart (Figure 25) shows the social media usage for market sector in Sri Lanka based on the survey done by institute of policy studies of Sri Lanka. Respondents were asked to rank each platform most used 1 to least used 8. According to the above chart WhatsApp is the most used platform. Compared to WhatsApp usage of Twitter platform is very low.

And most of the broker companies in the CSE used WhatsApp platform to share some hints with their closed groups. And also, CSE authorities are vehemently refuse depend on and promoting social media in the investment decisions. Mainly because of the difficulties that faced when verification the data sources and fraud.

iii. *Limited access to data.*

Access of the data is extremely limited in both tweets and stock prices at the CSE. When we are scraping the tweets from the Twitter something that caught our attention is, some CEOs of the leading companies in Sri Lankan stock market at least they do not have a Twitter accounts. Most of them are did not promote the social media when it come to the investments. Another points that take our attention is, most of the accounts, which are posts tweets that are related to the CSE most of the times they used company symbol in their tweets rather than the company name. Because of that people who are not aware of the CSE or new investors of the CSE may be faced some difficulties to identify the tweet exactly. Those tweets mainly used company symbols, because of that sometimes it is difficult to identify correct tweets.

In the CSE we can get only the data about the stock prices. Some days CSE closed early without operation normally. Some days stock market is closed. And there are some other reasons like inflation rates, interest rates, foreign exchange rate, and decisions that taken by the respective companies' board meetings are also can be affected to change the stock prices. Those data are not addressed in the CSE data. We have to reach those data differently.

Chapter 6

CONCLUSION AND RECOMMENDATIONS

This study was done to identify the impact of the Twitter posts to the CSE stock prices. We used python library called tweepy to scrape the tweets from the twitter and then perform sentiment analysis to identify the sentiment of those tweets. After perform sentiment analysis we get if the respective tweet's sentiment is positive, negative, or neutral. 49.87 % tweets have positive sentiment, 16.72% tweets have negative sentiment, 33.41% tweets have neutral sentiment. Tweets have spread subjectivity it means users have shared their opinion about the Colombo Stock Exchange, most of the tweets mid-range of negative sentiment tweets and with more weight on the positive sentiment tweets. "Market, share, stock, investor, daily, CSE" are the most frequent words in all tweets. "Good, stock, good, market, share, new, best, investor" are the some of most frequent words in positive tweets. "Price, will, company, due, risk" are the most frequent words in negative sentiment tweets. "Correct, media, daily, market, share" are the most frequent words in neutral sentiment tweets. We can use those words in future analysis.

Finally, we chose the random forest model to identify the impact of the tweeter on the CSE company's price. By the results obtained from the confusion matrix, the Accuracy Score, MSE and RMSE proves that the RF model is the best fitted model for this prediction.

When conducting this study, we faced some limitations and challenges. In this study we collect data from mainly in two sources. When scraping the data from the Twitter we faced some serious challenges. Most of the times those tweets are used company symbol name therefore difficult to understand that tweets are exactly related to the CSE. CSE authorities and some CEOs of the leading companies are vehemently refused depend on the social media when they are taking investment decisions.

As a result of the analysis, we are able to show that tweets have an impact on the oscillation of stock prices. As a consequence of this, a random forest model will be able to forecast stock price oscillations using the tweets that are posted on the social media site Twitter.

But in the CSE data only shows the fluctuation of the share prices. There may be various reasons for changing the share prices. Those reasons are not capture or not mentioned in the CSE. Some reasons are unable to gathered and some other reasons we have to reach those separately.

Chapter 7

BIBLIOGRAPHY

"Social media and the Sri Lankan stock market - Features | Daily Mirror," *www.dailymirror.lk*. https://www.dailymirror.lk/90899/social-media-and-the-sri-lankan-stock-market#sthash.oVdvrYAi.dpuf (accessed Jan. 30, 2023).

"Emotion and Sentiment Analysis: A Practitioner's Guide to NLP," *KDnuggets*. https://www.kdnuggets.com/2018/08/emotion-sentiment-analysis-practitioners-guide-nlp-5.html

"False social media msgs' drop Colombo shares," *Print Edition - The Sunday Times, Sri Lanka*. https://www.sundaytimes.lk/220130/business-times/false-social-media-msgs-drop-colombo-shares-470626.html (accessed Jan. 30, 2023).

"Social media and the Sri Lankan stock market." Accessed: Jan. 30, 2023. [Online]. Available: https://www.sec.gov.lk/wp-content/uploads/2021/08/Social-media-and-the-Sri-Lankan-stock-market-1.pdf

P. Shah, "My Absolute Go-To for Sentiment Analysis — TextBlob.," *Medium*, Nov. 06, 2020. https://towardsdatascience.com/my-absolute-go-to-for-sentiment-analysis-textblob-3ac3a11d524

R. Krishn Mishra and Y. Kumar Dwivedi, *Sentiment analysis and classification of Indian farmers' protest using twitter data*. International Journal of Information Management Data Insights, 2021.

C. Bhadanea, H. Dalalb, and H. Doshic, *Sentiment analysis: Measuring opinions*. International Conference on Advanced Computing Technologies and Applications (ICACTA2015), 2015.

Computer Literacy Statistics – 2020 (Annual). Department of Census and Statistics, Sri Lanka, 2021.

[1]"List of Sri Lankan public corporations by market capitalisation," Wikipedia, Jan. 29, 2023. https://en.wikipedia.org/wiki/List_of_Sri_Lankan_public_corporations_by_market_capitalisation (accessed Jan. 30, 2023).

"Sri Lanka Stock market return - data, chart," TheGlobalEconomy.com. https://www.theglobaleconomy.com/Sri-Lanka/Stock_market_return/

J. Chen, "Sector Breakdown," Investopedia, Feb. 23, 2022. https://www.investopedia.com/terms/s/sector-breakdown.asp

"Getting Started with the Twitter API," developer.twitter.com. https://developer.twitter.com/en/docs/twitter-api/getting-started/about-twitter-api

J. A. Ou and S. H. Penman, "Financial statement analysis and the prediction of stock returns," Journal of Accounting and Economics, vol. 11, no. 4, pp. 295–329, Nov. 1989, doi: 10.1016/0165-4101(89)90017-7.

Antweiler, W., & Frank, M. Z. (2004). Is all that talk just noise? The information content of internet stock message boards. The Journal of Finance, 59(3), 1259-1294. https://doi.org/10.1111/j.1540-6261.2004.00662.x

Bar-Haim, R., Dinur, E., Feldman, R., Fresko, M., & Goldstein, G. (2011). Identifying and following expert investors in stock microblogs. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 1310-1319). Association for Computational Linguistics. https://aclanthology.org/D11-1121

Bikhchandani, S., and Sharma, S. (2001),"Herd Behavior in Financial Markets", IMF Staff Papers, Vol. 47 No.3. https://doi.org/10.2139/ssrn.228343

Bollen, J., Counts, S. & Mao, H. (2011). Predicti ng financial markets: Comparing survey, news, twitter and search engine data. arXiv preprint arXiv:1112.1051. https://doi.org/10.48550/arXiv.1112.1051

Bollen, J., Mao, H., & Pepe, A. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. ICWSM, 11, 450-453. https://doi.org/10.1609/icwsm.v5i1.14171

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, 2(1), 1-8. https://doi.org/10.1016/j.jocs.2010.12.007

Breusch, T. S., & Pagan, A. R. (1979). A simple t est for heteroscedasticity and random coefficient variation. Econometrica: Journal of the Econometric Society, 1287-1294. https://doi.org/10.2307/1911963

Brown, E. D. (2012). Will twitter make you a bett er investor? a look at sentiment, user reputation and their effect on the stock market. Proc. of SAIS. https://aisel.aisnet.org/cgi/viewcontent.cgi?article=1006&context=sais2012

Bullard, C. G., Crossing, E. E. & Dunphy, D. C. (1974). Validation of the general inquirer Harvard IV dictionary. Harvard University Library.

Castillo, C., Gionis, A., Hristidis, V., Jaimes, A. & Ruiz, E. J. (2012). Correlating financial time series with micro-blogging activity. In Proceedings of the fifth ACM international conference on Web search and data mining (pp. 513-522). ACM. https://doi.org/10.2307/271007

Chalothorn, T., Ellman, J. (2014). TJP: Identifyi ng the Polarity of Tweets from Context. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), 657–662. https://aclanthology.org/S14-2117.pdf

Cheng, J., Hu, M. & Liu, B. (2005). Opinion obser ver: analyzing and comparing opinions on the web. In Proceedings of the 14th international conference on World Wide Web (pp. 342- 351). ACM. https://doi.org/10.1145/1060745.1060797

Chowdury, A., Jansen, B. J., Sobel, K. & Zhang, M. (2009). Twitter power: Tweets as electronic word of mouth. Journal of the American society for information science and technology, 60(11), 2169-2188. DOI:10.1002/asi.21149

Cortes, C., & Vapnik, V. (1995). Support-vector n etworks. Machine learning, 20(3), 273- 297. http://dx.doi.org/10.1007/BF00994018

Deng, X., Li, H., Li, Q., Liu, B., Mukherjee, A. & Si, J. (2013). Exploiting Topic based Twitter Sentiment for Stock Prediction. ACL (2), 2013, 24-29. https://aclanthology.org/P13-2005

Dragota, V., and Oprea, D. S. (2014),"Informational efficiency tests on the Romanian stock market: a review of the literature", The Review of Finance and Banking, Vol. 6 No. 1, pp. 15-28. https://doi.org/10.12795/anduli.2019.i18.10

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. The Journal of machine learning research, 3, 1289-1305. https://dl.acm.org/doi/10.5555/944919.944974

Fuehres, H., Gloor, P. A. & Zhang, X. (2011). Pre dicting stock market indicators through twitter “I hope it is not as bad as I fear”. Procedia-Social and Behavioral Sciences, 26, 55-62. 10.1016/j.sbspro.2011.10.562

Gao, X. Z., Huang, X., Lin, H. & Song, Z. (2007). A Self-organizing Fuzzy Neural Networks. In Soft Computing in Industrial Applications (pp. 200-210). Springer Berlin Heidelberg. DOI: 10.1007/978-3-540-70706-6_19

Grčar, M., Lavrač, N., Smailović, J. & Žnidarši č, M. (2013). Predictive sentiment analysis of tweets: A stock market application. In Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data (pp. 77-88). Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-39146-0_8

Hiemstra, C., & Jones, J. D. (1994). Testing for linear and nonlinear Granger causality in the stock price-volume relation. The Journal of Finance, 49(5), 1639-1664. https://doi.org/10.2307/2329266

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. In Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 168-177). ACM. https://doi.org/10.1145/1014052.1014073

Joachims, T. (1999). Making large-scale support v ector machine learning practical, 1999. Advances in Kernel Methods: Support Vector Machines. http://hdl.handle.net/10419/77178

Kivinen, J., & Warmuth, M. K. (1995). The percept ron algorithm vs. winnow: linear vs. logarithmic mistake bounds when few input variables are relevant. In Proceedings of the eighth annual conference on Computational learning theory (pp. 289-296). ACM. https://doi.org/10.1016/s0004-3702(97)00039-8

Krauss, J., Nann, S. & Schoder, D. (2013). Predic tive Analytics On Public Data-The Case Of Stock Markets. In ECIS (p. 102). 2.http://aisel.aisnet.org/ecis2013_cr/102

Kwak, H., Lee, C., Moon, S. & Park, H. (2010). Wh at is Twitter, a social network or a news media?. In Proceedings of the 19th international conference on World wide web (pp. 591- 600). ACM. http://dl.acm.org/citation.cfm?id=1772751

Lee, L., Pang, B. & Vaithyanathan, S. (2002). Thu mbs up?: sentiment classification using machine learning techniques. In Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10 (pp. 79-86). Association for Computational Linguistics. DOI:10.3115/1118693.1118704

Liu, B. (2010). Sentiment Analysis and Subjectivi ty. Handbook of natural language processing, 2, 627-666. DOI:10.1201/9781420085938-c26

Liu, B. (2012). Sentiment analysis and opinion mi ning. Synthesis lectures on human language technologies, 5(1), 1-167. https://doi.org/10.2200/S00416ED1V01Y201204HLT016

Liu, B., Mao, Y., Wang, B. & Wei, W. (2012). Corr elating S&P 500 stocks with Twitter data. In Proceedings of the first ACM international workshop on hot topics on interdisciplinary social networks research (pp. 69-72). ACM. https://doi.org/10.1016/j.najef.2022.101847

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, No. 1, p. 496). Cambridge: Cambridge university press. https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Pak, A., & Paroubek, P. (2010). Twitter as a Corp us for Sentiment Analysis and Opinion Mining. In LREc (Vol. 10, pp. 1320-1326). http://www.lrec-conf.org/proceedings/lrec2010/pdf/385_Paper.pdf

Rajamohan, S., and Muthukamu, M. (April, 2014), “Impact of Selective Corporate Events on Price Movements of Stocks of Bank Nifty Index”,Indian Journal of Applied Research, Vol. 4 No. 4, pp. 317-320. https://doi.org/10.15373/2249555X/APR2014/98

Ramesh, S., and Nimalathasan, B. (2011), “Bonus issue announcement and its impact on share price of Colombo Stock Exchange in Sri Lanka”, in 8th International conference of Business Management, University of Jaffna, Sri Lanka. https://doi.org/10.12795/anduli.2019.i18.10

Rao, T., & Srivastava, S. (2012). Twitter sentime nt analysis: How to hedge your bets in the stock markets. In State of the Art Applications of Social Network Analysis (pp. 227-247). Springer International Publishing. https://doi.org/10.48550/arXiv.1212.1107

Sandner, P. G., Sprenger, T. O., Tumasjan, A. & Welpe, I. M. (2014). Tweets and trades: The information content of stock microblogs. European Financial Management, 20(5), 926- 957. https://doi.org/10.1111/j.1468-036X.2013.12007.x

Sandner, P. G., Sprenger, T. O., Tumasjan, A. & Welpe, I. M. (2010). Predicting elections with twitter: What 140 characters reveal about poli tical sentiment. ICWSM, 10, 178-185. https://ojs.aaai.org/index.php/ICWSM/article/download/14009/13858

Sebastiani, F. (2002). Machine learning in automa ted text categorization. ACM computing surveys (CSUR), 34(1), 1-47. DOI:10.1145/505282.505283

Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). Biometrika, 52(3/4), 591-611. https://doi.org/10.2307/2333709

Vardavaki, A., and Mylonakis, J. (2013),"How A Specific Market Announcement May Impact The Stock Price Value Of A Particular Firm-An Event Empirical Study", Conflict Resolution & Negotiation Journal, Vol. 2013 No. 1, pp. 108-118. https://doi.org/10.1016/S2212-5671(15)01290-3

Zhang, H. (2004). The optimality of naive Bayes. AA, 1(2), 3. DOI: 10.4236/oalib.1105005