Strathmore University
SU+ @ Strathmore University Library

Electronic Theses and Dissertations

2018

Stock market price prediction using sentiment analysis: a case study of Nairobi stock exchange market

Victor K. Lwanga
Faculty of Information Technology (FIT)
Strathmore University

Follow this and additional works at https://su-plus.strathmore.edu/handle/11071/5996

Recommended Citation

Lwanga, V. K. (2018). *Stock market price prediction using sentiment analysis: a case study of Nairobi stock exchange market* (Thesis). Strathmore University. Retrieved from https://su-plus.strathmore.edu/handle/11071/5996

This Thesis - Open Access is brought to you for free and open access by DSpace @Strathmore University. It has been accepted for inclusion in Electronic Theses and Dissertations by an authorized administrator of DSpace @Strathmore University. For more information, please contact librarian@strathmore.edu

Stock Market Price Prediction Using Sentiment Analysis: A Case Study of Nairobi Stock Exchange Market

Lwanga Victor Kwome

Submitted in partial fulfillment of the requirements for the Degree of Master of Science in Information Technology at Strathmore University

Faculty of Information Technology

Strathmore University

April 2018

Declaration

I declare that this work has never been submitted for examination in any university.

Student's Name: Lwanga Victor Kwome.

Signature: _____

Date: _____

Approval

The thesis of Lwanga Victor Kwome was reviewed and approved by the following:

Dr. Joseph Orero (PhD)

Senior Lecturer, Faculty of Information Technology

Strathmore University

Signature: _____

Date: _____

Professor Ruth Kiraka

Dean, School of Graduate Studies

Strathmore University

Abstract

Stock market price prediction has become an area of research and interest for several years now due to the many challenges in making accurate price predictions due to the volatility of the data. However, the stock market is not easily predicted. Movement in the stock market is influenced by various factors such as personal fortunes, political events, individual tastes, preferences and natural disasters. People can express all these through their sentiments and opinions on the social media platforms, financial news, and blogs. The stock price does not only rely on the law of demand and supply. People's opinions and moods also have a substantial impact on the movement of the stock prices of a company.

Recently, efforts to increase the accuracy of stock market predictions by including data from social media such as Facebook and Twitter has received a lot of attention. Social media can be regarded as an indicator of sentiments, and these are known to influence the stock market. Current models lack a clear interpretation, and it is also difficult to determine, which data is relevant for stock market prediction since there is an abundance of the same on social media.

This study proposed the use of machine learning algorithms that will be utilized in Natural Language Processing (NLP) to get opinions and sentiments from social media on a particular company's stock to predict the stock market prices. Previous studies show that public mood, opinion, and stock market price have some relation to an extent. The research used Support Vector Machine with bigram feature to perform sentiment analysis which exhibited and accuracy of 83 percent and Artificial Neural Network in Stock price prediction which had a mean squared error of 5.6. This research has proven that sentiment analysis can be incorporated in stock price prediction.

Acknowledgment

I would like give my special thanks to my supervisor, Dr. Joseph Orero, for his guidance and support throughout the project and Dr. Bernard Shibwabo for his assistance during the Research.

I also would like to express my gratitude to my family for their continued support.

List of Abbreviations

ANN – Artificial Neural Network

API – Application Programming Interface

ARIMA – Auto-Regressive Integrated Moving Average

CDSC – Central Depository and Settlement Corporation

DJIA – Dow Jones Industrial Average

EMH – Efficient Market Hypothesis

NLP – Natural Language Processing

NN – Neural Network

NSE – Nairobi Stock Exchange

SVM – Support Vector Machine

S & P 500 index – Standard and Poor 500 Index.

VT OMNES VNVM SINT

Table of Contents

List of Figures

VT OMNES VNVM SINT

List of Tables

List of Equations

Chapter 1 : Introduction

1.1 Background

The prediction of stock market prices and trends is a problem of interest. The pricing of shares on the stock exchange has dynamic behavior often driven by the law of supply and demand for action. This dynamism attracts the attention of investors because it provides huge profits when investments are made in the best way at the right time. According to Rocha and Macedo (2011), investment in the capital market, the objective is always to buy shares when its price is low as possible and sell them when the price is much higher. In this way, expect the behavior of the stock market means generating profits and reduce risk and losses. This anticipation can also be referred to as prediction and can enhance the profitability of investment (Rocha & Macedo, 2011).

Uncertainty is the common characteristic that most stock markets have hence the long-term and short-term future states. Uncertainty in this section is undesirable for existing investors but remains so as this is unavoidable when using stock markets as investment tools. The solution in such scenarios us is the ability to reduce uncertainty levels. The process of reducing uncertainty entails the application of stock market forecasts or predictions. Past studies conducted in this subject relied on historical prices collected form companies following their fluctuation characteristics. The theories of efficient market hypothesis state that the movements in financial markets depend on current events, news and releases of products and their impact of the stock values of different companies. For this reason, the prices in the stock markets follow the random walk pattern resulting in inaccurate predictions of more than 50% (Pagolu et al., 2016).

Business and financial news brings us the latest facts about the industry's stock market. Previous Studies show that both financial and business news have a robust relationship with future stock performance. Therefore, extracting sentiments and opinions from business and financial

news is useful as it may assist in stock-market price prediction. It has been proven that the financial market is "informationally efficient" (Fama, 1965), stock prices reflect information and the price movement is in response to news or events.

As it is well known, emotional state can influence our decisions and no doubt such choices include stock market investment decision (Gilbert & Karahalios, 2010). When people are pessimistic or uncertain about the future, they will be more cautious to invest and trade. So capturing the collective mind of the peoples' mood becomes one possible way to predict the stock movement.

Antweiler and Frank (2004) in their study determined the association between activities in Internet message boards related to stock volatility and relevant trading volume. Gilbert and Karahalios (2010) used over 20 million posts from the Live Journal website to create an index of the US national mood, which they call the Anxiety Index. They found out that when the index rose sharply, the S&P 500 index needed the day marginally lower than expected. It shows there a correlation between how public mood or sentiments, financial news and people's opinion can affect a company's stock price.

Nowadays social media is representing the opinions and sentiments of the public about current events. Sentiment analysis entails having a complete understanding of an author's opinion expressed in a text. Elaborately, optimistic news and opinions in existing social media about a corporation would inspire people to capitalize in the stock of the specific company resulting in the stock prices of the corporation increasing. Behavioral finances hypothesis states that a public mood and any market performance are always correlated. The idea here is that when individuals are happy, in good moods or optimistic, they are most likely to surge investment, which in turn advances stock market price performance. (Makrehchi, Shah, & Liao, 2013)

This study is involves taking existing non-quantifiable statistics on financial news and public sentiments on companies. The collected data is used in predicting the trends on future stocks. The assumption here is that the opinions and news have a significant impact on the changes in the stock markets with an attempt to establish the correlations between public sentiments, opinions, stock trends and company news.

1.2 Problem Statement

Stock market prediction relies on factors such as interest rates, economic activities and related markets that influence the demand and supply of the trading volume. Currently, Stockbrokers who execute trades and advice clients, rely on their experience, technical analysis (price trends) or fundamental analysis in picking their stocks. These current methods are subjective and are usually short-sighted due to their limited capacity to crunch raw numbers. With the value of trade money involved, the improper investment could easily mean great losses for investors, especially if they keep making wrong decisions. Lack of guaranteed returns has also led to the reluctance by potential investors to participate in the market. It is therefore desirable to have a model that can guide on the most likely next day prices (prediction) as a basis for making any investment decision.

This study proposes text mining of financial news and public sentiments and opinions from social media such as twitter. The combination of market data and news features together helps improve the accuracy of predictions. Regardless, already existing systems have failed to effectively integrate news features together with market data. With this, the results obtained are converted into numeric forms that feed the prediction process.

1.3 Objectives

General Objective

This project aims at predicting future price movements of the stock market using financial news and peoples' opinions posted on the social media platforms hence getting sentiments that will aid in stock price forecasting.

Specific Objectives

i. To investigate how the stock market NSE operates.
ii. To analyze the current methods used in the stock market prediction
iii. To evaluate current methods, use in text mining and processing from social media.
iv. To develop a model for stock market price prediction based on sentiment on social media.
v. To test and validate the stock market price prediction model.

1.4 Research Questions

i. How does the NSE stock market operate?
ii. What are the current methods used in prediction of the stock market price?
iii. What are the current methods used in text mining and processing of data from social media?
iv. How to design and develop a model for stock market price prediction?
v. How will the prediction stock market price prediction model be validated?

1.5 Scope and Limitation

The project is limited to only the company's shares listed on the NSE. Additionally, the company should have traded for at least five years to ensure there is data consistency. The

languages to be used are English and Swahili in the sentiment analysis process. Use of slang in this case sheng' and vernacular language will not be considered. The assumption in this study is that there should be no form of manipulation that could have a bigger effect on the prices of stock movements by either the stockbrokers or any other affected parties.

1.6 Justification

Currently, in the Kenyan context, the stockbrokers use methods based on trend patterns which may not be effective. These methods do not have the predictive ability, and they are based on demand and supply. In forecasting the stock market prices, other critical factors will influence the price of the stock and also the demand and supply. Predicting the behavior of shares in the stock exchange is not a simple task as it involves variables not always known and can undergo various influences from the collective emotion to high profile news. By Incorporating the news and sentiments of people who in this case are considered as key external factors that are not presented in a quantifiable format, can aid in providing accurate results in the stock market price prediction.

VT OMNES VNVM SINT

Chapter 2 : Literature Review

2.1 Introduction

This section of the document reviews existing related literature and previous research and studies on predicting the stock exchange market using sentiment analysis and other methods.

2.2 Stock Market in Kenya

2.2.1 Nairobi Stock Exchange (NSE)

The NSE is an African exchange that offers trading facilities to investors looking for exposure to the Kenyan economies; it also lists equity and debt securities. It was founded in 1954, demutualized and self-enlisted in 2014. It operates under the Capital Markets Authority of Kenya (Nairobi Securities Exchange, 2017). The NSE is a member of the (WFE) World Federation of Exchange, which is the founder member of the ASEA (African Securities Exchanges Association) and the (EASEA) East African Securities Exchanges Association. It is also belongs to the Association of Futures Market and is a partner exchange in the United Nations-led SSE initiative (Nairobi Securities Exchange, 2017).

The NSE trades in both shares and bonds. The identified shares in NSE are classified into sectors that are commercial, agricultural, financial sectors, allied sectors and services. The sectors are displayed alphabetically for ease of location by the interested investors who can view daily trading from a public gallery. The bonds traded by NSE are treasury bonds which are issued by the Kenyan government and the corporate bonds which are issued by trading companies. Shares are equities while bonds are referred to debt instruments (Nairobi Securities Exchange, 2017). The NSE also provides five types of live (real-time) data: Real Time Listed Equity Securities Data, Real-Time Listed Debt Securities Data, and NSE Live Ticker on corporation websites, FTSE NSE Equity directories and FTSE NSE Bond Index.

The NSE enables the money markets segment to be productive by providing meetups between lenders of money and borrowers but at a lower cost. The lenders or savers end up being the investors. Their role is to invest in the market and expect financial rewards through profits. The borrowers or issuers borrow money from the market lenders and pay them in form of profits after a set duration. It also provides education to the public and its users regarding the highest profits available on shares and bonds; including the buying and selling process. They also teach the community on how to do investments as a group. It also provides financial answers to the most known problems. The shares and bonds are recognized as guarantees for loans in co-operative societies as well as bank loans. The two can be prearranged, with the assistance of money managers, to pay for children's tutoring fees, hospital bills, car services and other assurance schemes such as pension and retirement plans (Nairobi Securities Exchange, 2017). Through the two, the local government, cooperatives societies, small and big companies, and similar establishments can raise funds to increase their commercial activities, make profits, create employment opportunities and support the growth of the economy.

2.2.2 Capital Markets Authority (CMA) of Kenya

The self-governing public organization which was established through the Act of Parliament, under the finance ministry. The established authority took power on December 15th, 1989, after it was passed and established in office in 1990 (Capital Markets Authority, 2017).

The CMA and associated industry operates in a set regulatory framework guides the actions of the industry players in all the activities. Following its inception; the act has worked towards broadening and deepening the CMA by developing new regulatory frameworks that also facilitate development of new products and services as well as institutions. This is achieved through research and fairness coupled with orderliness in the industry (Capital Markets Authority, 2017).

2.2.3 Central Depository and Settlement Corporation (CDSC)

CDSC provides quality services in settlement and clearing services in the Kenya Capital Markets. It offers a secure central custody that is simplified, safe and swift transfer of investors' value to the right place. One way of boosting investor confidence the in CMA market is through creation of customized solutions that ensure the investor is made aware of all transactions taking place in an individual's central depository system (CDS) account (Central Depository & Settlement Corporation, 2017)

The services offered are; online account access, Investor services- SMS services, email statement services and statements of the accounts. Depository services such as securities accounts, deposits, transfers, pledges, and releases. Clearing and settlement services include guarantee fund, trade reporting, and clearing. Issuer's services; trading rights, AD HOC reports, dividends, bonus issues and initial public offers (Central Depository & Settlement Corporation, 2017).

The Central Depositories Act helps by laying down a regulatory and legal agenda through which the founding and processes of the CDSC are anchored. The organization also functions under the governing oversight and supervision of the CMA. It is a limited liability company permitted by the CMA and authorized to warrant the efficiency of the delivery process, clearing procedures, and settlement of purchases securities in all capital markets of Kenya. In this regard, CDSC is also managed by the Capital Markets Act as well as the rules and regulations (Central Depository & Settlement Corporation, 2017).

CDSC being an integrated financial market infrastructure plays an important role in the competent functioning of both domestic and regional monetary markets. The CDS Rules guide its day-to-day management. The CDSC policies and procedures describe the processes and descriptions of how all the stated functions should be performed by participating parties. Any

changes made in the policies and procedures must always be approved by the Capital Markets authority and the government where necessary (Central Depository & Settlement Corporation, 2017).

CDSC also has the function of entering enters into new contractual relations and agreements with interested stakeholders for the delivery of selected services based on individual preference. The essential elements are the contracts signed by CDSC and CDAs. This is accompanied by new agreements btween settlement bankls and CDSC. The most important agrreement is that one between CDSC and the central bank of Kenya such that all securities transactions are settled here (Central Depository & Settlement Corporation, 2017).

2.2.4 Stock Market Trading

Stock trading is classified as either day trading, medium-term, short-term and long-term based on the duration of the stock holding process. In day trading, the buying and selling of financial instruments usually done on the same selected day with all the trading closed before the end of day. The traders that are hired to trade in the day trading are referred to as day traders or active traders. Short-term trade is that which involves trading of one to few days; maybe a week. Medium-trading is that which takes place in few weeks to months. Long-term trading on the other hand goes on from months to years depending on the need (Zhang, 2013).

2.3 Stock Market Prediction Theories

In any financial derivation, two main principles are considered according to Hellstrom (1998) and Lawrence (2002). One principle is that profit is not generated from anything and the arbitrage principle which states that “no opportunities for arbitrage that is there is no possibility of generating profit without any associated risks”.

2.3.1 Random Walk Theory

This is a theory that works with the conclusion that changes in stock price have the same level of distribution and are always independent of each other. In this case, past movements and trends in stock prices or any markets cannot be used to forecast any future movement Hellstrom, (1998). The theory's formula is:

$$v(t) = v(t-1) + c(t)$$
$$\Delta v(t) = \frac{v(t) + d(t) - v(t-1)}{v(t)}$$

Equation 2:1: Random Walk Equation

Where v (t): the price of stock at the time t

v (t – 1): the price of stock at the time t -1

Δ v (t): change in the price of the stock at time

d (t): dividend at the time t

c (t): adjustment term at the time t

Since c (t) is the actual impact of all the privately and publicly available information on the stocks, which predicts **Δ** v (t) before a difficult task.

2.3.2 Efficient Market Hypothesis

This theory states that all market price mirrors the assimilation of all the information available. When the generated information enters a market, the system enters the unstable state and forecasts the new price eliminates correct change. From the available information, it is

impossible to predict any future prices of the stocks (Burton,2003). The principle is that all investors react immediately to any form of informational advantages available thus eliminating any possible profit opportunities. The prices at all times reflect the available information with the conclusion that no profits are generated from the information-based trading (Lo & MacKinley, 1999).

Fama (1970) mentioned that there exists three forms of competent marketplaces based on the information used to predict a future price. The first one is the weak form; only the historical price or past information is used. The second form is the semi-strong formula which includes past prices as well as the publicly accessible information. Lastly, the strong form which includes both the public and privately available information, it also includes insider information. One should note that efficient market hypothesis and the random walks never amount to the equivalent thing. A random walk in the stock-prices fails to suggest that the stock market prices are resourceful with normal investors. A random-walk is always defined by the fact that prices and their relevenat changes are independent of each other always (Brealey et al., 2005).

2.4 Stock Market Prediction

There exists four stock market prediction methods: Fundamental analysis, Technical analysis, Machine learning and Time series analysis and (NeuroAI, 2013).

2.4.1 Technical Analysis

Technical analysis is the numerical time series methodology used in the prediction of stock markets founded on historical data with charts being introduced as the primary tools (Pring, 1991).It is a technique used in the evaluation of securities by interpreting the figures produced by activities in the stock market such as previous prices and volume traded. The aim of using technical

indicators in prediction is to get the trends and patterns, which should then inform a direction of future prices.

This method has three major assumptions that are taken into account. The first assumption is that the market discounts everything. Technical analysis has been heavily criticized for not considering the fundamental factors as it only takes the price movement into account. Technical analysts have confidence in that everything from a corporation's fundamentals to the broad market is already priced into the stock hence there is no need to consider other influencing factors. The other assumption is the trend in price movement. In technical analysis, the price movement is believed to follow certain trends. It implies that once the trend is established, the future price movements are likely to follow the same trend. The last assumption is that the past has a habit of repeating itself regarding price movement. The repetitive nature of all the movements in market prices is attributed to all market psychologies, which tend to be very foreseeable based on the human emotions such as fear and excitement. Technical analysis also uses chart and graph patterns to analyze market movements and understand trends (Huang et al., 2011). Technical analysis process deals with past price movements to forecast a pattern that guides future investment decisions. All these indicators must be calculated and their values used for guiding the prediction.

2.4.2 Fundamental Analysis

The fundamental analysis process is referred to as a study of different factors affecting the supply and demand (Thomsett, 1998). The theory works with the idea that data assembly and its consequent interpretation is the foremost process involved in the prediction of the stock prices. The trading opportunities of the analysis uses the gaps between the existence of a new event and the consequence market responses towards the event. The central data used in the process of fundamental analysis is company data including annual reports, quarterly reports, balance sheets,

income statements and auditor's reports. News in the industry plays an important part in the analysis process as such news also reflect the existing supply and demand chains in the marketplace. Fundamental analysis tends to have an overview of the company from a top-level view and considers issues such as political, economic and the business environment of the company. The general requirement of fundamental analysis is to understand the company and decide on its prospects (Thomsett, 1998).

2.4.3 Time Series Method

Time series method utilizes past performances to forecast on a time-series measure. The time series is referred to as a system of experimented quantities from a selected observations, whereby discoveries such as a periodic dissemination can be established (Zhang, 2003). Other important methods in the time series prediction are auto-regression, linear regression, as well as ARIMA. A significant characteristic of time series data is the fact that it is dependent on time. For this reason; current observations have to depend on a past explanations in time. A typical prediction model, in this case, requires information external to the particular stock, which can be used to extrapolate the performance of the stock in question. Such information should be having a bearing on the stock of study. In time series forecast mathematical data series are placed successively, they take place in equivalent periods. In the method, there exists chains of numbers consisting of normal periods for a fixed time duration (Pang et al., 2002).

2.4.3.1 Linear Regression

Linear regression is a model that attempts to establish a connection between any two variables by accurately fitting a new linear equation to any collected or observed data (Pang et al., 2002). It can also be fitted with a quadratic equation, and still, it will be called linear regression.

In this context, one of the variables is the explanatory variable while the other is the dependent variable. A linear regression line with a linear equation is of the form: y = a + bx,

Where x; is the identified explanatory variable and y is the identified dependent variable. The slope of the line is b, and a is the intercept.

2.4.3.2 Auto-Regressive Integrated Moving Average

Zhang (2003), states that in an ARIMA model, the next future value of the variable is assumed to be the linear function of several historical observations as well as random errors. It can be represented in the form

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q}$$

Equation 2:2: ARIMA Equation

Where y_t is the differenced time series value, ϕ and θ are unknown parameters and e are independent identically distributed error terms with zero mean. Here, y_t is expressed as its past values and the current and past values of error terms.

The ARIMA model tends to combine three basic methods that are the autoregression (AR), Differencing (I-for Integrated) and Moving Average (MA). For auto-regression algorithm, the digits of the given time-series facts are regressed to their own lagged values, that is indicated by the “p” values in the model. Differencing this comprises of differencing the time series data to eliminate the identified trend and converting a non-stationary time series to a new stationary one. The “t” value indicates this in the model the moving average nature of the model is represented by the “q” value which is the number of lagged values of the error term.

2.4.4 Machine Learning

Machine learning is the last method and is extraordinary for AI solutions considering it is based on the principles of learning from continuous training and practices. The association models such as artificial neural networks are well fit for machine learning where new association weights are adjusted to progress the competence of a formed network.

2.4.4.1 Artificial Neural Network

The bio-inspired ML model has shown incredible success in the application and fields of artificial intelligence. Many scholars have shown that using the bio-inspired algorithm has improved the results of the research domain. The algorithms include artificial neural networks (ANN), artificial immune systems, evolutionary computation, fuzzy systems, and swarm intelligence (Andries, 2007).

An artificial neuron network takes the model of a biological neuron. The artificial neuron accepts signals or inputs from the other neurons or the surrounding environment. The signal will be fired given certain conditions, thus, transmitting the signal to all other connected neurons (Uhrig, 1995). Figure. 2:1 below is a representation of an artificial neuron. Here, there is an association between the numerical positive and the negative value which is associated with each neuron such that they either inhibit or excite inputs with each connection made to the artificial neuron. The activation functions in ANN are used to regulate the firing taking place in the artificial neuron. The neuron then collects all incoming signals by computing their net input signals as a function with the associated or given weights. These net input signals then serve as input to the activation function which calculates the output signal of the artificial neurons (Zupan, 1994). An ANN is a layered system containing of one or many artificial neurons. ANN components include the input layer, hidden layer and the output layer. Based on the interconnection of the components;

the ANN has been modelled with the ability to perform learning, generalize and map abilities to process information in parallel.

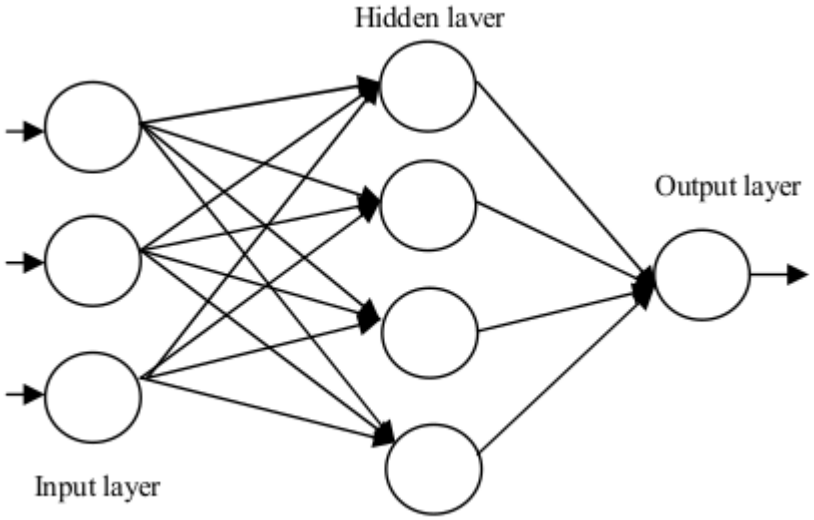


Figure 2:1: Artificial Neural Network

Several ANN architectures have been developed such as feedforward neural network, recurrent neural network, and spiking neural network. Also, there are also different types of neural network such as single-layer neural network, multi-layer perceptron (MLP), temporal neural network, radial basis neural function network, self-organizing neural network (Peterson & Rögnvaldsson, 1992).

2.5 Sentiment Analysis

A sentiment is a feeling, opinion or emotion that is formed by a person towards something, an idea or someone. Sentiment analysis in the studies is referred to as attitude mining, opinion mining studies the sentiments of people towards certain ideologies (Fang & Zhan, 2015). There exists different sentiment classification techniques machine learning approach, hybrid approach and lexicon-based approach and (Maynard & Funk, 2011). The ML approach employs the sue of

ML algorithms and simple linguistic features. In lexicon-based approach, the algorithm relies on different sentiment lexicons which are a collection of precompiled and known terms. This is classified into dictionary-based approaches and corpus-based approaches that utilize statistical or semantic techniques to find sentiment polarity.

2.5.1 Unsupervised Classification

Unsupervised learning has no definite explicit target outputs associated with the input, and consequent learning through manual observation. The purpose is to have the existing machine learn without giving any obvious instructions. The eminent approach to most unsupervised learning techniques is clustering, whereby similarities of features in the training data is discovered. Cluster resemblance parameter is well-defined upon common metrics such as the Euclidean distance. K-means, Gaussian mixture models, Hierarchical, Hidden Markov models and Self-organizing maps are examples of clustering algorithms (Batool et al., 2013).

2.5.2 Supervised Learning

2.5.2.1 Rule-based Learning

The rule-based learning classifier is based on the rule of incidences of sentiments in a text. If any selected word contains positive emotions, then the conclusion is that it is positive. If the word contains any negative emotions, the conclusion is that it is negative. The rule-based classifier has similarities with the fuzzy logic system that allows intermediate value to be well-defined between conventional evaluations like yes or no, true or false and others. (Bhardwaj et al., 2015)

2.5.2.2 Support Vector Machines

Support Vector Machines (SVM) or Support Vector Networks (SVN) are classification and regression examination techniques. Support vector machines are categorized as supervised

learning models for information analysis as well as pattern recognition. The common application areas for the SVM algorithms include image processing, bionformatucs and text analysis. The support vector machine constructs a hyperplane or a set of hyperplanes in a high or infinite-dimensional space. In many cases, the data is not linearly separable. With the use of an SVM learning algorithm, it is probable to create a room that is transformable. The model signifies the examples as points in space, maps separate categories and divides them as much as possible. The goal is to design a hyperplane that classifies all training vectors into two distinct classes, where the best choice is the hyperplane that leaves the maximum margin for both classes. (Platt, 1999) Recent research and state of the art approaches of Support Vector Machines show that using ensemble approaches can drastically reduce the training complexity while maintaining high predictive accuracy. This has been done by implementing the SVM without duplicate storage and evaluation of support vectors, which has been shared between consistent models (Marc et al. 2014).

2.5.2.3 Naive Bayes methods

The Naive Bayes methods are a set of supervised learning algorithms that is used for clustering and classification (Lowd & Domingos 2005). Methods employed are based on the application of Thomas Bayes' theorem with a simple assumption of there being an independence between every pair of selected features. The Naive Bayes classifiers are known as linear classifiers and are able to perform well, simply and are very efficiently (Zhang, 2004). For small sample sizes, naive Bayes classifiers can outperform more powerful alternatives. However, non-linear classification problems can lead to poor performances of naive Bayes classifiers. These methods are used in a several of different fields such as diagnosis of diseases, classification of RNA sequences in taxonomic studies and spam filtering in e-mail clients (Raschka, 2014). Research of Naive Bayes has previously been proved to be an optimal method of clustering and classification,

no matter how strong the dependencies among the attributes are. If the dependencies distribute evenly in classes or if they cancel each other out, Naive Bayes performs optimally (Zhang 2004). Recently, Naive Bayes theorem has been applied to image classification algorithms, where the Local Naive Bayes Nearest Neighbor algorithm increases classification accuracy and improves its ability to scale to bigger numbers of object classes. (Lowe, 2012)

2.5.2.4 Decision tree classifiers.

Decision trees employ a hierarchical decomposition of training data in which certain conditions on an attribute value are used to classify and divide data (Quinlan, 1986). The predicate and conditions used implies the absence and presence of more than one word. In decision trees; the process of dividing data takes place recursively until all the leaf nodes have a minimum number of records that show a detailed classification.

2.5.3 Lexicon Based Approach

The lexicon-based approach aims at attaining an effective cross-domain performance. The methodology works with the assumption that the total of the sentiment orientations of all words make contextual sentiment orientations. Here, words that are opinionated are used on the classification tasks. The positive opinions are employed in describing desired states while the negative opinions express the undesired states. There exists several opinion idioms and phrases that are known as lexicons. There exists different approaches when it comes to compiling and collecting the opinions used in the word list. Since the manual approach is time consuming; it is used together with other faster approaches that are automatic with the aim of checking out for any errors and mistakes. The common approaches are discussed in the section below (Bhardwaj et al., 2015):

2.5.3.1 Dictionary-Based Approach

In a dictionary based approach; sets of words or opinions are collected manually based on set subjects. The same grows through searching for more words in a corpora WordNet or thesaurus for simple synonyms as well as antonyms. The words obtained are often added to a seed-list resulting in the formation of new iterations. The process of iteration stops when the system fails to find new words. At the end of the process; there is manual inspection done with the purpose of eliminating any existing errors. The key disadvantage of this approach is the inability to find any opinions or words that are context or domain specific orientations (Mohammad et al., 2009).

2.5.3.2 Corpus-based Approach

The Corpus-based method helps to resolve the difficulty of finding opinion words given context-specific orientations. Its methods depend on syntactic patterns or patterns that occur together along with a seed list of opinion words to find other opinion words in a large corpus (Medhat et al., 2013). The Corpus-based approach tries to find co-occurrence patterns of words to determine their sentiments. This approach is based on seeding list of opinion words and then find another opinion words which have a similar context. This method is used to assign happiness factor of words depending on the frequency of their occurrences in "happy" or "sad" blog post (Bhardwaj et al., 2015).

2.5 Twitter Analysis

For any platform to be feasible as a predictor of stocks, the platform itself must be appropriate for the gatheoring of data. Twitter offers a comprehensive search API, up to seven days back in time, but also offers the opportunity to query against tweets in real-time, through its streaming API (Arafat et al., 2013). The Twitter API is convenient since it removes the need to

batch data gathering and management, and offers a whole new aspect to Stock Prediction due to the high accessibility of data. A major drawback using the Twitter Search API is the limitation on complexity where overly complex queries are restricted, and the limitation on the availability of data older than a set number of days, seven days to be precise. This is because the Search API makes use of indices that only contains the most recent or popular tweets, according to the developer's page on the Twitter website (Twitter, 2017). Furthermore, it is explained that the Twitter Search API should be used for relevance and not completeness and that some tweets and users might be missing in the query results.

The Twitter Search API Developers Page propose that the Streaming API is more suitable for completeness-oriented queries which would be the case of gathering data for the Sentiment Analysis where high completeness is required to analyze the whole picture rather than specific chunks of data (Twitter, 2017). The Streaming API is also favored by existing research on the subject (Choi & Varian 2012).

2.6 Empirical Review

2.6.1 Prediction of movies success using sentiment analysis of tweets

The researcher tried to predict the popularity of movies from twitter sentiment analysis on the movies. He manually labels tweets to create a training set and train a classifier to classify the tweets into positive, negative, neutral, and irrelevant. He further developed a metric to capture the relationship between sentiment analysis and the box office results of movies. He finally predicted the Box Office results by classifying the movie as three categories: hit, flop, or average (Jain, 2013). The prediction was of eight movies which had just been released.The prediction outcome were five movies to be a hit and one to be a super hit, one to be average and he could not determine the success rate for one due to it data unavailability. Comparing his prediction results with box

office results he found his prediction model to be exact for in four cases; a case was on the borderline between hit and average and for another one he could not find data to check the prediction confidence. (Jain, 2013).

2.6.2 Other Related Works

Kihoro and Okango (2014) used an artificial neural network (ANN) model in predicting stock market prices of Equity Bank in Kenya. They used the company's historical data then fitted it in an ARIMA model to identify the best input lags into the ANN model. The best combination of the lags was taken as input lags. The historical data used was obtained from Nairobi's Security Exchange financial and investment segment, comprising 487 daily share price for the bank. They observed that ANN could effectively model the stock market prices. The model was able to discover non-linear relationship in the data which was evident in the fact that the mean-squared misclassification between the predicted share price and the desired share price was very minimal. The ANN architecture gave the best results in terms mean squared error.

The proliferation of documents online and user-generated texts led to the recent growth in exploration in the field of sentiment analysis and their relationship with financial markets. The authors discuss the application of Twitter as a corpus for the achievement of sentiment analysis. The discussion is on the methods used in the gathering and processing of tweets from twitter. The writers use emoticons to formulate a training set used in sentiment classification; a machine learning technique that significantly reduces physical tweet tagging. The training set in the study was split into two sets of positive and negative samples that were based on sad and happy emoticons. Additionally, they analyze a few accuracy improvement methods. Similarly, Albert and Eibe (2011) presents an interesting discussion on streaming the Tweet mining process and the process of sentiment extraction as well as opinion mining. Bollen and Mao (2010) offered the

primary indications that there may be an existing correlation between stock market prices and Twitter sentiments. In the study, a sentiment result is connected with the DJIA and fed into an ANN algorithm to predict future market movements. The study uses a mood-tracking instrument; Opinion-Finder to find the mood in six dimensions (Alert, Calm, Sure, Kind, Vital, and Happy). Thereafter, they relate the mood-time-series with DJIA final values by means of a Self-Organizing Fuzzy NN. Using the techniques, the researchers measured a possible improvement in DJIA's prediction accuracy. After successful publication, the paper launched abundant of the research in the determining the relationship between Twitter and existing market sentiments.

2.7 Conceptual Model

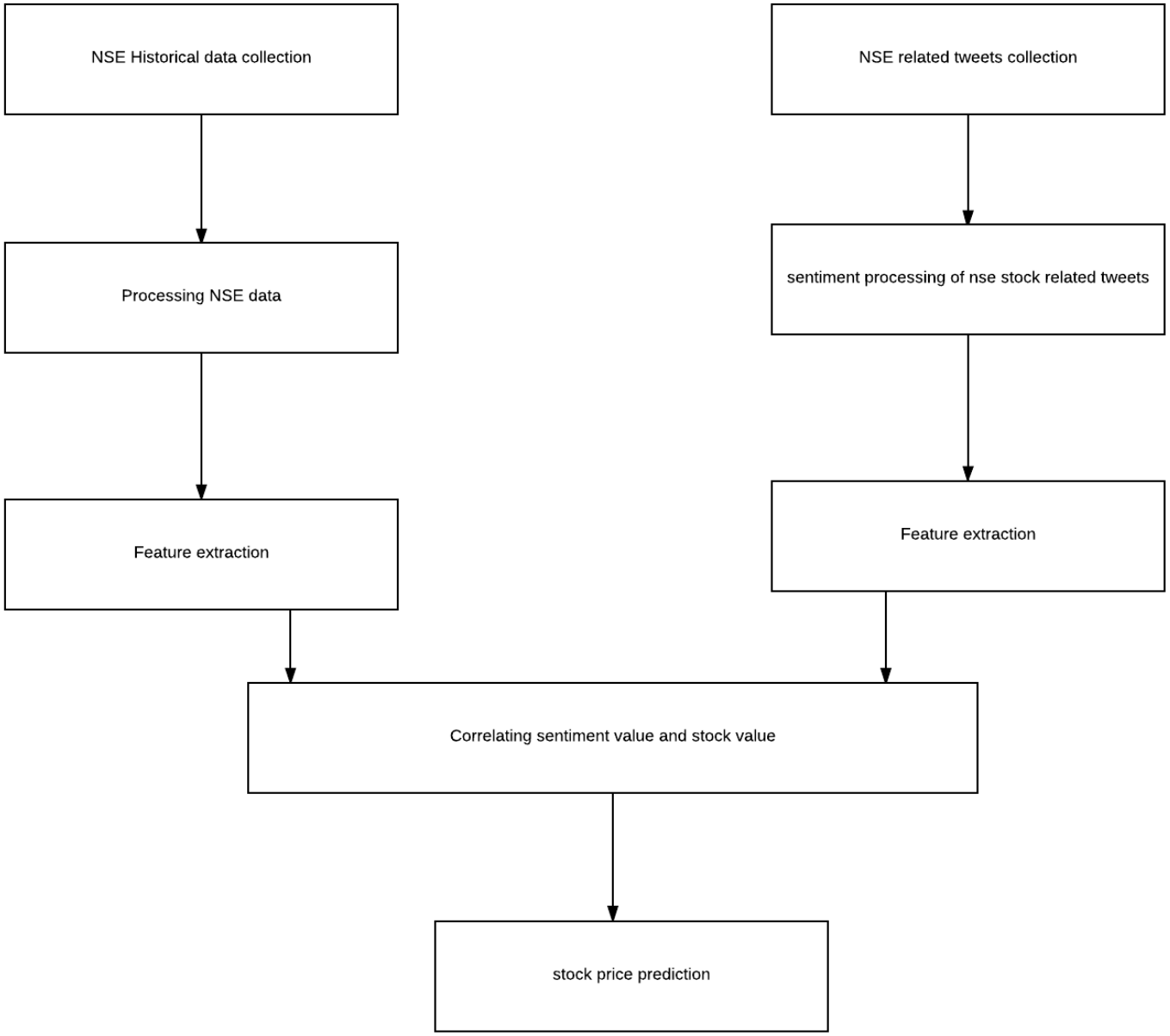


Figure 2:2: Conceptual Model

Data mining of the mood, opinion, and financial news will be retrieved from Twitter. Then the tweets will be processed for them to be classified to be as either positive, neutral or negative. Lastly, the classified tweets will generate the overall mood and use the previous prices the model will be able to predict the future prices. The historical data or NSE will be collected then processed to extract features for prediction. The association between the values of stock and the sentiment value will be generated which will later be used in our model for stock prediction.

Chapter 3 : Research Methodology

3.1 Introduction

According to Bhatnagar and Singh, research methodology can be defined as the process of systematically solving problems. This project will use experimental research to create an artificial intelligence tool based on a model and test its performance on a practical problem. The tool will have a model that will learn from previous shares prices, sentiments from the public and financial news found on social media, in this case, Twitter, that will be used to predict the daily prices for future of a particular stock. The project's objectives stated in chapter one will guide the research.

3.2 Research Design

The research design is the arrangement of conditions for collection and analysis of data in a manner that aims to combine relevance to the research purpose with economy in procedure. It constitutes of the conceptual structure within the research conducted. This research will take an experimental design approach to develop the different models for predicting the stock prices with different data samples which can give the best performance and thus the best result. For purposes of designing and evaluating the model, the research will need data from a typical stock exchange market NSE, and sentiment and opinion data from Twitter

3.3 Target Population and Sampling

The target population is defined as the total number of units in a study environment from which a sample may be selected. In this study, Twitter posts and comments related to stock market prices and companies will be used.

3.4 Data Collection

The process of data collection in sentiment analysis entails collecting of sentiment related data. Different statistical learning methods, adequate data sets for tweets and stocks are necessary.

3.4.1 Twitter Data

For tweet collection, Twitter provides a rather robust API. There are two possible ways to gather Tweets, using the Streaming API or the Search API. The Streaming API lets users obtain real-time access to tweets from an input query. The user first requests a connection to a stream of tweets from the server. Then, the server opens a streaming connection and tweets are streamed in as they occur, to the user. A limitation of the streaming API is that one cannot specify the language, i.e., English or Swahili.

3.4.2 Python

The programming language that will be used for collecting the data through the Twitter Streaming API is python. Python is a programming language commonly used for statistical computing and computer graphics. Data miners and statisticians extensively use it for data analysis. The reason why python is chosen for computing the data is primarily its powerful tools and large community. Python programming also has vast libraries for performing statistics and machine learning.

3.4.3 Stock Data Collection

The stock data will be collected using web scraping, which is the act of extracting information from the web. The data will table the format shown below.

Date	Company	Lowest Price of the Day	Highest Price of the Day	Closing Price	Previous Day Closing Price	Volume Traded
1/3/2017	Eaagads Ltd	25.5	25.5	25.5	25.5	2500
1/3/2017	Nation Media Group	83	88.5	85	87	15500
1/3/2017	Standard Group Ltd	18.75	18.75	18.75	18.75	---
1/3/2017	Centum Investment Company Ltd	34.5	35.75	34.5	35	13600

Table 3:1: NSE Data

3.5 System Development Methodology

According to Berman (2006), with Rapid Prototyping, also known as Rapid E-learning, learners or subject matter experts interact with prototypes and instructional designers in a continuous review and revision process.

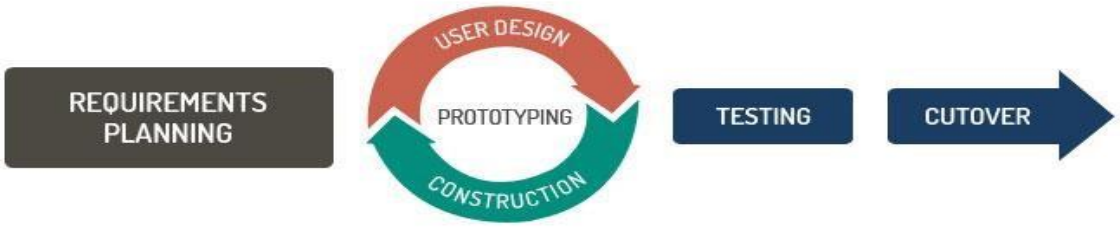


Figure 3:1: RAD Development Cycle

Phases of RAD Development

(i) Planning Phase- It involves getting the system requirements and doing a quick analysis. In this phase, all the tools and necessary materials will be gathered, and the plan for the whole process will be created.

(ii) Prototyping phase- In this phase, the designs and models for the prototype are created. The development follows after.

(iii) Testing- This phase involves the validation of the models created. Unit, integration and system testing are done.

(iv) Cutover- This is the final phase which involves the including data conversion and deployment of the system. The development of a prototype is the first step and analysis is continuous throughout the process. This strategy has many potential benefits including reduction in production time and the cost of late development revisions (Jones & Richey, 2000).

3.6 Research Quality

Validity is the degree to which a concept is accurately measured in a quantitative study. The second element in measuring the research quality is reliability or accuracy. This is the extent to which a research instrument consistently has the same results if it is used in the same situation on repeated occasions (Heale & Twycross, 2015).The reliability of the sources of information of the data that will be used in the study, the research instruments, and any other concerned research aspect will be guaranteed and accredited. In all datasets, there will be no missing values because all companies' stock prices are posted daily. In the validation of the model, a confusion matrix will be used. A confusion matrix contains information about actual and predicted classifications done

by a classification system. Performance of such systems is commonly evaluated using the data in the matrix (Kohavi and Provost, 1998).

		Actual class	
		Positive sentiment	Negative sentiment
Predicted class	Positive sentiment	TP	FP
	Negative sentiment	FN	TN

Table 3:2: Confusion Matrix

Evaluation Metrics of the confusion matrix according to Kohavi and Provost (1998)

True positives (TP): is the number of correct predictions that an instance is positive.

True negatives (TN): is the number of incorrect predictions that an instance is negative.

False positives (FP): is the number of incorrect of predictions that an instance positive.

False negatives (FN): is the number of correct predictions that an instance is negative.

Accuracy is the percentage of the total number of predictions that were found to be correct. It is determined using the equation:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Equation 3:1: Accuracy

Misclassification Rate: Overall, how often is the model wrong?

$$\text{Misclassification Rate} = \frac{\text{FP+FN}}{\text{Total}}$$

Equation 3:2: Error Rates

This is also known as the error rate

True Positive Rate: is the proportion of positive cases that were correctly identified, as calculated using the equation:

$$\text{True Positive Rate} = \frac{\text{TP}}{\text{Actual Yes Values}}$$

Equation 3:3: Recall

This is also known as sensitivity or recall.

False Positive Rate: is the proportion of negatives cases that were incorrectly classified as positive, as calculated using the equation:

$$\text{False Positive Rate} = \frac{\text{FP}}{\text{Actual No Values}}$$

Equation 3:4: False Positive Rate

Precision: is the proportion of the predicted positive cases that were correct, as calculated using the equation:

$$\text{Precision} = \frac{\text{TP}}{\text{Predicted Yes Values}}$$

Equation 3:5: Precision

Chapter 4 : System Design and Architecture

4.1 Introduction

This chapter reviews the proposed architecture, analysis and design of the stock market price prediction model. The system design and architecture was achieved through UML diagrams; use case diagram, sequence diagrams and the data flow diagrams. The diagrams provide detailed descriptions of the components of the proposed system and their interaction at each level.

4.2 Requirements Analysis

Based on the objectives as well as the user requirements, this section outlines the various requirements to be met in the research.

a. Functional requirements

These are functions or processes the proposed system and its components must perform. They are a definition of what users of the system expect form it. For the system, the functional requirements include:

a. The system should allow a user to select the stock to be predicted.
b. The system should crawl the historical and current price on a company's stock price
c. The system should retrieve financial news and sentiments pertaining the company's stock from twitter
d. The system should perform preprocessing of the tweets to clean then and store them in a comma separated values (csv) file.
e. The system should be able to generate an approximate share price for the next trading day

b. Non-functional requirements

Unlike the functional requirements, non-functional requirements place constraints or limits in how the proposed system will achieve its functional requirements. They describe how well the system does its functions and are classified based on the needs of the users. The non-functional requirements of the system include:

a. Usability- The intended users of the proposed system are the stock brokers from different accredited trading firms. The interaction with the system will be simple to allow stock price prediction.
b. Reliability - The reliability of the model will highly depend on the accuracy of the data collected (stock). As this data will be used to train the model which will be used in prediction.
c. Interoperability – This is the degree to which the developed system will be able to facilitate of couple the different interfaces with other systems.
d. Response time – this is defined as the time between the end of a request by a user and start of the response. For the proposed system, the response time should be fast.
e. Scalability – This describes the degree in which the system is able to expand its processing or functional capabilities outward or upward with the aim of supporting business growth and user requirements.
f. Persistent storage- the proposed system components and devices should be able to retain data or information after device's power have been shut down or eliminated.

4.3 System Architecture

System architecture outlines the structure of a system and its behavioral components. The proposed stock prediction system comprises of the classifier, the machine learning predictor, pre-

processor component and historical stock prices data. The only users of the system will be the stockbrokers for the different companies. The stock price prediction begins with the user or stockbroker entering keywords to retrieve tweets related to stock prices of a particular company. Once the tweets have been obtained, the twitter search API matches the keywords and sends them to the pre-processor for cleanup. The processed tweets are transformed into a document-term matrix that is suitable for machine learning algorithms and the tweet classifier. Based on the Machine Learning algorithm employed in the study; the tweets are classified as either positive (1), negative (-1) or neutral (0). Based on the classification, the tweets are matched against historical stock prices in the database hence the prediction of the stock price for the period under review. The result is then presented to the stockbroker for analysis or decision making. The figure 4:1 below illustrates the architecture of the proposed system:

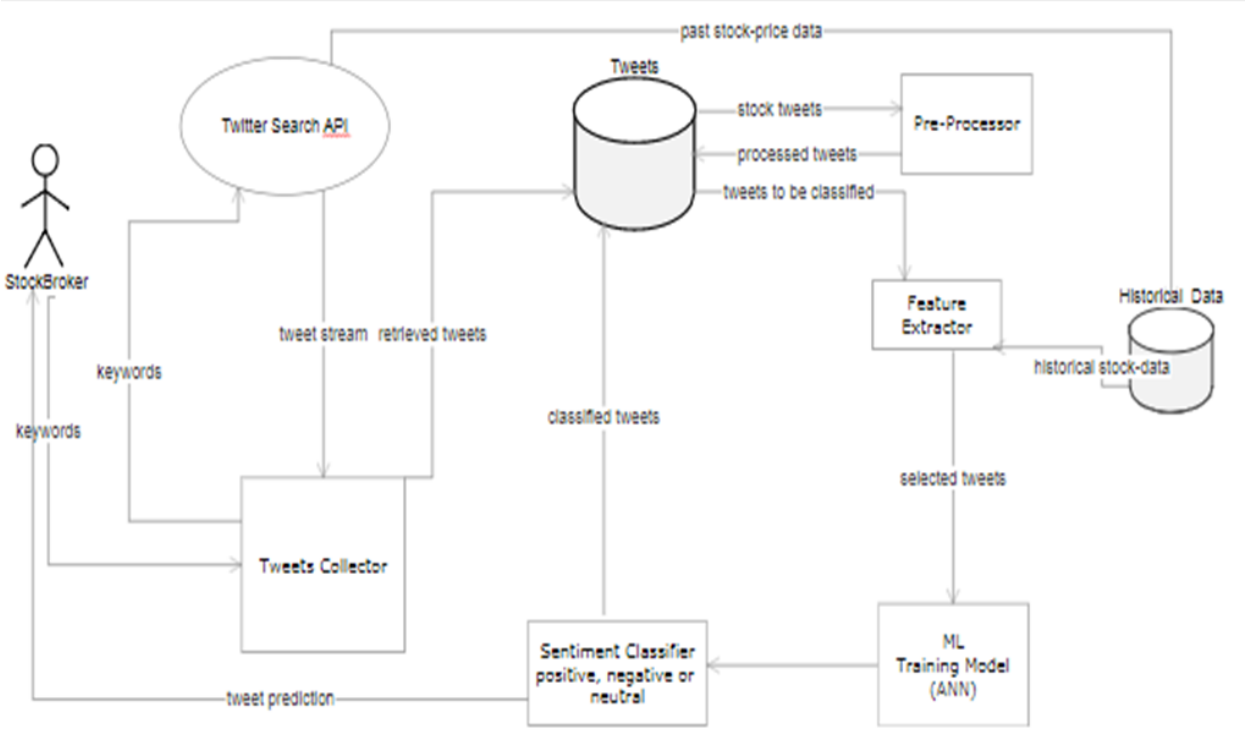


Figure 4:1: System Architecture

4.4 System Analysis

4.4.1 Use Case Diagram

The use case diagram in the analysis phase is used to describe the interactions between the system users and system itself. The most common relationships captured in a use case diagram are those between the actors, use cases and system. In the stock prediction model using sentiment analysis, the actors in the system are the stockbrokers, twitter search API, Historical data Module and the prediction model as illustrated below I figure 4:2:

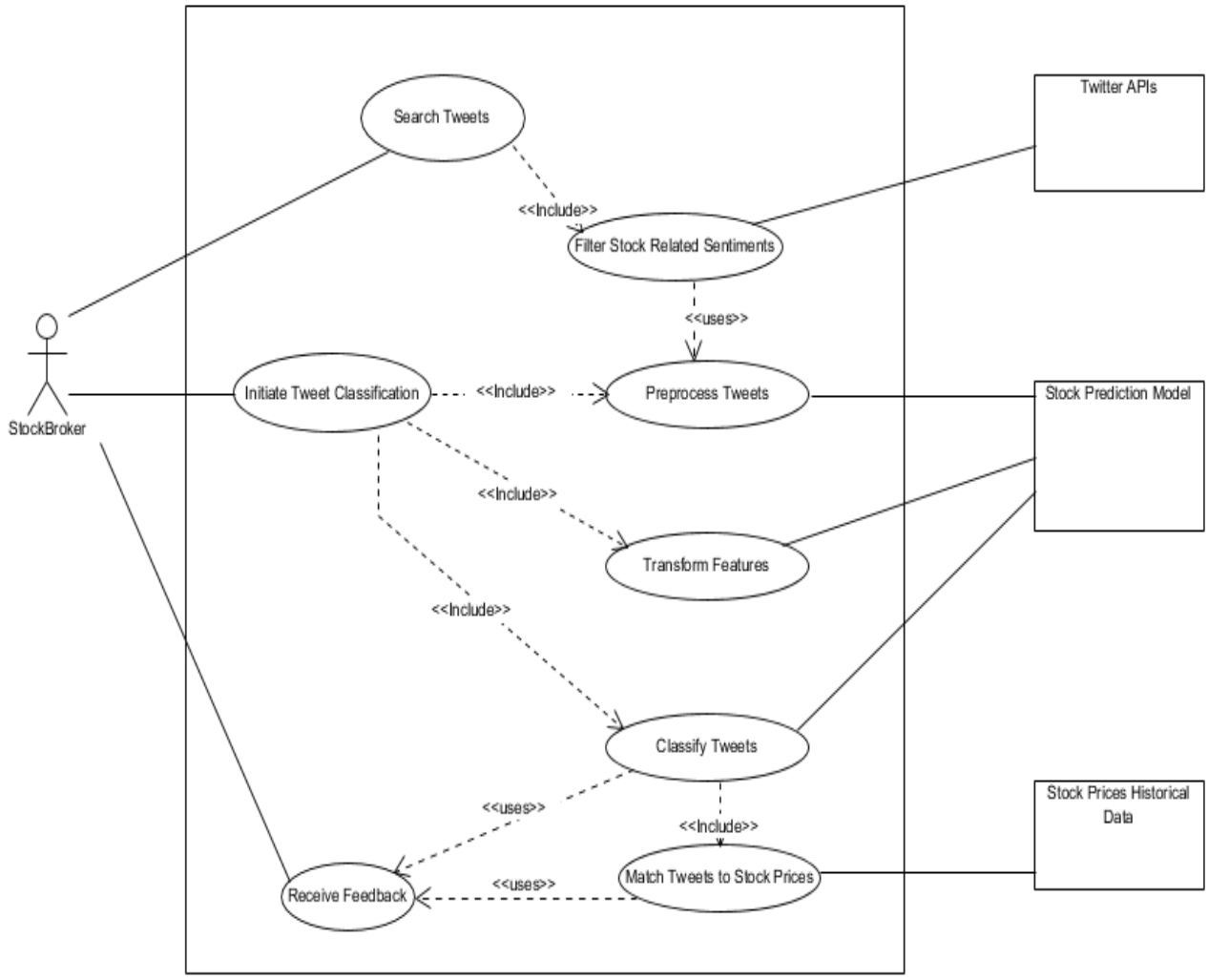


Figure 4:2: Use Case Diagram

4.4.2 Use Case Scenarios

Also known as use case narratives, is a detailed text-based and step-by-step dialogues and interactions between the actors and the system. In system analysis, the use case narrative is used to explain a complete business transaction successful or unsuccessful. The use case narrative for the proposed system is as below:

a. Scenario 1: Crawl Company's Share Prices

Use Case: Crawl Company's Share Prices

Primary Actor: User/Stock Broker

Precondition: The Company selected is listed in the NSE

Post Condition: System fetches the company's daily share price and saves them in a CSV file

Main Success Scenario:

Actor	System Responsibility
1. User enters the company whose share price is to be crawled	
	2. The system takes the company's name and crawls the share/stock price
	3. The system retrieves the share prices of the company
	4. The share prices are saved in a CSV file

5. Views the CSV file with the collected tweets

Extensions:

At any time the system fails to retrieve tweets: the user must confirm that there is internet access and restart the system

b. Scenario 2: Predicting Stock Prices

Use Case: Predict company stock prices

Primary Actor: User/Stock Broker

Precondition:

1. Company's financial news tweets are stored successfully in a CSV file
2. Company's share price data are stored successfully in a CSV file

Post Condition:

1. The system fetches the financial news tweets and the company's and uses them to predict next day and future share/stock prices of the specific company.

Main Success Scenario:

Actor	System Responsibility
1. The user selects the company share price to be predicted	

2. The system fetches the financial news tweets and performs the classification of each day's financial news to either positive(1), negative (-1) or neutral (0)
3. The system combines the aggregated sentiment and historical share price and uses the model to predict the future share/stock prices.
4. The company saves the predicted stock/share price
5. The user views the predicted share price

Extensions:

At any time the system fails to provide predictions user should repeat the process until successful prediction of stock prices.

c. Scenario 2: Search for Financial Sentiment News

Use Case: Search for Financial Sentiment News

Primary Actor:

1. User/Stock Broker
2. Twitter Search API

Precondition:

1. The company whose financial sentiment news is being retrieved is set or determined and listed on the NSE

Post Condition:

1. System fetches financial sentiment tweets of the related company using the twitter search API

Main Success Scenario

Actor	System
1. The user enters the keywords to be used to find financial news about the company	
	2. Passes the keyword entered to the twitter search API
	3. Retrieves tweets from twitter search API based on the keyword
	4. Saves tweets in a CSV file
5. Views CSV file with the collected tweets	

Extensions:

At any time the system fails to load or provide information regarding stock price and company sentiments; the user should cancel and restart the process or repeat the same for clarification

4.4.3 Sequence Diagram

Sequence diagrams depict the chronological flow of events in the system. In essence they describe communication and relationships between objects together with messages that trigger the communications. The user or stockbroker enters keywords that are used as search parameters for Twitter through the web platform. When the keywords are obtained, the system passes them through a Twitter Search API which returns results that are later saved into a CSV file. The stockbroker initiates the classification of the sentiments or tweets related to the company's stock or share price.

The web platform passes a message cleanup_tweets() to the preprocessor which processes the retrieved tweets and returns tweets that match that of the company or stock prices. The obtain_tweet_features() message is sent to the feature extractor component. The result is passed to classify_tweets() message that classifies the tweets as either positive (1), negative (-1) or neutral (0) about stock market pricing of a Company. The system loops through the clean_tweets(), obtain_tweet_features() and classify_tweets() based on user requests. The process ends when the stockbroker requests for the results of the classification for the tweets received.

The sequence flow of events in the proposed system is as in figure 4:3:

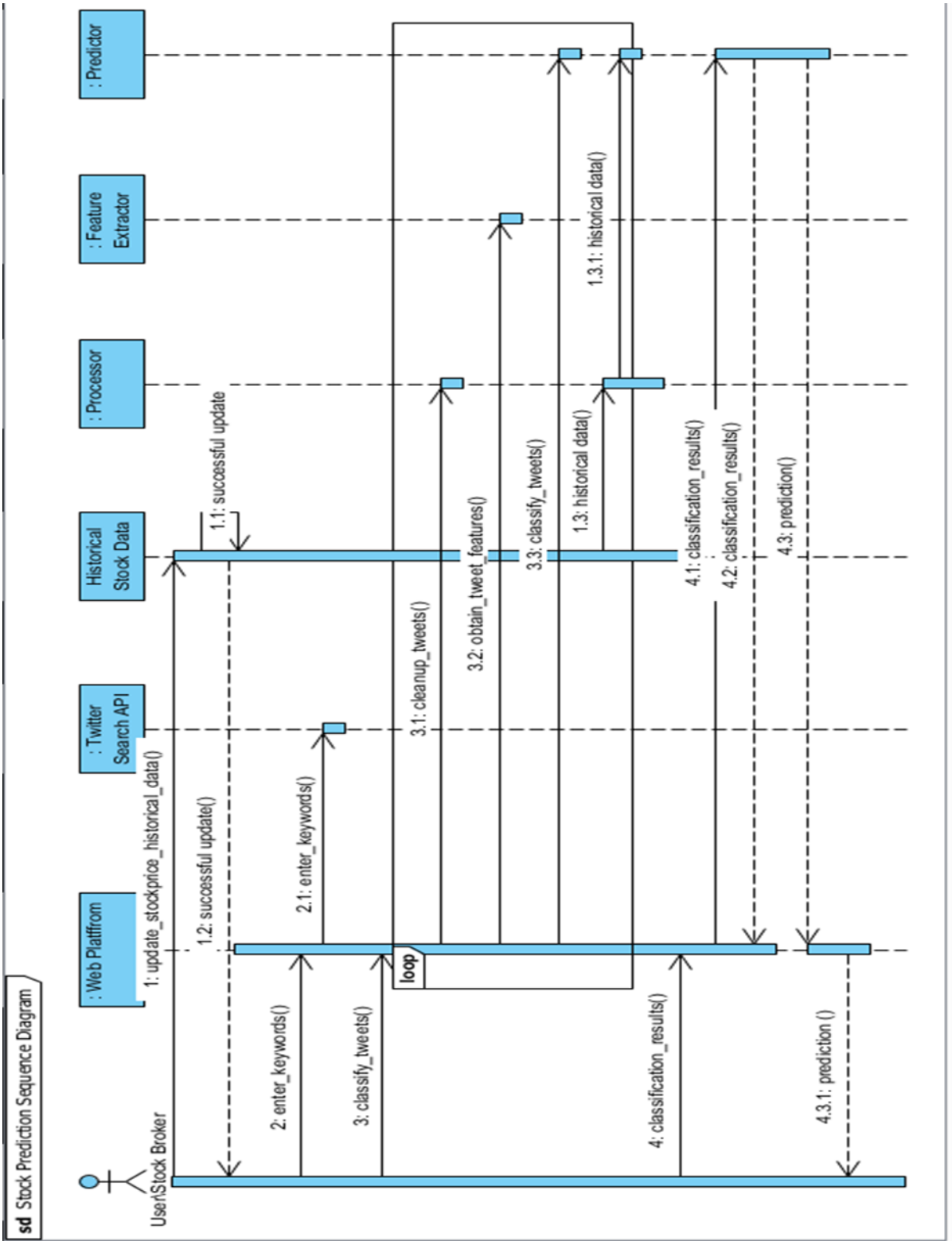


Figure 4:3: Sequence Diagram

4.5 System Design

4.5.2 Context Diagram

These are graphical representations of the flow of data through the information system. A DFD shows the flow of data from external entities into the system as well as movement of data from one process to another. Context diagrams is a DFD that represents the scope of a system consisting of external entities and system boundaries together with information flows between the system and the external entities. The context diagram shows participants or entities that will interact with the proposed system. The figure 4:4 below shows the context diagram of the stock market price prediction using sentiment analysis.

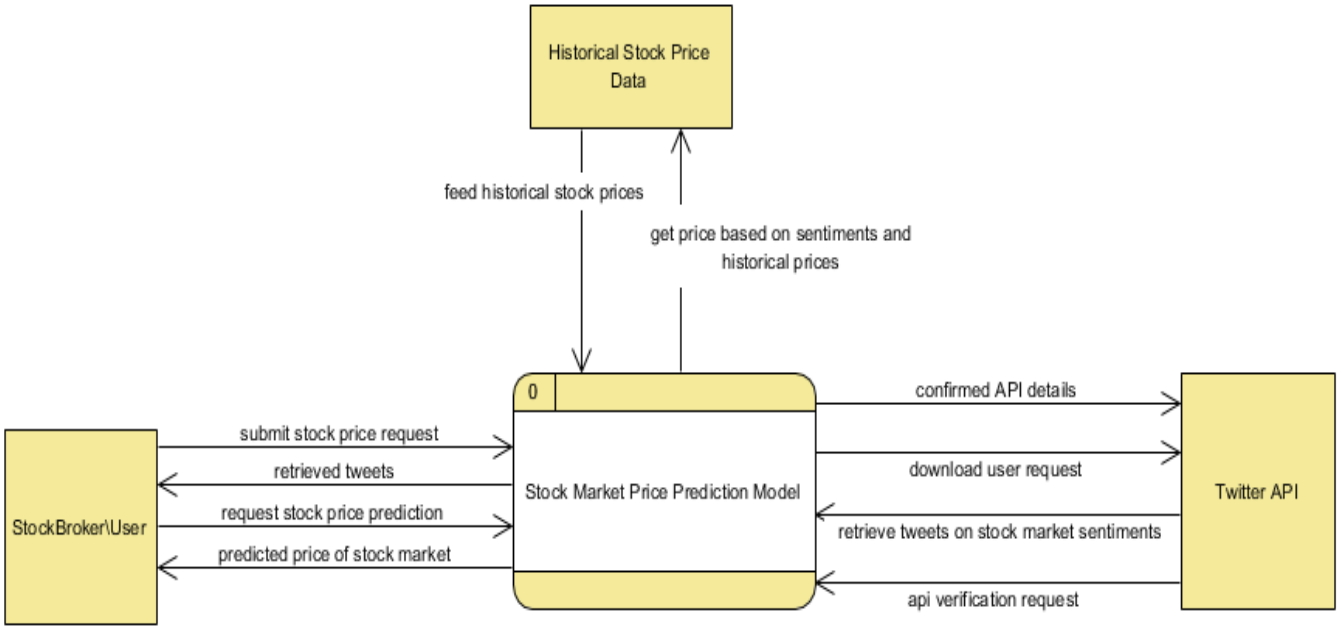


Figure 4:4: Context Diagram

4.5.3. Level 1 DFD Diagram

The context diagram is then expanded into several inter-related processes or levels. A level 1 DFD diagram represents the system's main processes, data stores and data processes with high-level details. With DFDs, it is easier for system users and non-users to understand how data flows through the system. The level 1 DFD is different from the context diagram as they illustrate the first level processes of the system. The level 1 DFD of the stock prediction model is as below in figure 4:5:

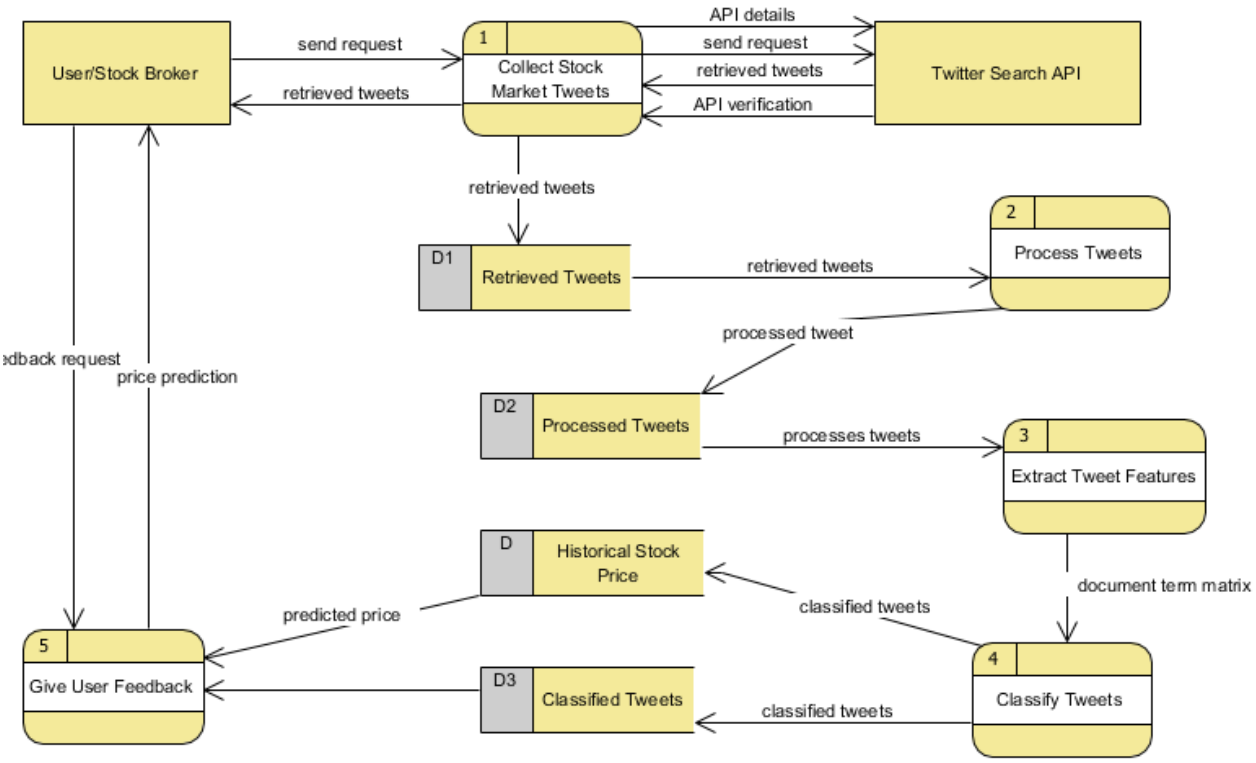


Figure 4:5: Level 1 DFD Diagram

.

Chapter 5 : System Implementation and Testing

5.1 Introduction

This chapter explains how the model of the prototype was developed and tested. It explains the whole process starting with the building of the sentiment analysis for financial news corpus which entails the way the data was obtained and the format. The next step is the preprocessing of the corpus then followed by the training of the model. The model is tested against 30% of the dataset to get the accuracy. The sentiment value is then correlated with the day's share price to build a forecasting model for 30 days in the future. Various experiments with different algorithms with different features were tested to pick the best model.

5.2 Sentiment Analysis

5.2.1 Building Sentiment Corpus

Tweets with financial news were crawled from twitter using the Twitter API with the help with tweepy which is a python library. The tweets were crawled and saved in a csv. Figure 5:1 and figure 5:2 illustrates a sample code used in crawling and downloading messages respectively.

```
def get_all_tweets(self,profile_name):
        # Twitter only allows access to a users most recent 3240 tweets with this method
        # initialize a list to hold all the tweepy Tweets
        alltweets = []
        # make initial request for most recent tweets (200 is the maximum allowed count)
        new_tweets = self.api.user_timeline(screen_name=profile_name, count=200)
        # save most recent tweets
        alltweets.extend(new_tweets)
        # save the id of the oldest tweet less one
        oldest = alltweets[-1].id - 1
        # keep grabbing tweets until there are no tweets left to grab
        while len(new_tweets) > 0:
            # print("getting tweets before %s") % (oldest)
            # all subsiquent requests use the max_id param to prevent duplicates
            new_tweets = self.api.user_timeline(screen_name=profile_name, count=200, max_id=oldest)
            # save most recent tweets
            alltweets.extend(new_tweets)
            # update the id of the oldest tweet less one
            oldest = alltweets[-1].id - 1
            print("...%s tweets downloaded so far" % (len(alltweets)))
        # transform the tweepy tweets into a 2D array that will populate the csv
        outtweets = [[tweet.id_str, tweet.created_at, self.cleaner.pre_cleaning(tweet.text)] for tweet in alltweets
        # write the csv
        with open('../data/twitter_data/raw_data/'+profile_name+'.csv', 'w') as f:
            writer = csv.writer(f)
            writer.writerow(["id", "created_at", "text","label"])
            writer.writerows(outtweets)
        pass
```

Figure 5:1: Crawler Code

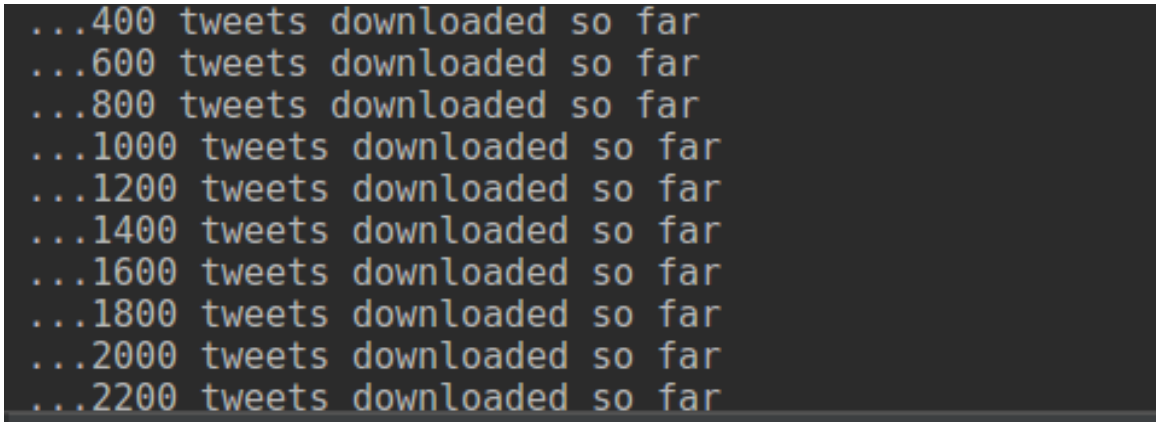


Figure 5:2: Downloading Tweets

5.2.2 Preprocessing

The text data obtained was in unstructured data format and this cannot be used in building a machine learning model. The data contains the tweet id, created at which is a timestamp and the text field. Text data contains more noisy words and symbols which are not contributing towards classification. The text data contains numbers, white spaces, tabs, punctuation characters, stop words, URL links, retweet symbols and others. All this needed to be cleaned by removing all those.

In preprocessing the first start was to harmonize the text by converting it to lowercase. The text may have different cases, and this may affect the classification. Then we remove the URL links, hashtags, usernames, punctuations and Twitter symbols. Figure 5:3 show the sample code used in the preprocessing of the data crawled. Figure 5:4. Shows the sample dataset that was crawled in its raw state. Figure 5:5 shows how the dataset looks after being cleaned.

```
20      def clean_tweets(self,tweet):
21          # harmonize the cases
22          lower_case_text = tweet.lower()
23          # remove urls
24          removed_url = re.sub(r'http\S+', '', lower_case_text)
25          # remove hashtags
26          removed_hash_tag = re.sub(r'#\w*', '', removed_url)  # hastag
27          # remove usernames from tweets
28          removed_username = re.sub(r'@\w*\s?','',removed_hash_tag)
29          # removed retweets
30          removed_retweet = removed_username.replace("rt", "", True)  # remove to retweet
31          # removing punctuations
32          removed_punctuation = removed_retweet.translate(self.remove_punctuations)
33          # remove spaces
34          remove_g_t = removed_punctuation.replace("&gt", "", True)
35          remove_a_m_p = remove_g_t.replace("&amp", "", True)
36          final_text = remove_a_m_p
37          return final_text
38
```

Figure 5:3: Text Cleaning Code

id	created_at	text
9.6691428E+17	2018-02-23 05:55:03	Kirubi completes buyback of Haco from South African firm https://t.co/vmX5BaFpYo https://t.co/an3Wr2sABE
9.6687098E+17	2018-02-23 03:03:00	TOP IN BUSINESS: Kenyans to pay Sh323bn interest on Eurobond II. Get a copy of Friday's BD to read this story and mâ€¦ https://t.co/WMI6zucF1r
9.6684456E+17	2018-02-23 01:18:00	BODO: Cash is still the king for many shoppers https://t.co/rQLZF8A5zF https://t.co/nSEMTalzZ5
9.6680983E+17	2018-02-22 23:00:00	"1
9.6677965E+17	2018-02-22 21:00:04	KRA loses Sh2.5bn tax claim from sugar importing firm https://t.co/9ihXTVWUJh https://t.co/lqSq4cmiMd
9.6675698E+17	2018-02-22 19:30:00	Africa must go into the cloud for unique gains of technology https://t.co/9c0kbVdYJt https://t.co/K0jVXVVDjF
9.6674796E+17	2018-02-22 18:54:10	TOP IN BUSINESS: Kenyans to pay Sh323bn interest on Eurobond II. Get a copy of Friday's BD to read this story and mâ€¦ https://t.co/9rQwzpTFfY
9.6674188E+17	2018-02-22 18:30:00	HR office for New World ought to change tack fast https://t.co/Q885EGlgx8 https://t.co/Vb5lxwe0LU
9.6673433E+17	2018-02-22 18:00:01	What professionals want in the law to police the Web https://t.co/1Qd6RonSc0 https://t.co/CMiMXDHaDp
9.667288E+17	2018-02-22 17:38:00	Exploiting client inexperience is a short-lived plan https://t.co/HqFkqHh9uQ https://t.co/xJgHCV1Q87
9.6671924E+17	2018-02-22 17:00:02	Why it is no longer a joke to wear the directorâ€™s hat https://t.co/4BdKbYHcTg https://t.co/RcHJW8CBd3
9.6671168E+17	2018-02-22 16:30:00	Building teams that are keen to win big https://t.co/MjRNiCHrHV https://t.co/0Vg597LOUp
9.6670414E+17	2018-02-22 16:00:03	Oserian seeks nod to produce solar power https://t.co/RZsvi2DGMS https://t.co/s9oXfggxSN
9.6669659E+17	2018-02-22 15:30:02	Import cover hits four-month high https://t.co/IbyEli4TxU https://t.co/zLmXra4h3i
9.6669407E+17	2018-02-22 15:20:00	Use value addition to cut post-harvest losses https://t.co/uV2Of1crpG https://t.co/jqy409ddaO
9.6668905E+17	2018-02-22 15:00:04	EDITORIAL: Fuel price surge worrying https://t.co/oyzVugGzlo https://t.co/TJHWt7O7gB
9.66684E+17	2018-02-22 14:40:00	Why advertising industry is now embracing big data https://t.co/MmdhER1Py3 https://t.co/S4rA4t0ruc
9.6667947E+17	2018-02-22 14:22:00	Digital techâ€™s role in Africa transformation https://t.co/GBofOXpWpu https://t.co/jNI32IMSIo
9.6667403E+17	2018-02-22 14:00:23	MANGENI: Integration steps Africa requires for trade glory https://t.co/th1xfKFzCA https://t.co/C5CZCpkrJZ
9.6666976E+17	2018-02-22 13:43:24	African nations urged to avoid hydropower reliance https://t.co/z09AQohxHP https://t.co/MJMxB03P2T
9.6666025E+17	2018-02-22 13:05:36	IMF warns over Kenya's rate in taking new debts https://t.co/zR0WK2bXUL https://t.co/AcNLruv6ZQ
9.6665605E+17	2018-02-22 12:48:55	Seven more KTDA plants ditch Kenya Power electricity https://t.co/WJt2zTySZ4 https://t.co/m8sctk6hFB
9.6664576E+17	2018-02-22 12:08:03	TECH GIANT Google enables M-Pesa Express payments on the Google Play Store allowing Kenyan Android users to pay forâ€¦ https://t.co/LDqXuqxGSi
9.6663607E+17	2018-02-22 11:29:32	ICYMI: Kenya raises $2bn in fresh Eurobond issue https://t.co/qhzUfVrNMv https://t.co/46TnD0B94L

Figure 5:4: Data Retrieved From Twitter

1	text
2	kirubi completes buyback of haco from south african firm
3	top in business: kenyans to pay sh323bn interest on eurobond ii. get a copy of friday's bd to read this story and mâ€¦
4	bodo: cash is still the king for many shoppers
5	kra loses sh2.5bn tax claim from sugar importing firm
6	africa must go into the cloud for unique gains of technology
7	top in business: kenyans to pay sh323bn interest on eurobond ii. get a copy of friday's bd to read this story and mâ€¦
8	hr office for new world ought to change tack fast
9	what professionals want in the law to police the web
10	exploiting client inexperience is a short-lived plan
11	why it is no longer a joke to wear the directorâ€™s hat
12	building teams that are keen to win big
13	oserian seeks nod to produce solar power
14	import cover hits four-month high
15	use value addition to cut post-harvest losses
16	editorial: fuel price surge worrying
17	why advertising industry is now embracing big data
18	digital techâ€™s role in africa transformation
19	mangeni: integration steps africa requires for trade glory
20	african nations urged to avoid hydropower reliance
21	imf warns over kenya's rate in taking new debts
22	seven more ktda plants ditch kenya power electricity
23	tech giant google enables m-pesa express payments on the google play store allowing kenyan android users to pay forâ€¦
24	icymi: kenya raises $2bn in fresh eurobond issue
25	fashion chain lc waikiki to open coast store april

Figure 5:5: Cleaned Twitter Data

5.2.3 Labeling

After the preprocessing, we need labels that will serve as a training and testing dataset. The labels are considered the most important to the model development. Each tweet of this dataset is tagged as either 1 if positive or -1 if negative or 0 if neutral. To classify these tweets. This process was done manually. Figure 5:6 below illustrates a sample dataset:

ten priorities for getting agriculture moving in zimbabwe	1
from the archives: the company that turned african telcos into insurance brokers	-1
from the archives listen: success depends on how well you understand yourself	0
in case you missed it: emerging market 2018 outlook via	1
newsletter: what's next for zimbabwe? making consumer goods affordable in africa	1
"from the archives: ""young african entrepreneurs should see themselves like the carnegies and rockefellers of americâ€¦ "	1
tourists still flock to cape town if the water runs out	1
africa's greatest economic opportunity: trading with itself	1
kenyas selina wamucii is empowering smallholder farmers through mobile tech	1
everyone's a winner? how 'impact investing' can make money â€“ and do good	1
from the archives â€“ listen: how social ventures can move from grant funding to real money |â€¦	1
ethiopia could be sitting on one of s great untapped gold deposits	1
newsletter: what's next for zimbabwe? making consumer goods affordable in africa	1
"from the archives: ""iâ€™m a control freak and bad with delegation. however	1
ten priorities for getting agriculture moving in zimbabwe via @madeitinafrica	1
zimbabwe â€“ where to now?	0
business opportunity: distribute and import premium wines throughout the african continent	1
from the archives â€“ listen: bottled-water company reveals secrets to franchising success |â€¦	1

Figure 5:6: Labeled Dataset

5.2.4 Training the model

The next crucial step is the creation of the model by training using our label dataset as shown in figure 5:7. SVM with bigram feature performed the best during the experiments, so the model was trained using it. The corpus was first shuffled then split into two; the training dataset which is 70% and the rest 30% is used as a testing dataset to measure the model's performance. Of the experiments carried out, SVM exhibited the best performance and using bigram feature increased its accuracy.

```
def svm_accuracy(X, y):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
    svm = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                    ('svm', SVC(kernel="linear", C=1))])
    svm = svm.fit(X_train, y_train)
    ypred = svm.predict(X_test)
    print("SVM metrics")
    print(metrics.accuracy_score(y_test, ypred))
    print(metrics.classification_report(y_test, ypred))
    drawrocSVM(y_test, ypred)
```

Figure 5:7: SVM Model Training Code

5.2.5 Testing

The testing dataset constituted 30% of the original dataset which was used to validate the model. The model was validated using a confusion matrix in table 5:1.

	Actual -1 (negative)	Actual 1 (positive)
Predicted -1 (negative)	207	241
Predicted 1 (positive)	59	1290

Table 5:1: Confusion Matrix

From the confusion matrix, we are able to get the values for true positive, true negative, false positive and false negative as illustrated in Table 5:2

True Negative	207
False Negative	241
True Positive	1290
False Positive	59

Table 5:2: Confusion Matrix Values

The metrics: accuracy, recall, precision, and f-score can then be calculated from the values in Table 5:3 or the confusion matrix on table 5:1. The accuracy of the model was computed to be 83%.

	Precision	Recall	F-Score	Support
Positive (1)	0.84	0.96	0.90	1349
Negative (-1)	0.78	0.46	0.58	448
Total Average	0.83	0.83	0.82	1749

Table 5:3: SVM Performance

Also a receiver operating characteristics (ROC) curve for the SVM model was drawn. This helps to visualize the performance of the classifier.

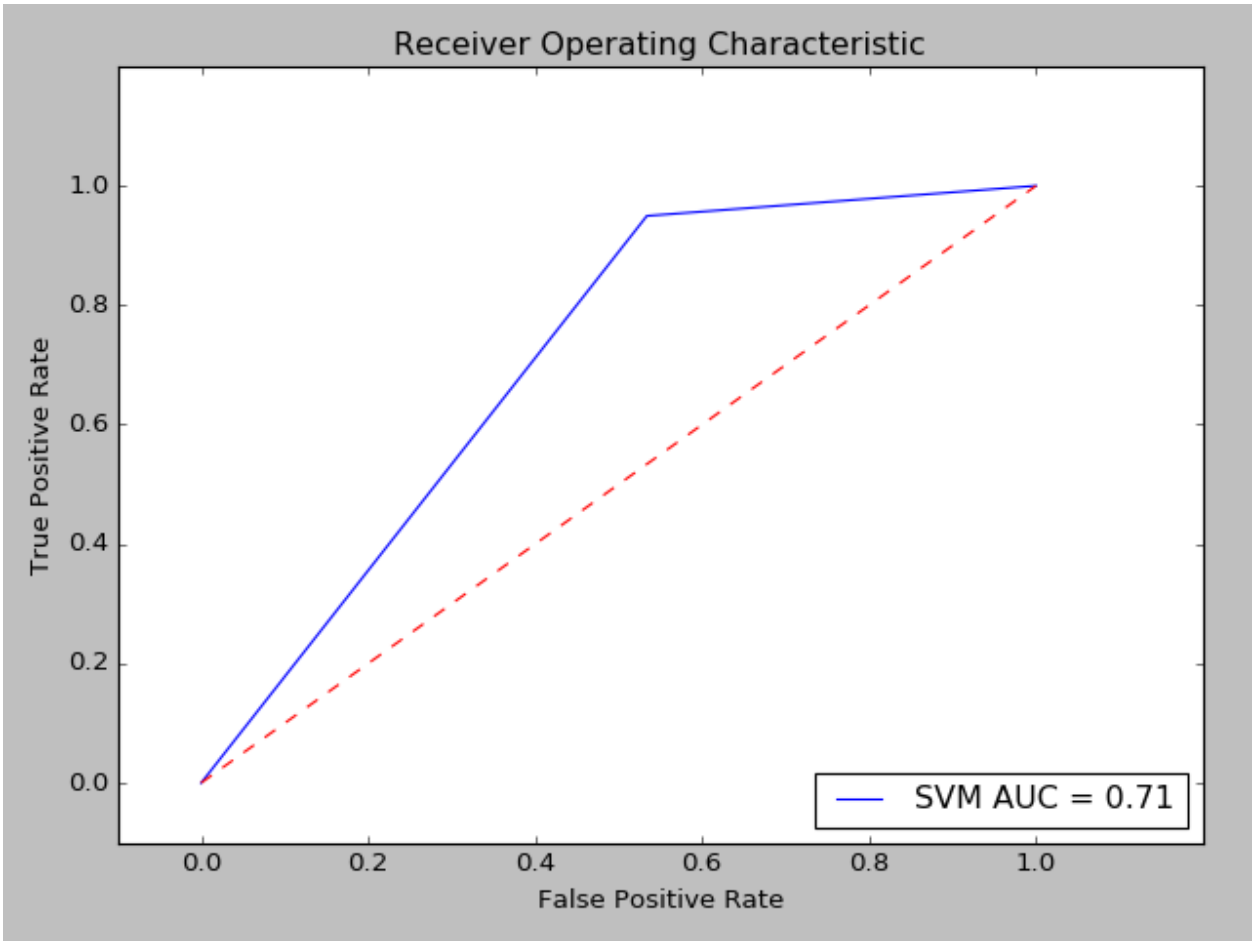


Figure 5. 1 ROC Curve

5.3 Share Price Prediction

We require the existing share prices which include the historical and current share prices and incorporate the day's financial news sentiment to help to predict future share price value. The sentiment values serve as input into the new model.

5.3.1 Building corpus

This involved crawling NSE historical data for the chosen company which is Equity bank. The data was daily data for the last one year and was saved in a csv file shown in figure 5:8.

	Code	Name	Lowest Price of the Day	Highest Price of the Day	Closing Price	Previous Day Closing Price	Volume Traded
2	EGAD	Eaagads Ltd	21.25	21.25	21.25	21.25	-
3	KUKZ	Kakuzi	355	355	355	355	-
4	KAPC	Kapchorua Tea Kenya Plc	74	74	74	72	100
5	LIMT	Limuru Tea Company Ltd	500	500	500	500	-
6	SASN	Sasini Tea and Coffee Ltd	25.25	26.5	26.25	26	5300
7	WTK	Williamson Tea Kenya Plc	149	150	149	150	1400
8	C&G	Car and General (K) Ltd	19	19	19	19	-
9	FIRE	Sameer Africa Ltd	2.5	2.5	2.5	2.5	100
10	BBK	Barclays Bank Ltd	10.4	10.8	10.75	10.6	2.39M
11	COOP	Co-operative Bank of Kenya Ltd	16.5	16.8	16.7	16.55	399300
12	DTK	Diamond Trust Bank Kenya Ltd	203	205	204	205	23700
13	EQTY	Equity Group Holdings Ltd	43	43.75	43.25	42.75	1.11M
14	HFCK	HF Group Limited	10.6	10.9	10.6	10.85	31600
15	I&M	I & M Holdings Limited	117	117	117	116	4300
16	KCB	Kenya Commercial Bank Ltd	45.25	46	45.5	45.25	257600
17	NBK	National Bank of Kenya Ltd	8.9	9	8.95	9	1700
18	NIC	NIC Bank Ltd	36.5	37.5	36.75	36.25	168300
19	CFC	Stanbic Holdings Ltd	81	85	82	81	6800
20	SCBK	Standard Chartered Bank Ltd	206	207	206	203	3800
21	DCON	Deacons (East Africa) PLC	2.85	2.85	2.85	2.85	-
22	XPRS	Express Ltd	3.75	3.75	3.75	3.75	-
23	KQ	Kenya Airways Ltd	15.5	17	15.65	16.15	55700
24	LKL	Longhorn Kenya Ltd	5.4	5.4	5.4	5.45	400
25	NBV	Nairobi Business Ventures Ltd	2.7	2.85	2.75	2.6	200
26	NMG	Nation Media Group Plc	95	106	103	103	16500

Figure 5:8: Share Price Dataset

5.3.2 Preprocessing

This involved converting the daily prices for all the company for that day to monthly or yearly share prices data format. Also, it involves incorporating the day's sentiments which comprise of the sentiment value, number of positive tweets and the number of negatives tweets. The day's sentiment values are derived by performing a K-Nearest Neighbor on the classified day' tweets, and a number of tweets that have the highest number of sentiment classification takes the final sentiment value i.e. if the negative values are ore than positive sentiments then the final sentiment for the day is negative.

DATE	HIGH	LOW	CLOSING	PREVIOUS DAY	VOLUME	SENTIMENT	POSITIVE TWEETS	NEGATIVE TWEETS
20170208	26.5	27.25	26.75	26	9690000	-1	50	100
20170210	27	28	27	27.25	3490000	1	82	20
20170216	27	28	27	26.75	3740000	-1	25	3
20170221	27	27.5	27	27	6510000	1	10	4
20170222	26.75	27.25	27	27	1060000	1	25	7
20170227	26.5	28	26.75	26.75	2049999	1	15	5
20170302	25.75	26.25	25.75	26	1130000	1	15	8
20170307	25.25	26	25.5	25.75	1730000	1	22	4
20170309	26	27	26.5	25.75	4970000	1	33	12
20170310	26.75	28.5	27	26.5	1430000	1	17	8
20170313	27	29.5	28.75	27	1370000	1	29	10
20170314	29	30.25	29	28.75	12130000	1	5	0
20170321	29.5	30	29.5	29.5	5880000	1	12	6
20170322	29.75	30	29.75	29.5	2790000	-1	5	19
20170327	30.25	31	30.5	30.25	5050000	-1	14	230
20170328	30.75	32	31.25	30.5	1360000	-1	9	19

Figure 5:9: Training Dataset

5.3.3 Training the model

Artificial neural network proved the best algorithm that yielded the best result for 30 days share price prediction. The dataset is split into two; the training dataset which is 70% and the rest 30% is used as a testing dataset to measure the model's performance.

```
X_train, X_test, y_train, y_test = cross_validation.train_test_split(X, y, test_size = 0.3)
# Training
clf = MLPRegressor()
clf.fit(X_train,y_train)
# Testing
confidence = clf.score(X_test, y_test)
print("confidence: ", confidence)
```

Figure 5:10: ANN Training Code

5.3.4 Testing the model

For testing, 30% of the dataset was used for testing. Then a measured error was used to compute the error between the predicted and the actual. The figure 5:11 shows the 30 days predicted by the model, and we have the current values for the 30days which were predicted.

```
30 Days prediction
[49.06130084 49.5535442  49.22728663 50.04950723 50.83500396 51.45379819
 52.21568315 52.56656562 50.51066652 51.07002578 51.31474374 51.55727741
 50.61274741 52.36487527 52.23036638 51.17194115 50.71991512 50.691316
 52.01999041 50.34107001 51.66423627 51.17421646 51.79814216 49.86420083
 51.79123309 52.08723244 51.14675118 50.58113557 51.08041749 52.84323908]
```

Figure 5:11: ANN Predicted Prices

The figure 5:12 shows a sample computation of the mean squared error

```
from sklearn.metrics import mean_squared_error
actual_values = [
    45.25,46,47,46.75,47.75,47.75,47.75,48,
    48.75,50,49.75,51,53,52.5,53.5,52.5,52.5,
    53,52.5,53.5,54,53.25,53,52.75,52.25,51.50,
    52.25,52.75,53.25,53.75]


#ANN
predicted_values = [
    48.9878508 , 49.57022879, 49.3529764 , 50.12779638, 50.9736035 , 51.6138593,
    52.44269939, 52.74885556, 50.68329415, 51.26848788, 51.50949479, 51.74168738,
    50.88472517, 52.60423564, 52.42768631, 51.40283603, 50.84071616, 50.95787582,
    52.18104572, 50.61284816, 51.81845599, 51.37957272, 52.00000527, 49.59193102,
    51.96610947, 52.31168079, 51.35734326, 50.85376892, 51.35767808, 53.03472337]

mse = mean_squared_error(actual_values,predicted_values)
```

Figure 5:12: Mean Squared Error Code

Table 5.4 illustrates the performance of the predicted value Mean squared error was performed.

Mean Squared Error	5.77

Table 5:4: SVM Mean Squared Error

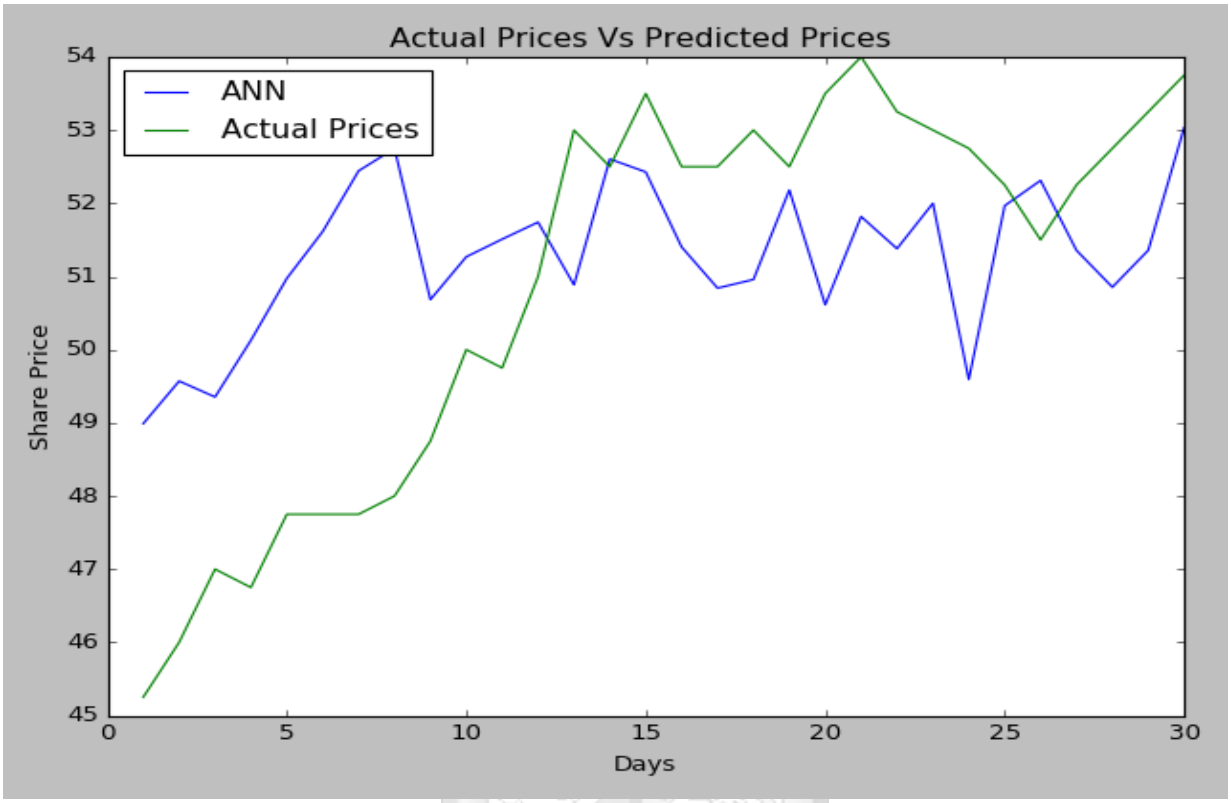


Figure 5:13: ANN Predicted Price and Actual Price Graph

VT OMNES VNVM SINT

Chapter 6 : Discussions

6.1 Introduction

This chapter discusses the experiments carried during this research to achieve the specified results in line with the objectives of the project. The main objective was to build a model that can use existing historical data and sentiment analysis to predict the share price for a company. For sentiment, analysis SVM seemed to perform better than the other machine learning method while artificial neural network performed better in the prediction of the share price.

6.2 Experiments on Sentiment Analysis

6.2.1 Using Different Classifiers

The purpose of this experimentation was to compare the performance of the different classifiers for sentiment analysis. Four machine learning methods were used namely SVM, Naive Bayes, Random Forest and K-nearest neighbor. The results are shown in table 6:1

Classifier	**Accuracy**	**Precision**	**Recall**	**F-score**
SVM	0.83	0.82	0.83	0.82
Naive bayes	0.77	0.81	0.78	0.70
Random Forest	0.813	0.80	0.81	0.79
KNN	0.801	0.79	0.80	0.77

Table 6:1: Classifiers Performance

6.2.2 Experiment 1: SVM with Different Feature Types

The main purpose of this experimentation was to determine the effect of using different feature types on the SVM and check how the accuracy of the model is affected. The features used were unigram, bigrams, and trigrams. The results in Table 6:2 shows that the best performance of the SVM classifier is obtained when using bigram feature.

Feature	**Accuracy**	**Precision**	**Recall**	**F-Score**
Unigram	0.833	0.83	0.83	0.82
Bigram	0.839	0.84	0.84	0.82
Trigram	0.838	0.83	0.84	0.82

Table 6:2: SVM Performance

6.2.3 Experiment 2: Naïve Bayes with Different Feature Types

This experiment is similar to experiment 2, using unigram, bigram and trigrams on Naive Bayes checking its performance. The results in Table 6:3 shows that the best performance of the Naïve Bayes classifier is obtained when using unigram feature.

Feature	**Accuracy**	**Precision**	**Recall**	**F-Score**
Unigram	0.768	0.81	0.77	0.69

Bigram	0.762	0.82	0.76	0.67
Trigram	0.759	0.82	0.76	0.66

Table 6:3: Naive Bayes Performance

6.2.4 Experiment 3: Random Forest with Different Feature Types

This experiment is similar to experiment 2, using unigram, bigram and trigrams on Random Forest checking its performance. The results in Table 6:3 shows that the best performance of the Random Forest classifier is obtained when using bigram feature.

Feature	**Accuracy**	**Precision**	**Recall**	**F-Score**
Unigram	0.802	0.80	0.79	0.79
Bigram	0.798	0.79	0.79	0.78
Trigram	0.783	0.78	0.78	0.77

Table 6:4: Random Forest Performance

6.2.5 Experiment 4: KNN with Different Feature Types

This experiment is similar to experiment 2, using unigram, bigram, and trigrams on K-Nearest Neighbor checking its performance. The results in Table 6:2 shows that the best performance of the SVM classifier is obtained when using bigram feature.

Feature	Accuracy	Precision	Recall	F-Score
Unigram	0.800	0.79	0.80	0.77
Bigram	0.797	0.79	0.80	0.76
Trigram	0.796	0.79	0080	0.76

Table 6:5: KNN Performance

6.3 Stock Prediction

6.3.1 Different Classifiers

The goal of this experiment was to compare the performance of the different classifiers for stock market price incorporating the sentiment value. Four machine learning methods were used namely support vector machine, Naive Bayes, Random Forest and artificial neural network. From the Table 6:6 it shows Artificial neural network was better regarding the values predicted as its mean squared error was the least while the SVM model was the worst as it had its mean squared value the highest.

Classifier	Mean Squared Error
SVM	129.40
Naive Bayes	43.59
Random Forest	74.96

Artificial Neural Network	5.77

Table 6:6: Classifiers Performance in Stock Price Prediction

Figure 6:1 shows the different 30 days prices predicted by the different models and also the actual prices that occurred in the 30 days. From the graph, ANN price curves are the closest to the actual prices in terms of the graph trend.

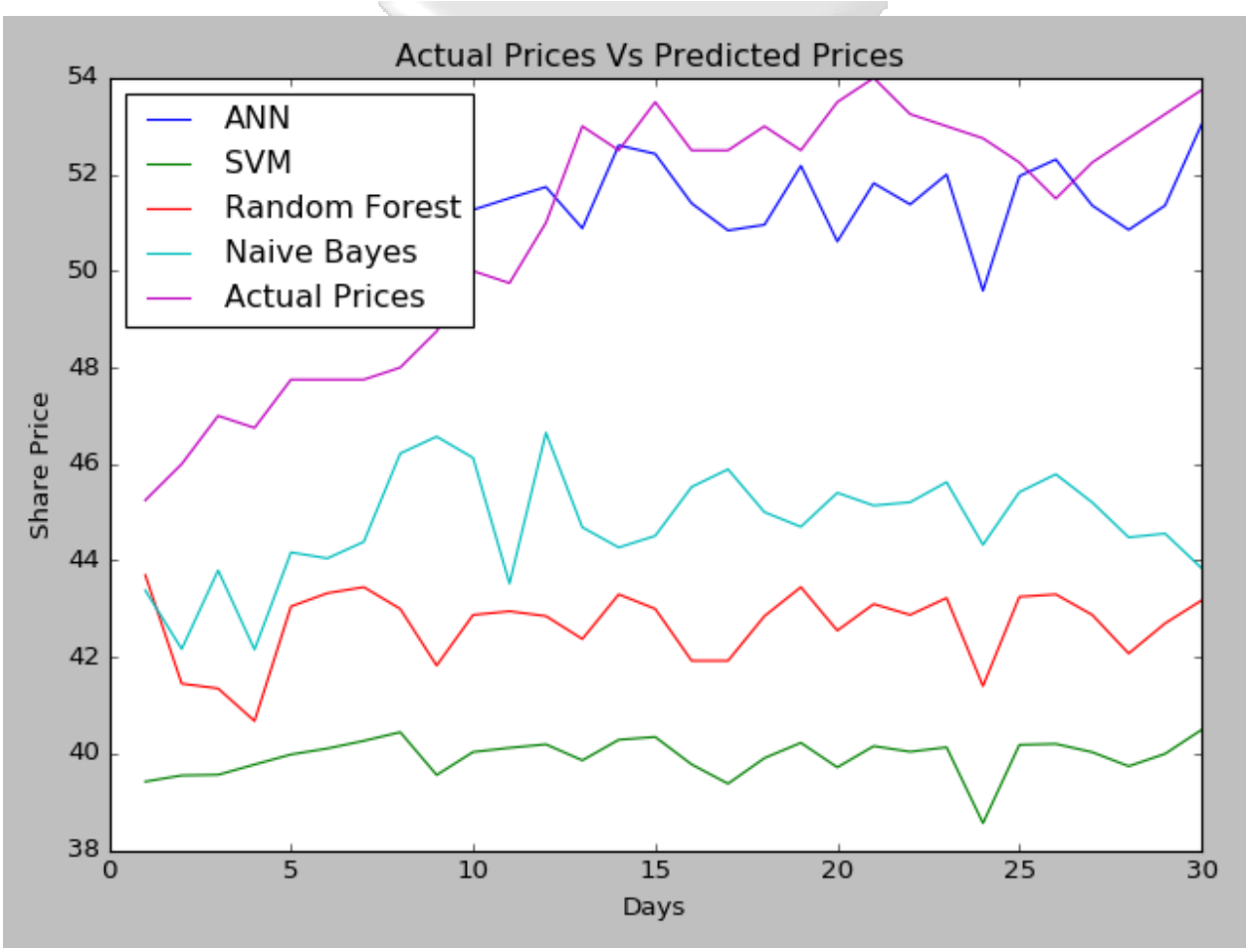


Figure 6:1: Different Classifiers Performance Graph

Chapter 7 : Conclusion and Recommendations

7.1 Conclusion

In this research, we have tried to predict equity's 30 days share price movement on Nairobi Stock Exchange by performing sentiment analysis on financial news tagged tweets on twitter about the company.

7.2 Recommendation

To increase the accuracy of the model and for it to be more reliable, the financial news should be collected from multiple sources such as digital newspaper headlines to avoid bias and to provide diversity in the news. Also, the data should be collected continuously for a duration greater than one year to avoid one been limited. On Twitter, one should be able to subscribe to the enterprise package though it's costly. This will help one retrieve historical tweets that may have a financial impact on that days share price that will help the building and increasing the accuracy of the model. Currently, Twitter API is limited to only 7 days tweets.

A method needs to be devised on how we can uniquely identify any company listed on Nairobi Stock Exchange market on the internet. An example is Stock listed on US stock exchange like Facebook and Google which are listed on NASDAQ. To identify them anywhere is the internet is easy as theyy have the symbol $FB and $GOOGL respectively, this makes one easy to crawl any news that has the initial provide for them. If this can be implement to NSE listed companies to that unique way of identifying them on the internet or social media, this will make it easy to crawl their data.

References

Albert, B., & Eibe, F. (2011). Sentiment knowledge discovery in twitter streaming data.

Alejandro Mosquera, Lamine Aouad, Slawomir Grzonkowski, & Dylan Morss. (2014). On Detecting Messaging Abuse in Short Text Messages using Linguistic and Behavioral patterns.

Andries, P. E. (2007). Computational Intelligence: An Introduction (2 Edition). Wiley Publishing.

Antweiler, W., & Frank, M. Z. (2004). Is All That Talk Just Noise? The Information Content of Internet Stock Message Boards. Journal of Finance, 59(3), 1259–1294.

Arafat, J., Habib, A., & Hossain, R. (2013). Analyzing Public Emotion and Predicting Stock Market Using Social Media. American Journal of Engineering Research, 265–275.

Batool, R., Khattak, A., Maqbool, J., & Lee, S. (2013). Precise tweet classification and sentiment analysis, 461–466.

Berman, P. (2006). E-Learning Concepts and Techniques.

Bhardwaj, A., Narayan, Y., Dutta, M., Vanraj, & Pawan. (2015). Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty.

Bollen, J., & Mao, H. (2010). Twitter mood as a stock market predictor. IEEE, 91– 94.

Brealey, R. A., Myers, S. C., & Allen, F. (2005). Corporate Finance (8 Edition). New York: McGraw-Hill Irwin.

Brownlee, J. (2016a, November 18). What is a Confusion Matrix in Machine Learning? Retrieved 28 August 2017, from https://machinelearningmastery.com/confusion-matrix-machine-learning/

Burton, M. (2003). The Efficient Market Hypothesis and its Critics. The Journal of the Economic Perspectives, 17(1), 59–82.

Capital Markets Authority. (2017). Retrieved 26 August 2017, from https://www.cma.or.ke/index.php/about-us/who-we-are

Central Depository & Settlement Corporation (CDSC). (2017, August 7). Retrieved from http://fib.co.ke/deal/cds/

Choi, H., & Varian, H. (2012). Predicting the Present with Google Trends.The Economic Record, 2–9.

Choudhury, M. D., Sundaram, H., John, A., & Seligmann, D. D. (2010). Can Blog Communication Dynamics be correlated with Stock Market Activity? Proceedings of the Nineteenth ACM Conference on Hypertext and Hypermedia.

Chu, T., Jue, K., & Wang, M. (2017). Comment Abuse Classification with Deep Learning.

Dawei Yin, Zhenzhen Xue, & Liangjie Hong. (2009). Detection of Harassment on Web 2.0.

Dinesh Sonachalam. (2015). Using Twitter to predict Stock Market Returns.International Journal of Scientific & Engineering Research, 6(10).

Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep Learning for Event-Driven Stock Prediction. Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015).

Dumais, S. (2001, November). Support Vector Machines. Retrieved 19 August 2017, from https://www.microsoft.com/en-us/research/project/support-vector-machines/

Fama, E. (1965). The behavior of stock-market prices. The Journal of Business.

Fama, E. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. Journal of Finance, 25(2), 383–417.

Fang, X., & Zhan, J. (2015). Sentiment Analysis Using Product Review Data. Journal of Big Data, 2.

Gilbert, E., & Karahalios, K. (2010). Widespread Worry and the Stock Market. 4th International AAAI Conference on Weblogs and Social Media (ICWSM).

Hellstrom, T. (1998). A Random Walk through the stock Market Licentiate. Umea Univeristy.

History of NSE - Nairobi Securities Exchange (NSE). (2017). Retrieved 16 August 2017, from https://www.nse.co.ke/nse/history-of-nse.html

Hossein Hosseini, Sreeram Kannan, Baosen Zhang, & Radha Poovendran. (2017). Deceiving Google's Perspective API Built for Detecting Toxic Comments.

Huang, C., Chen, P., & Pan, W. (2011). Using Multi-Stage Data Mining Technique to Build Forecast Model for Taiwan Stocks. Neural Computing and Applications.

Jain, V. (2013). Prediction of Movie Success using Sentiment Analysis of Tweets. The International Journal of Soft Computing and Software Engineering.

Jones, T. S., & Richey, R. C. (2000). Rapid prototyping methodology in action: A developmental study, 63–80.

Khatri, S. K., Singhal, H., & Johri, P. (2014). Sentimental analysis to Predict Bombay Stock Exchange Using Artificial Neural Network, 380–384.

Khatri, S. K., & Srivastava, A. (2016). Using Sentimental Analysis in Prediction of Stock Market Investment. 5th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO).

Kihoro, J., & Okango, E. (2014). The stock market price prediction using artificial neural networks: An application to the Kenyan Equity Bank share prices, 16.

Kohavi, R., & Provost, F. (1998). On Applied Research in Machine Learning. In Editorial for the Special Issue on Applications of Machine Learning and the Knowledge Discovery Process, 30.

Lawrence, S. (2002). A Model for Stock price Fluctuations Based on Information, 48.

Lo, A. W., & MacKinley, A. C. (1999). A Non-Random Walk Down Wall Street. Princeton: Princeton University Press.

Lowd, D., & Domingos, P. (2005). Naive Bayes Models for Probability Estimation.

Lowe, D. (2012). Local Naive Bayes nearest Neighbor for Image Classification. Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 3650–3656.

Makrehchi, M., Shah, S., & Liao, W. (2013). Stock Prediction Using Event-based Sentiment Analysis. International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT).

Makrehchi, M., Shah, S., & Wenhui Liao. (2013). Stock Prediction Using Event-based Sentiment Analysis. International Conferences on Web Intelligence (WI) and Intelligent Agent Technology (IAT).

Marc Claesen, Smet, F. D., Moor, B. D., & Suykens, J. A. K. (2014). EnsembleSVM: A Library for Ensemble Learning Using Support Vector Machines. Journal of Machine Learning Research, 141–145.

Maynard, D., & Funk, A. (2011). Automatic detection of political opinions in tweets. Proceedings of the 8th International Conference on the Semantic Web, 88–99.

Medhat, W., Hassan, A., & Korashy, H. (2013). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 1093–1113.

Minging, H., & Bing, L. (2004). Mining and summarizing customer reviews. Proceedings of ACM SIGKDD International Con- Ference on Knowledge Discovery and Data Mining.

Mohammad, S., Dunne, C., & Dorr, B. (2009). Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus.

Nairobi Securities Exchange. (2017). Retrieved 16 August 2017, from https://www.nse.co.ke/nse/about-nse.html

NeuroAI. (2013). Stock Market Prediction | Neuro AI. Retrieved 7 August 2017, from http://www.learnartificialneuralnetworks.com /stockmarketprediction.html

Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. Proceedings of LREC.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. Foundations and Trends in Information, 2, 1–135.

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Sentiment classification using machine learning technique, 79–86.

Peterson, C., & Rögnvaldsson, T. (1992). An introduction to artificial neural networks. Proc. 1991 CERN Summer School of Computing, 113–170.

Platt, J. (1999). Probabilities for SV Machines. Advances in Large Margin Classifiers. MIT Press, 61–74.

Pagolu, S. & Nayan R, Panda, G., Majhi, B. (2016). Sentiment analysis of Twitter data for predicting stock market movements. 1345-1350. 10.1109/SCOPES.2016.7955659.

Pring, M. J. (1991). Technical Analysis Explained.

Quinlan, J. (1986). Machine Learning

Raschka, S. (2014, October 4). Naive Bayes and Text Classification. Retrieved from sebastianraschka.com/Articles/2014_naive_bayes_1.html

Rocha, M., & Macedo, M. (2011). Previsão do preço de ações usando redes neurais. Congresso USP de Iniciação Científica Em Contabilidade.

Thomsett, M. C. (1998). Mastering Fundamental Analysis. Chicago: Dearborn Publishing.

Uhrig, R. (1995). Introduction to artificial neural networks. Proceedings of the 1995 IEEE IECON 21st International Conference, 1, 33–37.

Wanjawa, B. W. (2014, May). A Neural Network Model for Predicting Stock Market Prices at the Nairobi Securities Exchange. University of Nairobi.

Zhang, H. (2004). The Optimality of Naive Bayes.

Zhang, L. (2013). Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation. The University of Texas at Austin.

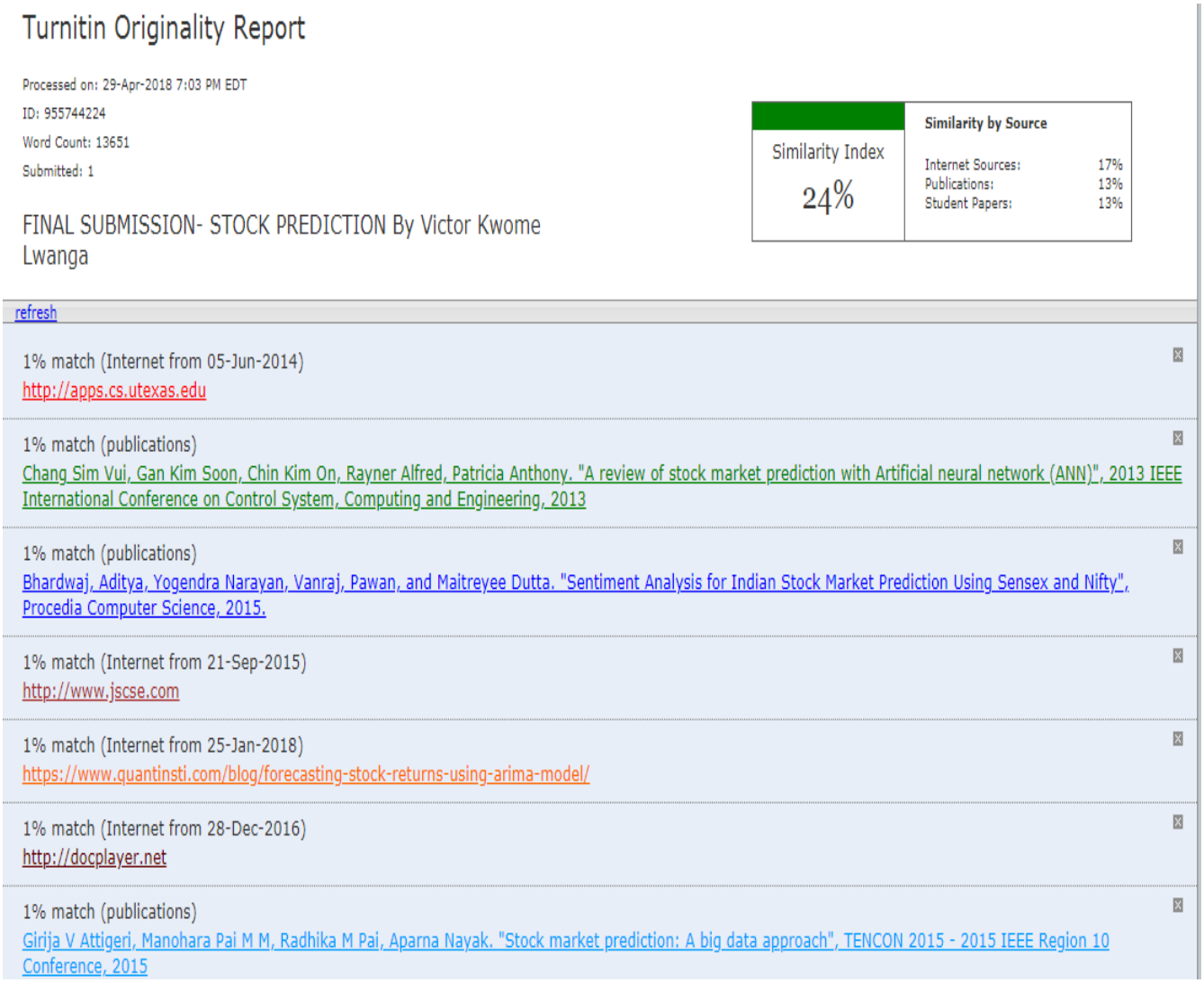
Zhang, Peter. (2003). Zhang, G.P.: Time Series Forecasting Using a Hybrid ARIMA and Neural Network Model. Neurocomputing 50, 159-175. Neurocomputing. 50. 159-175. 10.1016/S0925-2312(01)00702-0.

Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting Stock Market Indicators through Twitter.

Zupan, J. (1994). Introduction to Artificial Neural Network (ANN) Methods: What They Are and How to Use Them, 41–327.

Appendix

Appendix A: Originality Report



Turnitin Originality Report

Processed on: 29-Apr-2018 7:03 PM EDT
ID: 955744224
Word Count: 13651
Submitted: 1

FINAL SUBMISSION- STOCK PREDICTION By Victor Kwome Lwanga

Similarity Index	Similarity by Source	
24%	Internet Sources:	17%
	Publications:	13%
	Student Papers:	13%

refresh

1% match (Internet from 05-Jun-2014)
http://apps.cs.utexas.edu

1% match (publications)
Chang Sim Vui, Gan Kim Soon, Chin Kim On, Rayner Alfred, Patricia Anthony. "A review of stock market prediction with Artificial neural network (ANN)", 2013 IEEE International Conference on Control System, Computing and Engineering, 2013

1% match (publications)
Bhardwaj, Aditya, Yogendra Narayan, Vanraj, Pawan, and Maitreyee Dutta. "Sentiment Analysis for Indian Stock Market Prediction Using Sensex and Nifty", Procedia Computer Science, 2015.

1% match (Internet from 21-Sep-2015)
http://www.jscse.com

1% match (Internet from 25-Jan-2018)
https://www.quantinsti.com/blog/forecasting-stock-returns-using-arima-model/

1% match (Internet from 28-Dec-2016)
http://docplayer.net

1% match (publications)
Girija V Attigeri, Manohara Pai M M, Radhika M Pai, Aparna Nayak. "Stock market prediction: A big data approach", TENCON 2015 - 2015 IEEE Region 10 Conference, 2015

Appendix B: Python Source Code

```
import tweepy

import xml.etree.ElementTree as ET

class Credentials:

    def __init__(self):

        self.credential_xml = 'twitter-credentials.xml'

    def get_twitter_credentials(self):

        credential_xml_data = ET.parse(self.credential_xml).getroot()

        return (credential_xml_data[0].text,

                credential_xml_data[1].text,

                credential_xml_data[2].text,

                credential_xml_data[3].text)

    def authentinticate_twitter(self):

        twitter_credentials = self.get_twitter_credentials()

        auth = tweepy.OAuthHandler(twitter_credentials[0], twitter_credentials[1])

        auth.set_access_token(twitter_credentials[2], twitter_credentials[3])

        api = tweepy.API(auth)

        return api
```

The code below was used to clean the text and save it in a csv file.

```
import csv
import re
import string
import html
class Cleaner:
    def __init__(self):
        self.remove_punctuations = str.maketrans('', '', string.punctuation)
    def read_csv(self,csv_name):
        cleaned_text = []
        with open('../data/twitter_data/raw_data/'+csv_name+'.csv', newline='', encoding='utf-
8') as csvfile:
            reader = csv.DictReader(csvfile)
            for row in reader:
                text = row['text']
                clean_text = self.clean_tweets(text)
                cleaned_text.append(clean_text)
        self.save_cleaned_csv('cleaned_'+csv_name,cleaned_text)
    def clean_tweets(self,tweet):
        # harmonize the cases
        lower_case_text = tweet.lower()
        # remove urls
        removed_url = re.sub(r'http\S+', '', lower_case_text)
```

```
    # remove hashtags
    removed_hash_tag = re.sub(r'#\w*', '', removed_url)  # hastag
    # remove usernames from tweets
    removed_username = re.sub(r'@\w*\s?','',removed_hash_tag)
    # removed retweets
    removed_retweet = removed_username.replace("rt", "", True)  # remove to retweet
    # removing punctuations
    removed_punctuation = removed_retweet.translate(self.remove_punctuations)
    # remove spaces
    remove_g_t = removed_punctuation.replace("&gt", "", True)
    remove_a_m_p = remove_g_t.replace("&amp", "", True)
    final_text = remove_a_m_p
    return final_text
def pre_cleaning(self,text):
    html_escaped = html.unescape(text)
    final_text = html_escaped.replace(';','')
    return final_text
def pre_labeling(self,text):
    lower_case_text = text.lower()
    removed_url = re.sub(r'http\S+', '', lower_case_text)
    return removed_url
def save_cleaned_csv(self,name,tweets_list):
    with open('../data/twitter_data/cleaned_data/' + name + '.csv', 'w') as f:
```

```
            writer = csv.writer(f)
            writer.writerow(["text"])
            for tweet in tweets_list:
                writer.writerow([tweet,])
        pass
    def save_pre_labled_csv(self,csv_name):
        cleaned_text = []
        with open('../data/twitter_data/raw_data/' + csv_name + '.csv', newline='',
encoding='utf-8') as csvfile:
            reader = csv.DictReader(csvfile)
            for row in reader:
                text = row['text']
                clean_text = self.pre_labeling(text)
                cleaned_text.append(clean_text)
        self.save_pre_labeled_csv('unlabeled_' + csv_name, cleaned_text)
    def save_pre_labeled_csv(self,name,tweets_list):
        with open('../data/twitter_data/pre_labeled/' + name + '.csv', 'w') as f:
            writer = csv.writer(f)
            writer.writerow(["text","label"])
            for tweet in tweets_list:
                writer.writerow([tweet,])
        pass
```

The code below was used in SVM model training. This was used in sentiment analysis.

```
def svm_accuracy(X, y):
    X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=1)
    svm = Pipeline([('vect', CountVectorizer()), ('tfidf', TfidfTransformer()),
                    ('svm', SVC(kernel="linear", C=1))])
    svm = svm.fit(X_train, y_train)
    ypred = svm.predict(X_test)
    print("SVM metrics")
    print(metrics.accuracy_score(y_test, ypred))
    print(metrics.classification_report(y_test, ypred))
```

This is a sample artificial neural network code that was used in forecasting the store price.

```
import pandas as pd
import numpy as np
from sklearn import preprocessing, cross_validation
from sklearn.neural_network import MLPRegressor
df = pd.read_csv('../equity.csv')
df_close = df[[3]]
forecast_out = int(30) # predicting 30 days into future
df['Prediction'] = df_close.shift(-forecast_out) #  label column with data shifted 30 units up
# print(df.tail())
X = np.array(df.drop(['Prediction'], 1))
```

```
X = preprocessing.scale(X)

X_forecast = X[-forecast_out:] # set X_forecast equal to last 30

X = X[:-forecast_out] # remove last 30 from X

y = np.array(df['Prediction'])

y = y[:-forecast_out]

X_train, X_test, y_train, y_test = cross_validation.train_test_split(X, y, test_size = 0.3)

# Training

clf = MLPRegressor()

clf.fit(X_train,y_train)

# Testing

confidence = clf.score(X_test, y_test)

print("confidence: ", confidence)

forecast_prediction = clf.predict(X_forecast)

print('30 Days prediction')

print(forecast_prediction)
```