AN IMPROVED LEXICON USING LOGISTIC REGRESSION FOR SENTIMENT ANALYSIS

MAJOR PROJECT SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE AWARD OF DEGREE OF

Master of Technology

In

Information Systems

Submitted By:

KUNAL BHARGAVA

(2K15/ISY/12)

Under the Guidance

Of

Dr. Rahul Katarya

(Assistant Professor, Department of IT)

DEPARTMENT OF INFORMATION TECHNOLOGY

DELHI TECHNOLOGICAL UNIVERSITY

(2015-2017)

CERTIFICATE

This is to certify that **Kunal Bhargava (2K15/ISY/12)** has carried out the major project titled "**An improved lexicon using logistic regression for sentiment analysis**" in partial fulfilment of the requirements for the award of Master of Technology degree in Information Systems by **Delhi Technological University**.

The major project is a bonafide piece of work carried out and completed under my supervision and guidance during the academic session 2015-2017. To the best of my knowledge, the matter embodied in the thesis has not been submitted to any other University/Institute for the award of any degree or diploma.

Dr. Rahul Katarya
Assistant Professor
Department of Information Technology
Delhi Technological University
Delhi-110042

ACKNOWLEDGEMENT

I take the opportunity to express my sincere gratitude to my project mentor Dr. Rahul Katarya, Assistant Professor, Department of Information Technology, Delhi Technological University, Delhi, for providing valuable guidance and constant encouragement throughout the project. It is my pleasure to record my sincere thanks to him for his constructive criticism and insight without which the project would not have shaped as it has.

I humbly extend my words of gratitude to other faculty members of this department for providing their valuable help and time whenever it was required.

Kunal Bhargava

Roll No. 2K15/ISY/12

M.Tech (Information Systems)

E-mail: bkunal2707@gmail.com

ABSTRACT

Recently, stock market activities are becoming dependent intensely on social media interactions to deliver significant information for an extensive number of users. This obliges frameworks to scale expeditiously to suit the surge of new as well as existing users going to the recommendations based on information extracted from social media. In this work, we propose an approach for assigning scores to lexicon items using logistic regression based relative scoring to address both the proposal quality and the framework versatility. We propose to assign a rich range of scores to items in the lexicon, as indicated by their web usage history and corresponding effects. We utilise a Deep Learning way to deal with scores of the lexicon to space where the comparability amongst items and their favoured effects is maximised.

In this dissertation work, we will talk about a stock market lexicon that endeavours the sentiment regularities caught by a Logistic Regression model in microblog data. Most of the lexicon acquisition frameworks regard words as paired vectors under the exemplary sack of-words model; however, there is not an idea of relative comparability between words while depicting the same sentiment effect. This relativity is considered using the logistic regression model and the accuracy of the results is found to be improved significantly.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

SA	Sentiment Analysis
DM	Diebold- Mariano
AAII	American Association of Individual Investors
SOM	Self-Organising Maps
HRK	Hierarchical and Recursive K-means
SVM	Support Vector Machines
MR	Multiple Regression
VAR	Vector Auto-Regression
EA	Ensemble Averaging
POS	Part of Speech
DJIA	Dow Jones Industrial Average
S&P500	Standard and Poor's 500

Chapter 1

INTRODUCTION

Ever since the release and use of Twitter and other similar microblogs like StockTwits, users have been posting and following a lot of messages through such platforms. Recent online recommendations depend vigorously on information extracted from programmed models to recommend important content to a substantial number of users. This obliges frameworks to scale speedily to accommodate the stream of users visiting and considering online platforms surprisingly. In this work, we propose a logistic regression framework to address both the relative impact of words in tweets and the framework versatility. We propose to utilise a rich list of lexicon set, as indicated by their presence history among different tweets. We utilise a Deep Learning way to deal with items to a relative space where we can deter amongst items of the same sentiment with their favoured effect being maximised.

Recent advances in social media sites have provided the users with an easy way to convey more information in fewer words. Twitter, for example, allows its users to post short messages of 140 characters each only, which interpret more with the advantage of representing live situations. Until now major lexicons are just a collection of words where each word has its sentiment value (positive or negative) associated with it. These lexicons constituted to be the key element in Sentiment Analysis. For instance, the sentence “The effects of this event will be beneficial” can be assigned a positive sentiment if the term “beneficial” is present with a positive value in the lexicon. Although there can be words of same as well as different sentiments, with the same sentiment words having their own corresponding weight different than others.

In this work, we will talk about a scoring framework that exploits the semantics regularities caught by analysing terms in texts of social media messages. Numerous lexicon acquisition frameworks regard words as parallel vectors under the exemplary bag of words model. However, there is not an idea of relative comparability between words while depicting the same sentiment effect. This method exhaustively investigates text present in large volumes of social media data. Text classification techniques are employed to analyse the words present in

messages and associate the corresponding sentiment with each word. We contemplate and modify the current models, particularly lexicon acquisition models from a machine learning point of view. Logistic Regression, a classification algorithm which is used when the output variable has only two values (0 or 1).We utilised this profound learning way to deal with finding the semantic elements depicting similar sentiment and fit a relative scoring model to elevate sentiment analysis inclinations for stock market predictions.

Formal definition

Due to the volume and velocity properties of social media data, human analysis is impracticable and thus Sentiment Analysis is used to automatically mine the large amounts of data, both structured as well as unstructured present on the social media. This has proven to be potentially valuable for users in making decisions regarding the stock market activities [1].

Formally, Sentiment Analysis is a part of Natural Language Processing. It usually works with machine learning, data mining and computational linguistics. Usually, sentiment analysis handles analysis, automatic extraction and distinguishing of sentiments, beliefs, emotions or opinions communicated via written text. SA intends to decide the attitude of the user which may be judgemental or evaluation. The users' opinions are generally obtained in binary response like positive/negative. The embedding of SA in stock market prediction is essential because it easily reflects the changes in public viewpoints with respect to a specific stock. It helps the users decide whether the market will go up or down based on the aforementioned opinions on the web related to that stock. Naturally, this can be disintegrated into two sub-issues:

- Pre-processing, parsing and text refinement of the social media data.
- Analysing and scoring the items in the lexicon.

1.1 OBJECTIVE

To enhance the performance of the base financial lexicons, using logistic regression model based deeper and relative sentiment computation methods for items present in the lexicons.

1.2 MOTIVATION

The development and employment of information over the web have brought about the establishment of new research ranges in forecasting and prediction of stock market activities. A sentiment analysis framework, a totally mechanised framework which examines users' posts and anticipates user opinions, is one among them. The examination enthusiasm for this field is still high for most researchers because of the viable importance of the issue. The analysis of social media data may allow a deeper understanding of users' behaviour which can be utilised in the financial domain [2]. A great number of the online organisations have officially joined Sentiment Analysis frameworks with their administrations. Most of such frameworks incorporate information provided by microblogs such as Twitter and StockTwits.

To deter tweets' sentiment, one has to rely on text classification techniques. This subject has been considerably widened over the last ten years because of the multiplication of the data, both structured as well as unstructured, available online. The need for understanding the users' opinions has been decupled to help predict stock market variables. Many machine learning techniques can be used to assign classes to text documents. Unlike clustering used for unsupervised learning, these techniques allow for classification by training a computer on annotated data sets so that texts can be automatically labelled.

Stock market prediction strategies have as of recently gotten more noteworthy, with the help of technical analysis and fundamental analysis. Technical analysis is performed by applying machine learning techniques to stock market data that has been generated till date. On the contrary, fundamental analysis refers to the usage of human sentiment extracted from social media. It has effectively connected in otherworldly information examination and content mining. Until now, the greater part of the stock market organisations depends on the technical analysis's

machine learning models. Fundamental analysis is slowly being incorporated by organisations in order to maximise the accuracy of predictions since sentiments extracted from online users posts can prove to be very helpful as these posts tend to easily affect other users' opinions. Sentiment analysis is the key component in the fundamental analysis for stock market prediction. Various SA models are applied over a large set of microblog data in order to infer the opinion (positive or negative), which is then used in the prediction.

The goal of tweets classification is to use them as a time series supposedly correlated to another time series from stock market data. The time span may vary but frameworks usually consider one to six months of data. The function linking both time series is unknown, but it is possible to approach it by applying regression techniques. A very common tool is the Granger causality analysis, which helps to identify whether a time series has predictive information about the other or not.

Social media stages (e.g. Twitter, StockTwits, etc.) have numerous dynamic users who produce a tremendous measure of data by collaborating with stock market directives and with different users on the stages. For instance, up to December 2015, 320 million month to month dynamic users use Twitter and more than 40 million users utilise Stocktwits. These stages can document and utilise the delivered data to better serve their users. One of the administrations that a large part of the informal organisation stages gives is the recommendation administration.

Recommendation frameworks foresee the future inclinations of users depend on their past connections with the items. For instance, information extracted from data on past opinions of users can be utilised to make a recommendation on future stock market activities. The tremendous measure of data delivered by the users is utilised by a few distinct techniques to make recommendations, e.g. by supervised and unsupervised learning techniques.

1.3 GOAL

This thesis introduces an improved methodology for lexicon acquisition framework utilising logistic regression model which incorporates the profound learning of items scoring that takes relative sentiments into account. Logistic Regression Systems are known to be best suited

for dichotomous problems as these systems give a clear answer to problems via probabilities. In this work, we try to implement a logistic regression system for assigning relative scores to lexicon items with the help of manually labelled tweets. The main objective of this dissertation is to evaluate the efficiency of relative scored lexicon in stock market prediction.

1.4 ORGANISATION OF THE THESIS

Chapter 2 includes related works of lexicons with relative scores using logistic regression for a dataset. It also includes a literature review of sentiment analysis and the techniques that are being currently used for it.

In Chapter 3, Material and Methods, we firstly introduce about stock price prediction. After that, we explain about Microblogging data set, LDA-based method, JST-based method, and evaluation of these methods is fully explained. Then we also explain the role of Normalisation.

Chapter 4, Proposed Work, describes the analysis of regression techniques for stock prediction, after that, we focus on logistic regression with ROC curve analysis.

In Chapter 5, we discuss the data set used for the implementation. We elaborate the experimental results we obtained with implementations. Conclusion and future directions are presented in Chapter 6.

Chapter 2

BACKGROUND WORK

For almost ten years, microblogging has been spreading through the web and many researchers got an interest in this way of communicating. Because of its capacity to transmit ideas across people, a research identifies it as online word-of-mouth branding. Twitter has very often been considered the most straightforward choice by researchers for sentiment analysis and opinion mining on microblogging data. Indeed, it provides with a huge volume of opinionated data on a very broad range of subjects and has a free API for crawling tweets and users.

The first quantitative study on Twitter and its related information diffusion aimed at stock prediction date back to 2010. Since then, articles have multiplied and some authors have summarised and grouped disjointed research. In the recent years, many companies have found applications for global social media analysis but the research is still focused on particular markets and conditions. In this dissertation, we focus on authors using opinion mining and sentiment analysis for stock prediction applications or correlation analysis.

Considerable research has been done to acquire lexicons for sentiment analysis, below is a summary of researches conducted by various researchers.

An investigation on stock exchange information was made where at first they arranged the stock into groups and from that, a portfolio was assembled. The stock information utilised for bunching is the Bombay Stock Exchange information for the financial year 2007 to 2008. The portfolio was assembled utilising Markowitz view. They propose how to coordinate distinctive grouping procedures into portfolio administration. The hybridization of various procedures would additionally expand the portfolio proficiency [3].

Another mechanism expected to help clients in distinguishing collections of stocks that have comparative value development designs over some undefined time frame was developed

[4]. Here they group markets information and after that intends to utilise an unguided grouping calculation known as the Self-Organizing Map calculation. The dataset was taken from Yahoo fund graphs, S&P 100 stock information from Jan1, 2012 to Jul 30, 2012. The use of SOM is to delineate it in high measurements to two–dimensional plane. The calculation of first gains out weights on spatially sorted out hubs in an arbitrary manner and afterwards iteratively changes the weights. Euclidean Distance was utilised to decide the area of an information point. The real issue in utilising SOM calculation was its scalability.

Three noteworthy grouping calculations K-Means, Hierarchical grouping calculation and turn around K means were also considered and the execution of these three noteworthy grouping calculations on the building capacity for amending class acute group of calculation was evaluated [5]. An effective grouping strategy Hierarchical and Recursive K-means grouping (HRK) to anticipate the short stock value developments after the arrival of money related reports was proposed.

With the development of the Internet and Web 2.0 marvel, online networking is a critical huge information source. Clients spend a critical part of their time via web-based networking media. In this way, the investigation of this online networking information may permit a more profound comprehension of clients' state of mind that can be used for different purposes, including the budgetary area [6]. For example, Thomson Reuters Eikon and Bloomberg are cases of stock related administrations that incorporate supposition investigation of tweets.

The use of estimation and consideration pointers for stock exchange activities displaying and expectation is a dynamic subject. The opinion and consideration markers can be made utilising particular sources, supposition examination technique and mix technique used to blend particular sources. The stock related examination accepts a periodicity of the connected factors (e.g., day by day, month to month), sort of stock (Stocks, e.g., individual or portfolios), techniques used to display or anticipate and information used to fit the models. Probably the modern investigations (after 2011), perform a forecast that is portrayed by its information period and the factual tests used to confirm the measurable centrality of the slant and consideration

based pre-styles when contrasted with standard models, for example, the Diebold-Mariano (DM) test.

None of the related works endeavours to anticipate relative scores, which is tended to in this investigation. In the following couple of sections, we detail some of these measurements and clarify the curiosity of this work when contrasted with the related works. The prior examinations, from 1988 to 2010, embraced overviews, budgetary information, message loads and news to make the estimation and consideration pointers. After 2011, Web 2.0 administrations, for example, microblogs (e.g., Twitter, StockTwits) and Google seeks, have additionally been received. Some monetary measures and review values, for example, American Association of Individual Investors (AAII) and Investors Intelligence (II), are regularly utilised as an intermediary for the assumption. AAII and II are well-known supposition instruments that are made from surveys to speculators and bulletins made by showcase experts. Be that as it may, the pointers separated from writings (e.g., Twitter) have some advantages when contrasted with related works [22].

There are two principle approaches for the extraction of sentiment pointers from content: supervised and unsupervised. Many machine learning techniques can be used to assign classes to text documents. Unlike clustering used for unsupervised learning, these techniques allow for classification by training a computer on annotated datasets so that texts can be automatically labelled [7]. The majority of the connected dictionaries are area autonomous (e.g., General Inquirer, MPQA, SentiWordNet). Just two investigations utilise the stock related vocabulary made by Loughran and McDonald (2011). However, nonexclusive space autonomous dictionaries are ineffectual for surveying the assessment of stock exchange messages. For example, the expression "touchy" is frequently negative in casual settings yet can be certain inside the financial area ("hazardous ascent") [1]. Additionally, the financial vocabulary of Loughran and McDonald was made utilising vast content reports and it acquires low review esteems for short microblogging messages [2]. Accordingly, in this work, we utilise a current and broad dictionary, adjusted to microblogging stock exchange discussions and that should allow a more dependable estimation arrangement of any microblogging data.

It has also been stated that to assess the forecasting estimation of opinion for stock exchange factors, most of the examinations apply Multiple Regression (MR) and Vector Auto-Regression (VAR) techniques, which are summed up for study and monetary based supposition and exceptionally visit for the content based assumption. The use of more adaptable learning ML models, for example, SVM [8] [9] [10] or Neural Networks (NN), is all the rarer [11].

The forecast of overview estimation records can be extremely valuable for financial specialists. It might allow a significant suspicion of their esteems or constitute a shabby option measure. In spite of the fact that there are relapses of review feeling utilising contemporaneous estimations of different sources (e.g., money related measures, different studies) in a few examinations, we didn't discover their pre-lingual authority in view of slacked estimations of other slant intermediaries, particularly web-based social networking opinion [12]. In this investigation, we anticipate two well-known review assumption markers (AAII and II) utilising Twitter assessment pointers. With respect to acquired outcomes, content based opinion was considered helpful to settle on exchanging choices or foresee helpful stock exchange factors [9] [12]. For example, day by day or intraday estimations of stock costs, value bearings, returns, unpredictability and exchanging volume [8] [11] [23].

To the best of our insight, there are no examinations utilising microblogging information to anticipate factors with the help of relative scores, as executed in this work. Posting volume via web-based networking media administrations (e.g., microblogs and message sheets) has additionally been connected to foresee unpredictability or exchanging volume [13] [14]. The principle objective of this work is to evaluate the impact of microblogging information to the determining of stock exchange factors.

To foresee the future market cost of stocks numerous strategies have been proposed in the current days. In these systems, they have utilised distinctive sorts of strategies and some of them mix a blend of methods to foresee the future cost of a stock exchange. A neural system is a field of computerised reasoning where simulated neural system feedback calculation is utilised with the sustained forward neural system to foresee the cost of a stock exchange [15].

Hidden Markov Model has also been suggested for stock exchange value forecast. An HMM is a state machine for a framework follower to Markov handle with clandestine states. Particularly with respect to the time arrangement examination applications, on the off chance that we indicate the concealed state at time t as x(t) and the perception in the meantime as y(t) at that point the accompanying actualities are constantly valid in the HMM: x(t) is reliant just on x(t-1), y(t) is reliant just on x(t). In the way to forecast, first, there is a need to locate the most comparable day in stock exchange information for a particular day with the goal that it could be utilised to foresee the next day's nearby estimation. To do this first we have to figure the probability of earlier days in the coveted range. While having one day's stock information it is straight forward to figure the probability of that particular day from HMM [16]. Through utilising the probability resistance we bring a rundown of comparative days to yesterday's stock information and after that, we attempt to locate the best figure as the one that has the most noteworthy probability of all.

Chapter 3

MATERIAL AND METHODS

3.1 Introduction about Stock price prediction

Stock value expectation has dependably pulled in individuals inspired by putting resources into share market and stock trades as a result of the direct monetary advantages. It is additionally an imperative research theme in previous works. Forecast of stock exchange returns is an extremely complex issue and relies upon such a large number of elements. Nowadays stock costs are influenced because of many reasons like organisation related news, political, social practical conditions and cataclysmic events. A lot of studies were performed for the expecting of stock list organisations and in addition every day bearing of progress in the file. A large number of models were created for foreseeing the future cost of stocks however everyone has its own particular inadequacies. In this work, we research to foresee the stock costs utilising logistic regression model.

Forecast of Stock market returns is an essential issue and extremely complex in money related establishments. The expectation of stock costs has dependably been a testing assignment. It has been watched that the stock costs of any organisation don't really just rely on the money related status of the organisation but additionally relies upon the financial circumstance of the nation. Nowadays stock costs are influenced because of many reasons like organisation related news, political occasions, cataclysmic events and so on. Stock value forecast is a standout amongst essential issues to be examined in scholastic and monetary explores [23]. The quick information preparing of these occasions with the assistance of enhanced innovation and correspondence frameworks has made the stock costs change at an extremely fast rate. A lot of studies were performed to the expectation of stock file esteems and in addition the day by day heading of progress in the file. There are such a large number of models to foresee a cost of a stock exchange. To put cash in money markets we need a thought whether the costs of stocks will increment or reduction on the following couple of days. A few processing systems should be joined keeping in mind the end goal to foresee the idea of the share trading system. A

considerable measure of research has been occurring for a long time in anticipating the stock costs or stock record. It includes a presumption of key data that is openly accessible before that makes them anticipate connections to the future stock returns or lists. The specimens of such data incorporate monetary factors, for example, financing costs and trade rates, industry particular data, for example, customer cost, and organisation particular data, for example, salary explanations and profit yields. The estimation of the stock relies upon what number of individuals need to get it and what number of individuals are offering it. In the event that many individuals need to purchase a stock, the cost will go up. On the off chance that there are a bigger number of venders than purchasers, the cost will go down. Individuals typically purchase/offer offers in stocks with the assistance of a specialist. An agent additionally enables clients to use sound judgment in stocks. Most merchants have suggestions for the vast majority of the stocks. Generally, specialised investigation approach that predicts stock costs in view of authentic costs and volume, the Dow Theory, fundamental ideas of patterns, value examples and oscillators, is usually utilised by stock speculators to help venture choices. Progressed insightful strategies extending from immaculate numerical models and master frameworks to neural systems have likewise been utilised by numerous monetary exchanging frameworks for stock expectation. At last, the vast majority of the analysts have inferred the different strategies for anticipating future offer market costs utilising simulated neural system.

3.2 Microblogging data

Microblogging information has qualities which may be of potential incentive for determining stock exchange conduct. Administrations, for example, twitter and StockTwits have huge groups of investors sharing data about the stock exchange. These clients as often as possible communicate amid the day and respond promptly to occasions. Messages are typically exceptionally objective. Microblogging clients, for the most part, apply cashtags in securities exchange discussions to allude to the included stocks. Cashtags are made by a "$" character and the individual ticker (e.g., $AMZN). The use of cashtags licenses a simple and less loud determination of messages identified with particular stocks. The extraction of consideration and feeling from microblogging is quicker and less expensive than from customary sources (e.g., reviews) since information is immediately accessible effortlessly. We have used a set of 5000

manually labelled tweets for our analysis. The dataset is in CSV format and contains three columns namely ‘created at’, ‘tweet’, and corresponding ‘sentiment’ [2].

3.3 Microblogging sentiment

We selected data to utilise the primary distinction of the posting volume in light of the fact that there is a noticeable developing number of tweets amid the broke down day and age. To make the financial specialist assumption markers, we utilised the assessment scores delivered by notion investigation of all tweets. For example, a negative score demonstrates a bearish word and a positive assumption esteem shows a bullish word. To distinguish positive and negative fragments, we connected a similar approach utilised as a part of the vocabulary creation. The sentiment score of each tweet compares to the total of the estimation score of all vocabulary things show in the message. To satisfactorily confirm the nearness of vocabulary components, we executed the pre-processing assignments:

- Replace all cashtags by the tag "tkr";
- Replace all numbers by the tag "NUM";
- Replace all notices by "@user";
- Replace all URL addresses by "URL";
- Execute tokenization, Part of Speech (POS) labelling and lemmatization by applying Stanford CoreNLP (Toutanova, Klein, Manning, and Singer, 2003).
- Identify the positive and discredited fragments keeping in mind the end goal to apply the sufficient score.

The sentiment pointers are made utilising the scores delivered by the conclusion investigation [1].

3.4 Survey sentiment indicators

Survey sentiment pointers are often connected in learns about the examination of the assessment effect on stock exchange conduct. For example, AAII gives week after week estimations of the votes of their individuals to a survey scrutinising their opinion (bullish, bearish) on the share trading system for the following six months. Likewise, UMSC is a month to month assessment file developed from a buyer certainty review replied by an arbitrary gathering

of five hundred mainland US family units. Another case is the II week after week list in light of over large autonomous market brochures, with every bulletin being arranged as bullish or bearish. II measures might be more related to institutional estimation than AAII, in light of the fact that a number of the brochures' creators are showcase experts.

3.5 Models

3.5.1 LDA-based method

Latent Dirichlet Allocation (LDA) is an element extraction method for the lexicon. In this method, an improved LDA-based strategy is proposed that can beat the downside existed in the customary LDA techniques. It reclassifies the between-class scramble by adding a weight work as indicated by the between-class remove, which isolates the classes however much as could reasonably be expected. In the meantime, it anticipates the between-class dissipate into the invalid space of the inside class disperse that contains the most discriminant data. Consequently, the changing network formed with the Eigen vectors comparing to the biggest Eigen values of the exchanged between-class disperse can amplify the Fisher criteria [8].

We expelled the prevent words from messages. At that point, every one of the words is lemmatized by the Stanford Core NLP. We prepare the LDA on the preparation set and the points for concealed messages on the test set. Topics are construed by the Gibbs Sampling with 1000 emphases. We picked 50 as the quantity of points. From that point onward, the likelihood of every theme for each message is computed. Next, for every exchange information, the likelihood of every subject is characterised as the normal of the probabilities of that theme in the messages having a place with that exchange date. At that point, we coordinated the probabilities into the forecast show at the transaction dates t and $t-1$, respectively. Figure 1 demonstrates the graphical model portrayal of LDA. Notations in Figure 1 appear in Table 1.

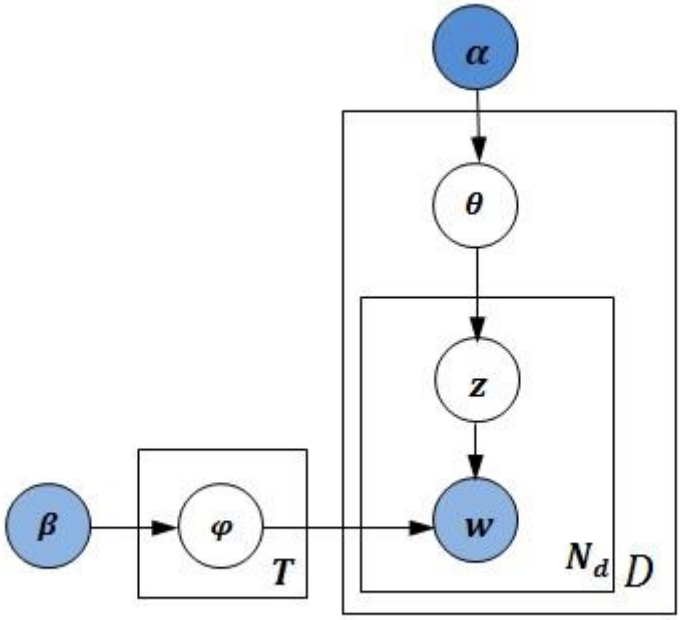

Figure 1. Graphical model representation of LDA [8].

Table 1. Notations in LDA

Notation	Definition
α, β	Hyperparameters
φ	The distribution over words
T	The number of topics
θ	The message specific topic distribution
z	A topic
w	A word in the message d
N_d	The number of words in the message d
D	The number of messages

3.5.2 JST-based method

The assessment is often expressed on a theme or viewpoint. At the point when people post the message on the web-based social networking to express their sentiment for a given

stock, they tend to talk their conclusions for a specific theme, for example, benefit and profit. In light of sets of subject feeling, they would surmise that the future cost of that stock goes up or down. From that instinct, we propose another element theme conclusion for the stock forecast shows. The first is an inactive point based model, the JST demonstrate [17]. The JST display was utilised to extricate points and assumptions at the same time. A word in the record is drawn from dispersion over the words characterised by the theme. After extraction of stop words and lemmatization, the JST demonstrate is prepared from the preparation set, and themes on the test set are construed by the Gibbs Sampling with 1000 cycles. We picked 50 as the quantity of themes and 3 as the quantity of suppositions. Next, the joint likelihood of each combine of subject and supposition is ascertained for each message. From that point forward, for every exchange date t, the joint likelihood of every theme match is characterised as the normal of the joint probabilities of that in the messages having a place with that exchange date. At that point, we incorporated these probabilities into the forecast display [8]. Figure 2 demonstrates the graphical model portrayal of JST. Notations in Figure 2 appear in Table 2.

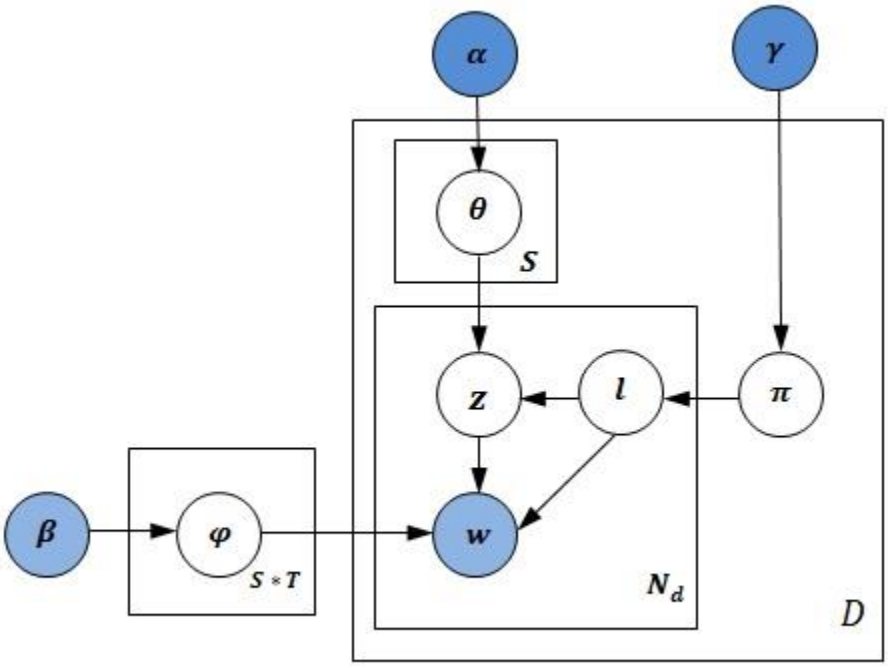

Figure 2. Graphical model representation of JST [8].

Table 2. Notations in JST.

Notation	Definition
α, β, γ	Hyperparameters
φ	The distribution over words
T	The number of topics
S	The number of sentiments
θ	The message and sentiment specific topic distribution
z	A topic
w	A word in the message d
l	A sentiment label
π	The message specific sentiment distribution
N_d	The number of words in the message d
D	The number of messages

3.6 Evaluation

3.6.1 Normalization

The goal of normalisation is to make an entire set of values have a particular property.

Min-max normalisation: Performs a direct change on the first information values. Assume that min_x and max_x are the base and greatest of highlight X. We might want to outline [min_x, max_x] into another interim [new_min_x, new_max_x] [18]. Thus, every value v from the first interim will be mapped into value new_v utilising the accompanying equation:

$$new\_v = \frac{v - min_x}{max_x - min_x} \cdot (new\_max_x - new\_min_x) + new\_min_x \quad (1)$$

3.6.2 Precision-Recall

In design acknowledgement, data recovery and parallel grouping, precision is the portion of significant examples among the recovered occasions, while recall is the part of pertinent occurrences that have been recovered from aggregate important cases in the picture.

As demonstrated in Figure 3, assume a PC program for perceiving mutts in photos recognises 15 pooches in a photo containing 24 puppies and a few felines. Of the 15 canines recognised, 10 really are puppies (genuine positives), while the rest are felines (false positives). The program's precision is 10/15 while its recall is 10/24.

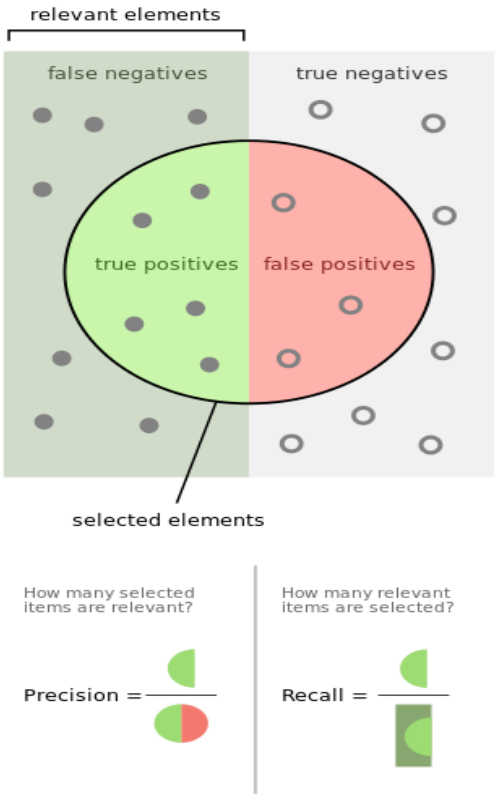

Figure 3. Precision and Recall

We can say that an increase in precision means that the output of an algorithm is more relevant than irrelevant, while an increase in recall means that an algorithm returned a greater number of the relevant outputs.

Chapter 4

PROPOSED WORK

4.1 Existing System Approach

Direct relapse is utilised to anticipate information and for displaying the connection between a scalar ward variable ‘y’ and at least one informative factors indicated ‘X’. There are progress in this field, yet the impediments continue as before. Basic Linear Regression is the one where just a single logical variable is utilised.

Disadvantages:

1. Considers just two sections of the dataset for examination.
2. The open value and close value is considered.
3. Be that as it may, the precision given is not palatable.

4.2 Proposed System Approach: Regression Techniques

In the factual illustration, regression evaluation is a preferable technique for assessing the relations among factors. It incorporates a number of techniques for displaying and evaluating a few factors when the emphasis is on the relation between a subordinate variable and at least one free factors. Regression evaluation helps in seeing how the common assessment of the required variable changes when any one of the autonomous variables is differed, while the other free variables are fixed. In regression evaluation, it is important to illustrate the difference in the reliant variable around the regression which can be shown by a likelihood dissemination [19].

4.2.1 Logistic regression

It is an assessing strategy for breaking down a dataset in which there are at least one autonomous factors that decide a result. The result is measured with a dichotomous variable (in which there are just two conceivable results).

In logistic regression, the reliant variable is paired or dichotomous, i.e. it just contains information coded as 1 (TRUE, positive, achievement, and so on.) or 0 (FALSE, negative, disappointment, and so on.).

The objective of logistic regression is to locate the best fitting (yet organically sensible) model to depict the connection between the dichotomous normal for intrigue (subordinate variable = reaction or result variable) and an arrangement of autonomous (indicator or informative) factors [20]. Logistic regression creates the coefficients (and its standard mistakes and criticalness levels) of a recipe to anticipate a logit change of the likelihood of quality of the normal for intrigue:

$$\text{Logit}\,(p) = \Theta_0 + \Theta_1 X_1 + \Theta_2 X_2 + \Theta_3 X_3 + \ldots + \Theta_k X_k \qquad (2)$$

In our work, we see the logistic regression framework as a calculation which takes a dataset of labelled tweets between an arrangement of tweets and an arrangement of lexicon items and endeavours to figure how a given items' permutation may affect different tweets. For instance, an item when used with a set of words may contribute to a positive sentiment, while the same item, when used with another set of words, may contribute to give negative sentiment. The centre data in the dataset for this situation would indicate the lexicon items. For a given tweet, a logistic regression framework would score a list of items <L1; L2… … Lk> which occur in this tweet, based on the relative presence of other items.

4.2.2 Regression coefficients

The logistic regression coefficients are the coefficients Θ_0, Θ_1, Θ_2 ... Θ_k of the regression equation:

$$\text{Logit}\ (p) = \Theta_0 + \Theta_1 L_1 + \Theta_2 L_2 + \Theta_3 L_3 + \ldots + \Theta_k L_k \quad (3)$$

The strategic regression coefficients demonstrate the change (increment when $\Theta_i>0$, diminish when $\Theta_i<0$) in the anticipated logged chances of having the normality for an increase for a one-unit change in the free factors [21].

At the point when the autonomous factors L_a and L_b are dichotomous factors (e.g. Smoking, Sentiment) the impact of these factors on the required variable can basically be thought about by looking at their regression coefficients Θ_a and Θ_b. Figure 4 below demonstrates the functioning of our proposed approach.

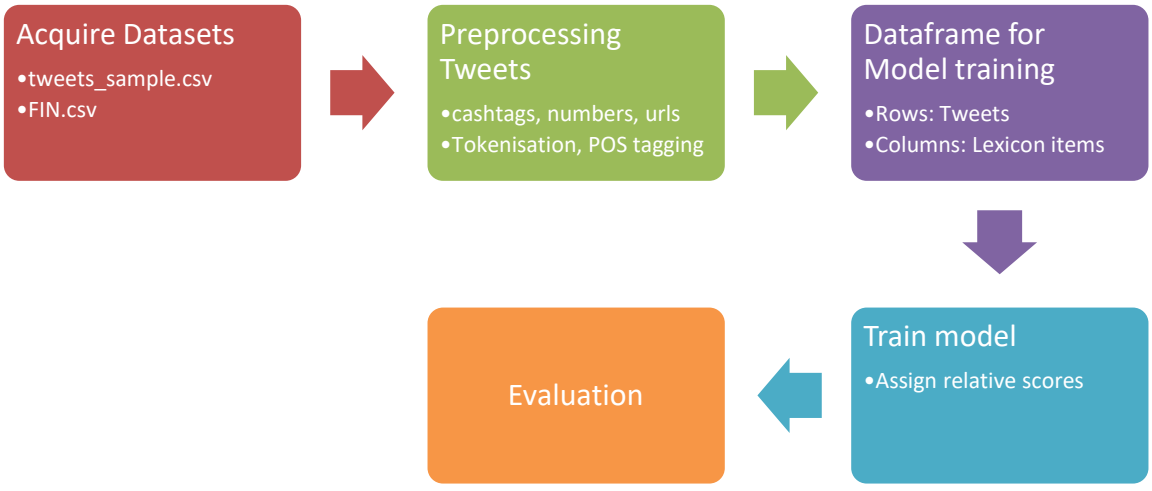

Figure 4. Flowchart of Proposed Algorithm

4.3Proposed System Algorithm

Input: Tweets dataset (tweets.csv), Lexicon dataset (FIN.csv)

Begin

1. Create an empty data frame with tweets as rows and lexicon items as columns.
2. Logistic regression performs operations on the data frame where the objective values have been characterised as of now.
3. The relations which regression sets up amongst indicative and target variables can make an example. This example can be utilised on different lexicon items with their sentiment values are not known.
4. The preparation informational collection is utilised to train a model and to evaluate the obtained coefficients of the auto regression state. These coefficients are assessed.
5. The assessed coefficients are utilised to anticipate the sentiment of an item.
6. Normalise the coefficients using min-max normalisation.
7. Then the coefficients are utilised to test the informational index and the examination is done between genuine sentiment and anticipated sentiment.

End

Output: A new lexicon (FIN_new.csv) having relative scores of items.

Interpretation of the fitted logistic regression equation

The logistic regression equation is:

$$\text{Logit (Sentiment)} = \Theta_0 + \Theta_1 * L_1 + \Theta_2 * L_2 + \ldots + \Theta_n * L_n \quad (4)$$

Where Θ_i represents the weight of i^{th} lexicon item L_i assigned during and after training the model.

ROC curve analysis

ROC curve analysis is a strategy to assess the calculated regression model. In this examination, the energy of the model's anticipated values to segregate amongst positive and negative cases is measured by the Area Under the ROC curve (AUC). The AUC is the value that changes from 0.5 (worthless) to 1.0 (excellent). To play out a full ROC curve investigation on the anticipated probabilities we can spare the anticipated probabilities and next utilise this new factor in ROC curve examination.

Chapter 5

RESULTS

5.1 DATA SETS

The dataset for tweets is taken from Nuno Oliveira's stock_market_sentiment repository: https://github.com/nunomroliveira/stock_market_sentiment/blob/master/twt_sample.csv.

This data set consists of:

- 5000 manually labelled tweets.
- Each tweet has its corresponding creation time.
- The next column is the tweet data itself.
- And the final column is the sentiment that is manually assigned.

The lexicon used is created by Loughran and McDonald and can be obtained from http://www3.nd.edu/~mcdonald/Data/Finance_Word_Lists/LoughranMcDonald_Negative.csv which contains all the negative words and http://www3.nd.edu/~mcdonald/Data/Finance_Word_Lists/LoughranMcDonald_Positive.csv which contain all the positive words.

5.2 EXPERIMENTAL RESULTS

Finally, the results were analysed. We carried out all our experiments on Intel® Core™ i5-3230 @ 2.60GHz, 4.0 GB RAM computer and run R on Windows 10 Enterprise edition (64 bit) to simulate the methods. We have used accuracy, precision, recall and f-measure for evaluation of the proposed work and comparison with the related works. The formulae for each of them is given below:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (5)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{F-1 Measure} = 2 * \frac{Precision*Recall}{Precision + Recall} \quad (8)$$

The experimental results for the methods compared and our own approach are shown in Table 3.

Table 3. Experimental Results

	Accuracy	Precision	Recall	F1-measure
JST	49	63	41	49
LDA	53	64	54	58
FIN_old	52	66	38	48
Proposed	57	67	72	69

The following figure shows the evaluation measures for JST based method, LDA based method, Old lexicon method and our approach i.e. improved lexicon with relative scores. Evaluation time for our proposed method is 27 seconds. We can conclude from the evaluation chart that the accuracy of the proposed method is increased in comparison to the other methods. The percentage increase in accuracy is 7.54%.

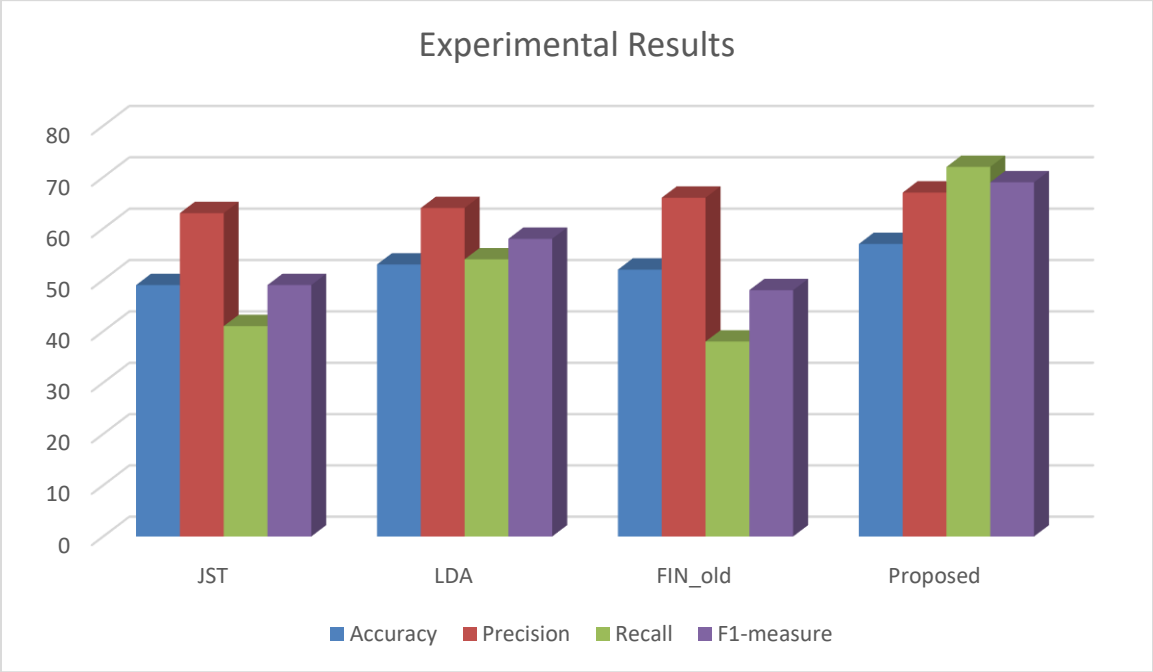

Figure 5. Experimental Results

Chapter 6

CONCLUSION AND FUTURE WORK

Stock exchange prediction is the strategy for deciding future estimations of an organisation's stocks and other money related values. Stock exchange forecast with the assistance of regression examination is the most productive mix to anticipate the stocks and the states of the market. The market does not have an effective programming where the correct proposals of accessible stocks and the best possible venture investigation are introduced in a productive way. The speculators ought to be guided and urged to put resources into the stocks soundly. The advancement of an energetic application for breaking down and anticipating stocks exchange costs is a fundamental device gone for expanding the rate of speculators enthusiasm for stocks exchanges. This theory clarifies the improvement and execution of a stock value forecast application utilising machine learning calculation and question arranged approach of programming framework advancement.

The proposed work utilises regression examination as an information mining procedure and builds up a framework for using time arrangement information in money related matters. An expectation framework has been constructed that utilises information mining procedure to deliver intermittently insights about stocks exchange costs.

There are many research headings which may be considered later on work. Enhancing the precision of the prescient models is one of them. Precision can be enhanced by considering an altogether extraordinary angle i.e., human assumptions. As the future extent of the stock exchange is boundless, the interest for its information investigation will be constantly expanding. By changing just the preparation information, the proposed framework can be utilised for any stock exchanges of different nations. With a few altercations, the framework can be utilised for different purposes, for example, foreseeing costs of wares like gold, anticipating the fuel utilisation of a vehicle and in addition observing soundness of a patient.

Foreseeing the stock exchange cost is extremely prevalent among speculators as financial specialists need to know the increments that they will get for their ventures. Generally, the specialised experts and intermediaries used to foresee the stock costs in light of verifiable costs, volumes, value designs and the fundamental patterns. Today the stock value expectation has turned out to be extremely unpredictable than before as stock costs are influenced because of organisation's budgetary status as well as due to socio conservative state of the nation, political environment and catastrophic events and so forth. We have demonstrated the examination between the anticipated cost and genuine cost. As it unmistakably obvious from the results that, our expectation cost nearly harmonises with the real stock cost. This strategy for anticipating the return on venture will help extraordinarily to budgetary organisations and stock agents to foresee the future cost in such indeterminate conditions. Further lexicons can be created with a domain oriented view in mind. The domains can vary from agricultural to technological to administrative etc.

REFERENCES

[1] N. Oliveira, P. Cortez, and N. Areal, “Stock market sentiment lexicon acquisition using microblogging data and statistical measures,” *Decis. Support Syst.*, vol. 85, pp. 62–73, 2016.

[2] N. Oliveira, P. Cortez, and N. Areal, “The impact of microblogging data for stock market prediction: Using Twitter to predict returns, volatility, trading volume and survey sentiment indices,” *Expert Syst. Appl.*, vol. 73, pp. 125–144, 2017.

[3] D. Lee and H. Seung, “Algorithms for non-negative matrix factorization,” *Adv. Neural Inf. Process. Syst.*, no. 1, pp. 556–562, 2001.

[4] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization.,” *Nature*, vol. 401, no. 6755, pp. 788–91, 1999.

[5] M. S. Babu, N. Geethanjali, and P. B. Satyanarayana, “Clustering Approach to Stock Market Prediction,” vol. 1291, pp. 1281–1291, 2012.

[6] W. Fan and M. D. Gordon, “Unveiling the Power of Social Media Analytics,” *Commun. ACM*, vol. 12, no. JUNE 2014, pp. 1–26, 2013.

[7] D. P. Bertsekas, “On the Goldstein — Levitin — Polyak Gradient Projection Method,” *IEEE Trans. Automat. Contr.*, vol. 21, no. 2, pp. 174–184, 1976.

[8] T. H. Nguyen, K. Shirai, and J. Velcin, “Sentiment analysis on social media for stock movement prediction,” *Expert Syst. Appl.*, vol. 42, no. 24, pp. 9603–9611, 2015.

[9] R. P. Schumaker, Y. Zhang, C. N. Huang, and H. Chen, “Evaluating sentiment in financial news articles,” *Decis. Support Syst.*, vol. 53, no. 3, pp. 458–464, 2012.

[10] S. Deng, T. Mitsubuchi, K. Shioda, T. Shimada, and A. Sakurai, “Combining Technical Analysis with Sentiment Analysis for Stock Price Prediction,” *2011 IEEE Ninth Int. Conf. Dependable, Auton. Secur. Comput.*, pp. 800–807, 2011.

[11] J. Bollen, H. Mao, and X. Zeng, “Twitter mood predicts the stock market,” *J. Comput. Sci.*, vol. 2, no. 1, pp. 1–8, 2011.

[12] A. Al Nasseri, A. Tucker, and S. De Cesare, “Quantifying StockTwits semantic terms’ trading behavior in financial markets: An effective application of decision tree algorithms,” *Expert Syst. Appl.*, vol. 42, no. 23, pp. 9192–9210, 2015.

[13] F. T. Jesus *et al.*, “Lethal and sub lethal effects of the biocide chlorhexidine on aquatic organisms,” *Ecotoxicology*, vol. 22, no. 9, pp. 1348–1358, 2013.

[14] T. O. Sprenger, A. Tumasjan, P. G. Sandner, and I. M. Welpe, “Tweets and trades: The information content of stock microblogs,” *Eur. Financ. Manag.*, vol. 20, no. 5, pp. 926–957, 2014.

[15] P. K. Sahoo and K. Charlapally, “Stock Price Prediction Using Regression Analysis,” *Int.*

J. Sci. Eng. Res., vol. 6, no. 3, pp. 1655–1659, 2015.

[16] D. Palguna and I. Pollak, "Mid-Price Prediction in a Limit Order Book," *IEEE J. Sel. Top. Signal Process.*, vol. 10, no. 6, pp. 1083–1092, 2016.

[17] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," *Proc. 18th ACM Conf. …*, pp. 375–384, 2009.

[18] C. Chisalita-Cretu, "A multi-objective approach for entity refactoring set selection problem," *2nd Int. Conf. Appl. Digit. Inf. Web Technol. ICADIWT 2009*, pp. 790–795, 2009.

[19] Y. E. Cakra and B. Distiawan Trisedya, "Stock price prediction using linear regression based on sentiment analysis," *ICACSIS 2015 - 2015 Int. Conf. Adv. Comput. Sci. Inf. Syst. Proc.*, pp. 147–154, 2015.

[20] M. Xu, Y. Lan, and D. Jiang, "Unsupervised Learning Part-Based Representation for Stocks Market Prediction," *Proc. - 2015 8th Int. Symp. Comput. Intell. Des. Isc. 2015*, vol. 2, pp. 63–66, 2016.

[21] B. S. Bini and T. Mathew, "Clustering and Regression Techniques for Stock Prediction," *Procedia Technol.*, vol. 24, pp. 1248–1255, 2016.

[22] Hosmer, David W., Stanley Lemeshow, and Rodney X. Sturdivant. Introduction to the Logistic regression model. John Wiley & Sons, Inc., 2010.

[23] Sabherwal S., Sarkar S.K., Zhang Y. "Do internet stock message boards influence Trading? Evidence from heavily discussed stocks with no fundamental news", *Journal Business Finance and Accounting*, 2011, vol.38, Issue 9-10, pp. 1209-1237