**Chapter 1**

**Introduction**

Social media has become an integral part of everyone’s life. Twitter is the most prevalent social networking service where a large number of users share vast amount of information on a daily basis. In this project we have used a method to improve the efficiency of the existing methodologies by including news articles along with tweets for scrutinizing the public sentiment and preprocessing the tweets by adding specific weights to each tweet based on authenticity and followers of twitter, impact of the tweet and the number of times it has been retweeted.

**1.1 Description**

In today’s era of digitization, where information technology is growing progressively, the incessant use of social media has reached miraculous levels. It has become the perfect platform for people to share their point of view. Public opinions and information is shared on a large scale daily.These opinions have a significant impact on the collective conscience of the society. Twitter, a social media platform with 275 million users worldwide is one of the most used social networking services. The information extracted from thesetweets can be considered highly reliable as an investor information source. Research shows that the collective conscience created by tweets can be used effectively for predicting the companyearningsandtheirsubsequentimpactonthestockprices.

Sentiment Analysis is the process of recognizing the sentiments expressed in the textual data. It classifies the data as negative, neutral and positive. For sentiment analysis emoticons are considered along with text for more efficient results. Further, it also considers the authenticity of the tweets i.e. the number of followers; whether the tweet is posted from a verified account and the number of times the tweets are retweeted to improve the accuracy of the sentiment analyzer. Twitter is a valuable resource for learning about users’ feelings, emotions, behaviors, and decisions that reflect their opinions.

**1.2 Problem Formulation**

## To predict the Stock Market movements based on Public Sentiments expressed on Twitter and News Articles.We will be integrating news data and twitter data to predict stock market in our system to improve the efficiency of the existing system by including relevant news articles along with tweet based on authenticity and followers of the user .

**1.3 Objective**

The Objective of this study is as follows:

* To detect public sentiments expressed on twitter and in news articles related to stock market.
* Toanalyzetheeffectofdifferentsentimentsonstockmarketmovements.
* To aware investors about stock market variations inadvance*.*

**1.4 Proposed Solution**

Following will be the outcomes after completion of this project -

* + - Predicting stock market movement in advance based on tweets andnews.
    - Providinginvestorsrecommendations,toinvestincompanyornot.
    - Smart use of tweets and news for smart investment.

**1.5 Scope**

The system will only focus on predicting stock prices for a particular group of companies.It is possible that other factors may affect the stock behavior but we are limiting are scope for predicting the stock by considering only tweets and news articles.The prediction considers only past 7 days twitter data.

**Chapter 2**

**Review of Literature**

Twitter is one of the effective tools for any business intelligence to get information about what people are talking and reacting about the topics that are roaming around the world. A twitter helps to engage the users and directly communicates with them and in response, users to provide word-of-mouth marketing for companies by discussing the product quality. With the limited resources and knowing about no one can target directly to the destination consumers, the business intelligence can be more efficient in their policy of marketing by being very selective about consumers choice.

News articles serve the purpose of spreading company’s information to the investors either consciously or unconsciously in their trading strategies on the stock market. Because of the immense growth of the internet in the last decade, the amount of financial articles has experienced a significant growth. It is important to analyze the information as fast as possible so they can support the investors in making the smart trading decision before the market has had time to adjust itself to the effect of the information.

**2.1Sentiment Analysis on Twitter Data-set using Naive Bayes Algorithm**

**2.1.1 Fetching Twitter Data**

Using Twitter API Develop a twitter API [4] for downloading the tweets. The Twitter API directly communicates with the Source and Sink. The source is twitter account and the sink is HDFS (Hadoop Distributed File System) where all the tweets are stored.

**2.1.2 Pre-processing of tweets.**

1) Removal of URL’s: If any user posted any link which is none of the use for sentiment analysis. Therefore, URL should be removed from the tweet.

2) Removal of special symbol: There are various types of symbols used by the user such as punctuation mark (!), full stop (.) etc. which does not contain sentiment. Therefore, special symbols should be removed from the tweet.

3) Converting emoticons: Table. I. shows the various emoticons used for conversion. Nowadays emoticons become a way for the user to express their views, feeling, and emotion. Emotions play a big role in the sentiment analysis. Therefore, convert the whole emoticons into its equivalent word by which we can do the analysis efficiently.

4) Removal of Username: Every Twitter user has a unique username, therefore, anything is written by a user can be indicated be writing their username proceeding by @. This type is

denoted as proper nouns. For example, @username. This also has to be removed for effective analysis.

5) Removal of Hash tag: A hash tag is a prefixed with the hash symbol (#). Hash tag are used for naming subjects or phrases that are currently in trend. For example, #google, #twitter.

6) Removal of additional white spaces: There may be consists of extra white space in the data and it needs to be removed. By removing white spaces the analysis to be done more efficiently.

**2.1.3 Applying Naive Bayes Algorithm**

**2.1.3.1 Map Phase**

The working of Map phase [5] consists of two major tasks. First,creating a hash map for retrieval of polarity of each word. Secondly, processing the overall polarity of the tweets by applying Naïve Bayes algorithm. The map () method in MapReduce phase reads the content of the SentiwordNet dictionary from a file and transform into the Hash map for keyvalue based polarity retrieval of words. From here, the polarity of each word is stored in the hash map for faster processing. Now, the map() method read tweets line by line from the file. Map method parses each and every word and generates tokens. Each token has polarity available in the hash map. The polarities is fetched for each word and calculate the overall polarity of a single tweets using probabilistic model.

**2.1.3.2 Reduce Phase**

The reduce() method collects the overall polarity of each tweets and transform into 5 different categories as extreme positive, positive, extreme negative, negative and neural. The reduce() method iteratively work to collect various sentiments and based on polarities it classify and write the output on HDFS.

**2.2 A Fuzzy based Soft Computing Technique to Predict the Movement of the Price of a Stock:**

IIbrahim M. Hamed et al. [6], proposed an intelligent system to recommend the price of the stock. They have used multilayer perceptron-based ANN for the research. The model proposed in their research had several stages. Input selection, pre-processing of data and signal detection were some of the critical stages in the model. The algorithm was used in the model update the weights between the neural network neurons to minimise the error of the prediction results. ANN is a slow learner, and moreover, the multilayer perceptron leads to taking more time for the prediction of results. Monica

Sunil Kumar Khatri et al. [7] have proposed a sentiment analysis to predict Bombay stock exchange using ANN. Data of blogs/forums and E – media were the inputs for the system and pre-processed into four different moods like happy, hope, sad and disappointing. The ANN model was trained with sample data from a detailed set and produces the final prediction from the

test data. The output of the ANN model prediction will be positive, negative/neutral. The research entirely based on the sentiment of the users in the forum and there was no relation with the real price of the stock. The accuracy is the primary concern of the research as it depends on the sentiment shared in the forum.

**2.3 Using News Articles to Predict Stock Price Movements:**

According to the “efficient market hypothesis”, in financial markets profit opportunities are exploited as soon as they arise, hence stock prices follow a random walk and are extremely difficult to Task, data, and system overview since profit opportunities in the stock market are

present for only an extremely short period of time, high frequency information is essential to profitable trading strategies.

Indicators can be of two types: those derived from textual data (news articles), and those derived from numerical data (stock prices.We obtain indicators derived from textual data by learning a naïve Bayesian text classifier for higher level, relative price movements of stocks. We then use this trained naïve Bayesian classifier to compute the probability for every new, stock-specific news article that that particular news article belongs to a class representing a particular movement

**2.3.1 Identification of movement classes:**

Unlike in the usual text classification framework in our task news articles are initially unlabeled. Defining classes and obtaining labels for training examples is crucial for any classification task. The following subsections describe in detail our approach to accomplish this task.

**2.3.2 Aligning of news articles:**

One commonly used method to evaluate the performance of a particular stock is based on the volatility of the stock, which is known as the ß-value. In short, this value describes the behavior or movement of the stock relative to some index, and is calculated using a linear regression on the data points (∆index-price, ∆stock-price). Hence a stock with a ß-value of 1 has the characteristic that whenever the percent change for the index price is δ the percent change for the stock price is expected to be δ as well. Similarly a stock with a ß-value of 2 has the characteristic that whenever the percent change in the index price is δ the percent change in the stock price is expected to be 2δ. Stocks with a ß-value greater than 1 are relatively volatile, while stocks with a ß-value less than 1 are more stable. Labeling news articles A movement of zero at a particular time and for a particular window of influence means that the movement of the stock at that time is as predicted or expected