

Scope of Sentiment Analysis on News Articles Regarding Stock Market and GDP in Struggling Economic Condition

Sandipan Biswas¹, Ahona Ghosh^{1#}, Srabanti Chakraborty², Sandip Roy^{1*}, Rajesh Bose^{1s}

¹Brainware University, India, sandipan_diet@rediffmail.com

²Elite Institute of Engineering & Management, India, srabanti2k@gmail.com

[#]ahonaghosh95@gmail.com, ^{*}sandiproxy86@gmail.com, ^sbose.raj00028@gmail.com

ABSTRACT

The present paper studies the scope of sentiment analysis on news articles on impact of economic downfall of the primary international stock markets, comparing it with that of previous terrorist attacks or epidemic (9/11, Covid-19). For properly embedding our investigation in the theoretical outline, first we present an overview of existing studies focusing on the consequence of terrorism on economic markets. The practical part is mainly in terms of measuring whether the return of the New York Stock Market's main index due to the attacks vary statistically regarding the deviations of the four months before and after the attack. We study the internet data, contrasting the direct impact of news spread through news media affected the index of the New York stock market comparing with gold and crude oil price. The result analysis discusses interpretative hypotheses involving a behavior evolution of the stock markets as a result of the terrorist attacks or epidemic which find out the scope of more future works with sentiment analysis on news data in stock market.

Key words: Economic downfall, GDP, Lexicon based approach, Machine learning, News articles, Pandemic, Sentiment analysis, Stock market, Terrorist attack.

1. INTRODUCTION

Stock Prices are very dynamic in nature and able to adapt quick changes due to its fundamental nature of the monetary area. Researchers in financial domain conclude that the news article, blogs, and stock market prediction is important topics in many business incomes. Stock market forecasters focus on developing a successful approach for forecast the values or stock prices. The stock market prices are more fluctuating that's fall the stock price or raising the stock price. The aim is to earn high profit using well defined different trading strategies. The overall mood of social data according to the company that can be important variables which affect the stock price of the company. The different on-line social network sites that helps to make availability of large amount of data. Therefore, comparing information from social media

data with historical price can progress the forecast and accuracy of a system.

Twitter is presently the mostly used social microblogging stage [1] which allows its operators to send and read precise messages containing up to 140 characters, called tweets. It provides different services to create and share massive amount of data. From Twitter analysis of company's product can improve relationship and trust between customer and producer according to need of customer. Sentiment analysis is a study that addresses opinion or view based natural language processing. Such studies include feeling and mood detection, ranking, significance measuring related to [7]. G-POMS, Google N-gram, Lingmotif, LIWC, POMS, Opinion Finder, SentiStrength are the different analytical tools provide to calculate sentiment analysis of given data. The problem definition is to develop a model for sentiment analysis within big data distributed platform for stock forecast. To apply clustering and SVM classifier on sentiment score to improve accuracy and implement the model in distributed environment to speeds up the performance. The dataset must be filtered by adding metadata such as exact location of a person, the number of re-tweets, the number of followers to selected dataset. To parallelize the computation using Map-reduce distributed environment.

The next section reviews state of the art. Section 3 discusses different areas where data science and sentiment analysis applications lie. Section 4 and 6 describes sentiment classification techniques and data collection methods used in the recent works of our concerned domain whereas in section 5, graphical representation of worldwide stock market record downfall has been there. Data processing components are described in section 7 and after analyzing the existing approaches in section 8, section 9 presents case study analysis of the proposed/future work. Concluding statements and possible future enhancements are discussed in Section 10 finally.

2. RELATED WORK

In the current decade, sentiment analysis has been a very emerging concept in different research areas. Many researchers study and aim to identify a method to predict

sentiment analysis on various fields. Social media is popularized source of data, which collect useful data like micro-blogs, blogs, Facebook, Twitter etc. Skuza et al. in [1] covered the evaluation of a system to predict future stock price based on analysis of social media data. Real time Tweets are saved in using Twitter Streaming API. The large amount of data to be classified using Naive Bayes method for quick training of a huge volume of training data. The stock market prediction should be calculated by using linear regression technique. Pagolu et al. in [2] presented two different word-based representations, N-gram and Word2vec to analyze different sentiments reflected in different tweets. The author used supervised machine learning techniques and sentiment analysis (such as logistic regression, random forest, SMO) to tweets extracted from twitter and analyzing the correlation between tweet sentiments and stock market movement of company. A data can be extracted from twitter API of Microsoft using keyword \$MSFT, #Microsoft, etc. Ding et al. in [3] created a forecasting system of stock market activities of a particular day, depending on time series data and marketplace sentiment study. They collect values of S&P 500 between January 2008 and April 2010 from Yahoo! Finance into Excel spreadsheet. For analyzing the sentiment, they got Twitter Census stock Tweets dataset from Info-chimps, a private company offering a “data marketplace”. Naive Bayes Classifier used to analyze sentiment in the tweet data set. The SVM, Logistic and Neural network techniques would be used for predicting market movement.

Sumbureru et al. in [4] focused on the forecasting of regular stock activities of three businesses mentioned in the National Stock Exchange (NSE). The Support Vector Machine (SVM) has been applied for prediction of the stock market. The tweets collected were of 5-month period having 200000 tweets. The tweets were collected directly from twitter using

Twitter API and filtered using keywords for example #airtel. The relevant stocks were downloaded directly from yahoo finance. Soni et al. in [5] have performed sentiment analysis of a product by extracting tweets about products and classifying the tweets that can be as positive and negative sentiment. A mixed approach combining unsupervised learning and clustered tweets and then performing supervised learning techniques for classification has been observed in this paper, 1200 tweets were collected for the company “Apple” for analysis. The proposed model would be compared with SVM, CART, Random forest, Logistic regression. The predicted and actual value can be compared by means of confusion matrix. Zhang studied the efficiency of several machine learning methods [6] to provide the sentiment reflected in a tweet in terms of positive or negative sentiment. The author applied support vector machine, Naive Bayes, Maximum entropy, etc. and compared them. A correlation among stock prices and twitter sentiments has been observed and the words in tweets which correlate to change in stock price has been determined by performing a tweet analysis related to price change and tweets. From the literature study,

it can be concluded that for sentiment analysis of bigger dataset made to be accurate and efficient we need to make use of distributed approach. In this paper, a distributed model with supervised and unsupervised technique to improve accuracy and performance has been introduced.

3. DATA SCIENCE AND SENTIMENT ANALYSIS APPLICATION

Data Science is controlling almost every industry worldwide nowadays. No industry presently in the world can ignore the use of data. So, data science is becoming the fuel of several industries. Industries like education, e-commerce, transport, manufacturing, finance, banking etc. use data science and several data science applications have been eventually available related to those. In this paper, the transformations of the whole world due to different data science applications have been studied by us. The way of data perceiving has been revolutionized and at the end, different situations where data has been used to improve the quality of industries will be discussed by us as shown in Figure 1.

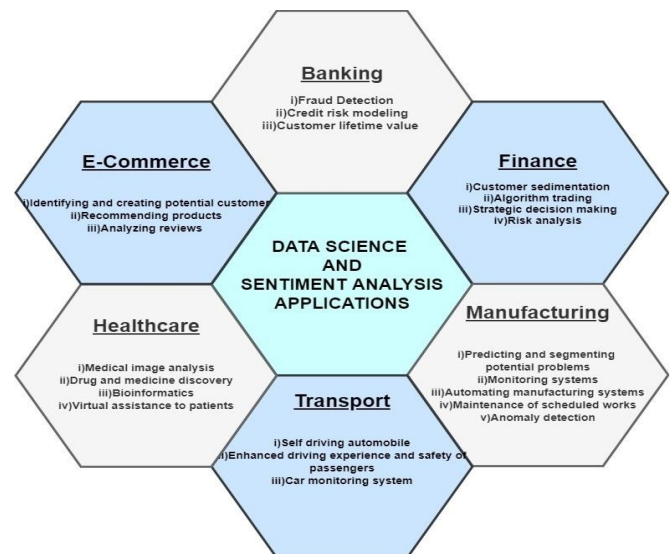


Figure 1: Application of Data Science and Sentiment Analysis

3.1. Banking

Banking and investment have been one of the major application areas of Data Science and it has allowed banks to be active in the competition. Using Data Science, banks are able to control the concerned resources effectively; additionally, banks are able of taking smarter decisions through by detecting fraud, organization of client data, modelling risk, real-time forecasting, client division, and so on. Evaluating the client era, banks display the number of clients they have. It provides them some forecasts, which the corporate bank will originate through the clients. For detecting fraud, banks permit the businesses to identify frauds involving accounting, insurance, and credit card. Banks get able of analysing investment outlines and series of clients to offer several suggestions according to them.

Moreover, banks can model the jeopardy by applying data science where the performance assessment can be done and

banks get the capability of tailoring customized marketing satisfying the customer requirements. Real-time predictive analytics used in banking sectors apply machine learning algorithms for improvement of their analytics plan. Furthermore, the real-time analytics help to understand the fundamental problems, which obstruct the system and degrade their performance.

3.2. Finance

Data Science has been playing a key role in automation of different monetary sectors from the last decade. Just like the way, banks have been able to automate risk analytics; data science has also helped finance industries in this purpose. Strategic planning in a company needs to be carried out using data science and machine learning helps to recognize, display and arrange the risks. Several machine-learning procedures improve cost efficacy and model sustainability by training a hugely obtainable client data. Similarly, monetary organizations apply machine learning for prognostic analysis, which lets the companies to forecast client life era and the concerned stock market changes.

Data Science too becomes a vital part in algorithmic transaction. Over severe data analysis, monetary organizations get able to make data-driven choices. It is also playing a vital role in making the customer practices improved for the consumers. Through wide analysis of client knowledge and alteration of likings, monetary organizations are able to generate a personalized association with their clients. This is further increased by the real-time analysis of clients, which increases the customization. Through several customer sentiment analysis techniques and machine learning procedures, we can increase the social media communication, increase their feedback and analyse customer evaluations. In addition, the extra machine learning methods like natural language processing and data mining have contributed to the alteration of information for keener governance, which helps to raise the effectiveness of trades.

3.3. Manufacturing

In 21st century, experts of Data Science are playing the role of workshop labours. They are holding a crucial role in the industrial sectors. Wide application of data science in industrial sectors to optimize manufacture and reduce costs and simultaneously increasing the profits. Moreover, with the accumulation of tools such as Internet of Things (IoT), it has allowed the businesses to forecast possible glitches, screen the systems and analyse the unceasing data stream.

Additionally, with data science, industries have been able to screen and control their costs and to improve their manufacturing duration. With a detailed study of client feedbacks, data scientists always support the trades in making improved choices and enhance the efficiency of the goods. Additional vital feature of it in productions is Automation. Using existing and real-time data as well, the industries are being capable of producing self-directed systems which can help to enhance the construction of industrial lines. The redundant tasks have been removed by it and powerful

technologies applying machine learning approaches including reinforcement learning have been introduced.

3.4. Transport

Data science's additional significant application area is transportation. Here, Data Science is vigorously creating its spot to make harmless driving surroundings for the motorists. It is also used to optimize the means of transportation performance and adding better self-sufficiency to the drivers. Furthermore, Data Science has actively augmented its application by introducing self-driving cars. Through widespread analysis of fuel consumption configurations, driver performance and lively vehicle controlling, data science has produced a strong position in the transport sector. The self-driving car is one of the most trending areas in the world today. By introducing autonomy to vehicles by reinforcement learning, vehicle producers have been able to create intelligent motors. Moreover, the industries have been able to generate improved logistical ways with the help of data science. Using different parameters including customer profile, position, financial pointers and logistics, sellers are able to optimize the transport routes and deliver a proper distribution of needy resources.

Moreover, several transport corporations like Uber is using data science to optimize price and provide improved practices to their clients. Applying strong prognostic tools, they are being able to correctly forecast the price built on factors like weather outline, transport obtainability, clients and so on.

3.5. Healthcare

In also healthcare domain, data science is doing its significant hikes. The different healthcare trades applying data science include

- Health bots or virtual assistants
- Predictive Modelling for Diagnosis
- Drug Discovery
- Genetics and Genomics
- Medical Image Analysis

3.6. E-Commerce

Data science has helped E-commerce and retail businesses to be enormously promoted in the following ways

- To detect a possible client base, data science is being hugely applied.
- To predict goods and services, use of forecasting analytics
- To identify patterns of common goods and predict their tendencies.
- To optimize various companies' pricing patterns for clients.

Serious application of data science is also there in collaborative filtering, where it acts as the mainstay of progressive recommendation system. Applying this method, the e-commerce podiums deliver perceptions to the clients depending on their past consumptions and acquisitions made by similar people. This hybrid recommendation system, containing collaborative and content-based filtering as well

help the businesses to deliver improved client services. In addition, businesses apply sentiment analysis for the analysis of client feedback. Natural language processing gets used here for analysing texts and reviews. Fraud Detection in trades is custom-made to find out scam dealers and deceptions in wire-transfers.

4. SENTIMENT CLASSIFICATION TECHNIQUES

Lexicon based sentiment classification technique and Machine learning are basically the two widely used techniques in this domain as described in Figure 2.

4.1. Machine Learning

Machine Learning (ML) is the training the machine such that it becomes capable of taking decision itself. Among the two types of ML, Training set gets used by supervised classifier for learning and training itself and performance of the classifier is tested using test dataset [8]. There are many kinds of classifier under supervised learning; most common among them are probabilistic classifier and linear classifier.

4.1.1. Probabilistic classifier

Probability classifier is a generative and combination model where every class is individual component. It determines the sampling likelihood for that module. Probabilistic classifiers are of three types – maximum entropy, Naive bayes and Bayesian network.

4.1.1.1. Naive Bayes classifier

It is an easy and popular classifier. Results derived using naïve bayes are usually good. It is based on Bayes theorem and uses Bag of word feature extraction technique. This type of classifier is suitable for text organization as it calculates the prior and later likelihood of a class, which is based on how the phrases in a document are distributed [8]. When the feature fits to a particular label, then Bayes equation is given

$$P(l | f) = \frac{P(l) * P(f | l)}{P(f)} \tag{1}$$

Where P(l) represents the previous likelihood of a label or a random feature set the label, P(f|l) denotes the previous probability of a assumed feature that is being classified as label and P(f) denotes previous likelihood that a given feature has occurred. In general, above equation could be rewritten as

$$P(f | lb) = \frac{d(w, e(f | lb))}{sl(d(w, e(f | l)), f)} \tag{2}$$

Kang and yoo [9] proposed an improvisation in Naïve bayes algorithm to increase the average accuracy. They successfully reduced the difference between the positive accuracy and negative accuracy using unigrams and bigrams as feature for increasing average accuracy. Pak and Paroubek [10] using naïve bayes algorithm for classifying the tweets as positive or negative proposed a model. Twitter corpus was created using Twitter API for collecting tweets containing emotions. POS-tags and N-gram feature extraction techniques was used. But the model turns out to be less efficient as training set considers

the tweets which contain emotions. Po-Wei liang and Bi-Ru Dai [11] designed a system called for polarity identification which integrates machine learning techniques and domain specific data of tweets collected using Twitter API. Unigram Naïve Bayes algorithm was used and Mutual Information and Chi square feature extraction techniques were used. But the proposed model did not give satisfactory accuracy.

Table 1: Description of the healthcare industries using data science

Types	Descriptions
Medical Image Analysis	Data science has influenced analysis of medical images like X-ray, MRI, CT-Scan etc. Earlier doctors and medical examiners used to manually search various clues in the medical images, whereas improvements in computing technologies and rise in the amount of data, have been able to generate machines which is self-sufficient to identify issues in the imagery. Strong image recognition techniques have been introduced for allowing doctors to have a detailed view of complicated medical imagery.
Genomic Data Science	Genomic Data Science sometimes uses some statistical methods to genomic sequences, which allows the bioinformatics specialists and geneticists to figure out the errors in genetic structures. It is also helpful in classifying diseases that are genetic in nature. Gene’s reaction to different types of medicines can be analysed using data science. Moreover, different big data tools like MapReduce have been able to significantly minimize the processing time of genome sequencing.
Drug Discovery	In drug discovery, medicine generation according to the requirement of new candidates is a tedious and most of the time, complicated process. Data Science helps in simplifying the technique and provides a view of the success rate of the new drug discoveries. With Machine Learning, we can also analyse several combinations of drugs and their effect on different gene structure to predict the outcome.
Predictive Modelling for Diagnosis	Improvements of predictive modelling techniques are enabling data scientists to predict the result of some disease where different existing data of some patients are available. Practitioners have been able to analyse data, generate correlations among the concerned data variables with the use of data science and can provide detailed view to doctors and medical persons.
Natural Language Processing	Natural Language Processing (NLP) is a part of machine learning which by integration with data science is concentrated to analyse basically textual data. Using NLP, intelligent bots which are able to answer different user queries can be created. The application area is extended to the healthcare section also where we can reply to patients queries and deliver proper diagnostic guidelines.

4.1.1.2. Bayesian network

Bayesian network is a type of probabilistic classifier that uses directed acyclic graph i.e. DAG to represent variables and

their conditional dependencies. Computation cost of Bayesian network [12] [13] in text mining is very high; therefore, they are not frequently used. Hernandez and Rodriguez [14] considered a practical problem, which determines the behaviour of the author using three different but connected target variables. They extend the Multi-dimensional classification framework to the semi-supervised domain so that unlabelled data can also be taken into account. This approach shows better performance than the other common approaches of sentimental analysis. In a Bayesian network, quoting directly from [Nilsson, 1998], each node of the graph are "conditionally independent of any subset of the nodes that are not descendants of itself given its parent". If it is denoted as: V representing a node in graph, $non(V)$ representing any collection of its non-descendant nodes, and $par(V)$ is denoting the collection of the direct parent of V , then $non(V)$ provisionally does not depend on V for $par(V)$, or

$$P(V | non(V), par(V)) = P(V | par(V)) \tag{3}$$

Therefore, the joint probability of all the nodes ($V_1, V_2 \dots V_n$) residing in the network can be denoted as

$$P(V_1, V_2, \dots, V_n) = \prod_{i=1}^n P(V_i | par(V_i)) \tag{4}$$

From the obtained joint probability, the essential probability can be extracted by using marginalisation. For measuring the joint probabilities, the conditional probability between itself and its parent should be calculated, for all the nodes in network. After that chain rule can be used to measure the joint probability functions of network.

4.1.1.3. Maximum Entropy classifier

Maximum Entropy [15] [16] also referred as MaxEnt is a probabilistic classifier which selects the model having highest entropy. Max Entropy assumes that the features are dependent to each other based on some condition. MaxEnt classifier is used to convert label feature set to vectors. Weight can be calculated for each feature using encoded vectors which can be shared to identify the most probable label aimed at a feature set.

Maximum Entropy Model

NLP: POS Tagging, LM, PP attachment, Text Classification, Parsing, ...

POS Tagging Features

$$f_i(C, t) = \begin{cases} \text{ifword}(C) = \text{Moody} \& t = 1 - \text{ORG} \\ 0 \text{ otherwise} \end{cases}$$

Model

$$p(t | C) = \frac{1}{Z(C)} \exp\left(\sum_{i=1}^n \lambda_i f_i(C, t)\right) \tag{5}$$

4.1.2. Linear classifiers

Linear classifiers are amongst the most practical classification methods. In sentiment analysis, linear classifier associates a coefficient with a count of each word in the sentence.

4.1.2.1. Support Vector Machine

One of the simplest linear classification approaches is Support Vector Machine. It is based on separating different

classes by finding a hyperplane, which can best maximize the margin among different classes. Li and li [17] designed a framework, which can identify the trendy topics and can classify the opinions. This approach is performed on tweets collected using twitter API. Research shows that user credibility and opinion subjectivity are considered essential for the aggregation of micro blog opinions. This framework effectively discovers market intelligence (MI) to support decision-makers.

A method was proposed by Chen and Tseng [18] to evaluate the quality of information extracted as something the reviews extracted are useless and even noisy. The evaluation of information quality is considered a classification problem and an effective information quality framework is used which extracted review features. This proposal used two multiclass SVM built methods: One-versus-All SVM and single-Machine multi-class SVM and improved than the state-of-the-art approaches. The SVM is implemented using different kernels. The most commonly used kernel used to implement the SVM is called the linear SVM kernel. The linear kernel SVM is the most usually applied text classifier because of its ease and less operational cost. The aim of it is finding out the largest margin for defining the hyperplane [19, 20]. For linear SVM, the kernel gets defined as the resemblance or a distance amount between a new data point and the linear combination of the support vectors [21]. Here we have described the SVM method briefly. Most of the time your data will be composed of n vectors x_i . Each x_i will also be associated with a value y_i indicating if the element belongs to the class (+1) or not (-1). Note that y_i can only have two possible values -1 or +1. Moreover, most of the time, our vector x_i ends up having many dimensions. We can say that x_i is a p -dimensional vector if it has p -dimensions. So, our dataset is the set of n couples of elements (x_i, y_i) . The more formal definition of an initial dataset in set theory is

$$D = \{(x_i, y_i) | x_i \in R^D, y_i \in \{-1, 1\}\}_{i=1}^n \tag{6}$$

4.1.2.2. Neural Network

Neural network is a group of interrelated neurons where neuron is the elementary cell unit. For non-linear boundaries multilayer neural network is used [22] where one layer's output is accepted by some succeeding layer but its training method is complicated as errors need to be back – propagated. Duncan and Zhang [23] used feed forward neural network to analyse tweet sentiments collected using twitter API. Memory becomes a constraint of feed forward training when the vocabulary becomes too huge.

A neuron is made-up of deciding the output and returns some other. This function is named as activation function denoted by $f(z)$, where z represents the combination of all the inputs. The basic types of activation include non-linear and linear. When $f(z)=z$, then $f(z)$ is called linear where nothing new is observed. The others are known as non-linear as the following

$$x_1 w_1 + x_2 w_2 + \dots + x_n w_n = f\left(b + \sum_{i=1}^n x_i w_i\right) \tag{7}$$

Where b is termed as bias, x represents the input of neuron, w denotes the weights, n is the quantity of inputs from the inward layer and i is a number that counts from 0 to n .

4.1.3. k-nearest neighbors' algorithm (k-NN)

It is a non-parametric technique used to classify and regress in machine learning. In both the scenarios, k number of nearest training features is the input [24] [25] and the output is based on the application area of k-NN, i.e. regression or classification [26]. In the classification problem, the class membership is the output. A feature is classified by votes by its neighbors, having the object allocated to the mostly voted class by its k closest neighbors where k is a positive small number. When $k = 1$, the object then is just allocated to the class of that particular nearest neighbor. In k-NN regression, the output is the mean value obtained from k nearest neighbors. k-NN is a kind of instance-based learning, or lazy learning as the function is approached just locally and every calculation is delayed before evaluating the function.

k-NN as a supervised learning algorithm example is a new instance classification based on the widely held K-nearest neighbor group. No model is used here for fitting because it just depends on memory. If all examples having m attributes are categorized to any of the two categories, positive or negative, provided a query instance x_q , k amount of training examples nearest to it, denoted by $N_k(x_q)$, are obtained. By

$N_k^+(x_q)$, and by $N_k^-(x_q)$, indicate the sets of positive and negative instances in $N_k(x_q)$ separately. If

$|N_k^+(x_q)| > |N_k^-(x_q)|$, x_q is labeled as positive, otherwise labelled as negative class. Temporarily, KNN is easily modified to estimated continuous-valued goal functions. For achieving the said thing, the estimation algorithm approximates the likelihood of an instance to be categorized in positive class properly by following the equation.

$$p(x_q) = \frac{|N_k^+(a_i(x))|}{|N_k^-(a_i(x))|} \tag{8}$$

4.1.4. Rule-based classification/machine learning (RBML)

It is integration of some machine learning technique which detects, learns, or changes 'rules' for storing, manipulating or applying [27] [28] [29]. The feature of a rule-based machine learner is the detection and using a set of relational rules which jointly signify the information taken by the scheme. This is in difference with other machine learners that commonly identify a singular model which is able to be generally applied to any instance to forecast. In the framework of a software quality categorization [30], the metric x_j for a given unit i , is either

$$(x_{ij} \geq L_{ji}) \cap (x_{ij} < U_{ji}) \text{ or } (x_{ij} \geq L_{j2}) \cap (x_{ij} < U_{j2}) \text{ or } \dots \text{ or } (x_{ij} \geq L_{jN}) \cap (x_{ij} < U_{jN})$$

where, L_{jk} and U_{jk} are the lower and upper limits of the j^{th}

characteristic in the k^{th} interval. If m software intricacy metrics are chosen as the most significant features for the model, then maximum I^m rules can be designed each rule having $2m$ terms, where I denote the number of intervals. For example, if $m = 2$, and $I = 2$ then maximum four possible rules can be created as listed in Table 1, in which a ' \cap ' denotes Boolean AND. Each instance in the fit data set can be classified by only one of the I^m rules. When a module satisfies a given rule, it implies that all the $2m$ individual Boolean terms of the rule are True. If we denote n_{fp} and n_{nfp} as the number of fp and nfp modules that satisfy a given rule, we can determine the probability of a rule as being fp or nfp . We can use the following equations to compute these probabilities:

$$p_{fp} = \frac{n_{fp}}{n_{nfp} + n_{fp}} \tag{9}$$

$$p_{nfp} = \frac{n_{nfp}}{n_{nfp} + n_{fp}} \tag{10}$$

While training the model, if a rule does not grasp any units, it gets removed from the collection of rules. Once the likelihood of all the rules are determined, p_{fp} and p_{nfp} , the rules get ranked in ascending order of p_{fp} . An advantage of organizing the rules is that an optimal threshold denoted by θ of p_{fp} can be obtained to distinguish the fp from the nfp rules. Each rule having $p_{fp} \geq \theta$ is categorized as fp rule, and rules having $p_{fp} < \theta$ are categorized in nfp rule. The classification guidelines to detect a module, x_i , the existing rule having features of the unit, belonging to either fp , or nfp is shown below.

$$Classs(x_i) = \begin{cases} fp & p_{fp}(rule^r(x_i)) \geq \theta \\ nfp & p_{fp}(rule^r(x_i)) < \theta \end{cases} \tag{11}$$

where $rule^r(x_i)$ implies that module i satisfies rule r ; $p_{fp}(rule^r(x_i))$ represents the probability of rule r , classifying the i^{th} module, being fp . The rule index r , varies from 1 to I^m . A specialist differs θ as per the requirement of the project.

4.1.5. Decision Tree in Machine Learning

A decision tree is a construction like flowchart where every inner node signifies a feature test (During tossing a coin, head comes up or tail), each leaf node signifies a class label (conclusion made by calculating every feature) and branches signify combinations of structures which results in different class labels. The routes between root and leaf signify classification guidelines. Decision tree is a prognostic modelling method applied in machine learning, data mining and statistics built by an algorithmic method which is able to detect the dataset splitting depending on several conditions. It is one of the most extensively used and practical methods for supervised learning. It is a non-parametric supervised learning method used as classifier and as well as regressor. Tree structures when the goal variable accepts a collection of

discrete values are known as classification trees. Decision trees taking a collection of continuous values as the goal variable (some real numbers) are named as regression trees.

Among several applications of decision trees, the most common one is C5.0 [31]. It has become the business standard to build decision trees as it performs good for most of the cases straight out of the box. Compared to more progressive and classy machine learning techniques like Neural Networks and Support Vector Machines, decision trees having C5.0 algorithm achieve same performance but are simpler for deployment and understanding. It measures the ailment in the set of attributes where effectivity and efficiency of that attribute using entropy and information gain get focused. The C5.0 activities on dataset can be denoted by the following

1. The entropy of the data is calculated using

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i \quad (12)$$

Where E(S) stands for the concerned entropy, *c* signifies the number of classes at the scheme and the number of instances proportional to class *i* gets denoted by *p_i*

2. The information gain for an attribute *C* can be measured by the following in a group *S*, having *E(S)* as the entropy of the entire system, *S_w* as the group of instances which provide value *w* for the feature *C*.

$$G(S, C) = E(S) - \sum_{w \in \text{values}(C)} \frac{S_w}{S} E(S_w) \quad (13)$$

4.2. Lexicon – based approach

The method is of three types: corpus-based approach, dictionary based and manual. Manual based approach takes much time and limited to some lexicon. It is prone to errors. To make it more efficient, it is integrated with two more automated approaches.

4.2.1. Dictionary based approach

A collection of opinion-based words is gathered manually having identified polarity. Then, this set rises using well defined thesaurus [33] or corpora WordNet [32] which finds similar and opposite words. The newly obtained phrase or word gets added to the list and the repetition proceeds again. This process continues until no new word is there. Taboada applied lexicon-based method for sentiment fetching from micro blogs while handling negations and intensifying words [34]. Dictionary based methods are actually categorized as unsupervised where it is assumed that positive (negative) adjectives are observed often close to a positive (negative) sense (Harb et al., 2008) [35]. An unsupervised learning procedure to classify feedbacks (like or dislike) has been used in [36]; Wang and Araki, 2007) [37].

A feedback classification can be calculated by the mean semantic alignment of the considered phrases containing adverbs or adjectives. The semantic alignment of a phrase is figured applying the joint information among the given phrase and the word excellent minus the mutual information between the given phrase and the word poor. Therefore, a

phrase is said to have a positive semantic alignment in case of having good relations and a negative semantic alignment in case of having bad relations, as following.

$$SO(\text{phrase}) = \log \frac{\text{hits}(\text{phrase NEAR excellent}) \cdot \text{hits}(\text{poor})}{\text{hits}(\text{phrase NEAR poor}) \cdot \text{hits}(\text{excellent})} \quad (14)$$

To learn the arguments used for expressing sentiments, positive and negative phrases (e.g. good, excellent, bad) are applied for extraction of adjectives close to the phrases.

4.2.2. Corpus based approach

Corpus based method is typically designed to induce domain specific semantic lexicon from a collection of domain specific text. When Contemporary dictionary approach is applied to specialized vocabularies, it exhibits some serious validity problems. Rice and Zorn [38] therefore developed a class of “minimally- supervised” so that sentiment dictionary can be created from corpus text. Its usefulness is shown in an application, which is used in U.S. federal appellate court decision.

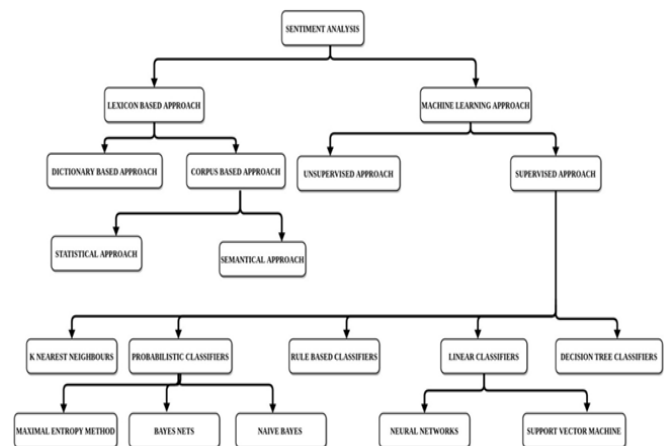


Figure 2: Sentiment Classification methods

4.2.2.1. Corpus-Based Statistics-Oriented (CBSO) approach

This kind of method [39] typically shows the following features. (1) It applies high level concepts long approved in traditional morphology, instead of surface strings, for modeling the stochastic language conduct, so that the parameters in the linguistic can be significantly condensed. (2) Parameterized statistical method is used to solve the uncertainties and ailments, where the linguistic-processing job can be accurately improved and the mandatory knowledge like the parameter values, can be learnt automatically and reliably. (3) It is generally more robust with contrast to the pure statistical methods as statistical optimization is used on higher layer of concepts; where statistical features are probable to be general for unobserved data improved than surface strings. The CBSO methods as hybrid methods, take benefit of rule-based as well as genuine statistical methods. A CBSO method creates statistical linguistic models on high-layer concepts like annotated syntax trees. A CBSO model for machine conversion has been introduced in [40] [41] [42] depending on high layer concepts like normal forms

and parse trees; the translation event is modeled like optimizing model that chooses the finest translation to maximize the translation score as follows

$$\begin{aligned}
 (1) & P(T_i | S_i) \cong \\
 & \sum_{I_i} \{ [P(T_i | PT_i(i)) \times P(PT_i(i) | NF1_t(i)) \times P(NF1_t(i) | NF2_t(i))] \quad (15) \\
 (2) & \times [P(NF2_t(i) | NF2_s(i))] \\
 (3) & \times [P(NF2_s(i) | NF1_s(i)) \times [P(NF1_s(i) | PT_s(i)) \times P(PT_s(i) | S_i)]]
 \end{aligned}$$

where (S_i, T_i) denotes the i^{th} source-target conversion pair (PT_s, PT_t) are the parse trees for the source-target sentences, $NF1_s$ and $NF2_s$ signify syntactically and semantically normalized parse trees, known as normal forms of the original sentence, normal forms of the target sentence are denoted by $NF1_t$ and $NF2_t$, and the summation of the likelihoods is taken over all such in-between depictions, I_i . The three equations (1), (2) and (3) describe the creation, transmission and investigation models of a transfer-based MT system following CBSO manner; which is able to be simplified during implementation in future.

4.2.2.2. Semantical approach

Corpus-based semantic methods approve the common idea of application of alike words in alike contexts and signify the word/concept as a sequential list of standards which precise in a way, the background of the word, which is the collection of contexts where the word appears. In this context, some basic findings have been presented. The most usual estimation can be mentioned when some distance metric on some semantic vector space gets described, there the vectors consistent with semantically unconnected words should be further separate than those for connected words should. As there exist several normalization artefacts which may arise during comparing vectors resulted by several values of the above factors like various quantities of baseline noise, the natural dimension free amount to relate is the connectivity ratio

$$R = \frac{\text{Mean distance between control words}}{\text{Mean distance between related words}} \quad (16)$$

The greater the ratio, the comparatively closer are the linked words, and improved the semantic symbols. Illustrative collection of 100 pairs of words which were refereed by human subjects as close substitutes [43] and for every pair eight frequency matching arbitrary couples of words for controlling the whole system. The co-occurring vectors were built eventually and ratio denoted by R using simple Euclidean distances for these couples in a series of parameters.

5. WORLDWIDE STOCK MARKET RECORD DOWNFALL (COUNTRY AND CAUSE (PANDEMIC OR TERRORIST ATTACK WISE) WISE)

Capitalization of Market (called as market price) is the share price times the number of shares remaining (including numerous classes) for registered national corporations. Companies, unit trusts, and investment funds, where only

commercial aim is to hold shares of other listed companies are omitted. End of year values are the considered data. Stock market capitalization, billion USD, 2020 - Country rankings: The mean for 2020 depending on 65 countries was 1040.04 billion U.S. dollars. The maximum value was observed at USA: 30436.31 billion U.S. dollars and the lowest was at Algeria: 0.37 billion U.S. dollars. The pointer is available from 1975 to 2020 as shown in Figure 3.

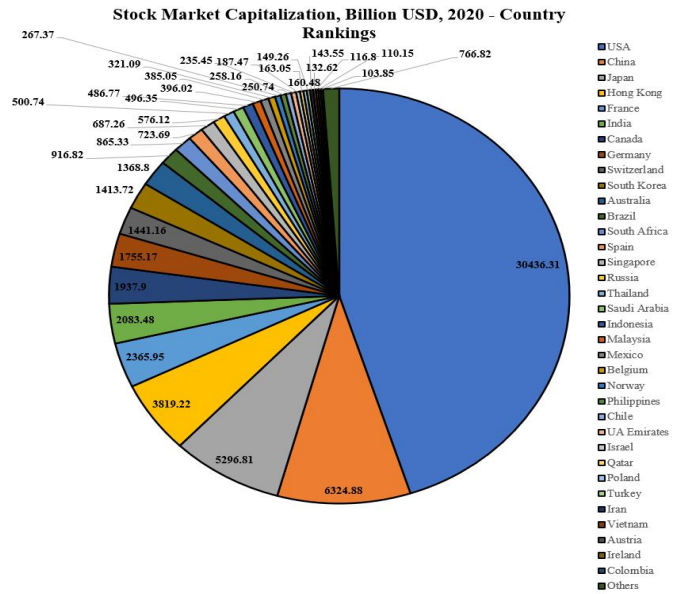


Figure 3: Stock Market Capitalization, Billion USD, 2020 - Country Rankings

6. DATA COLLECTION

The quality and amount of data has a great impact on the concerned work and Figure 4 shows the various sources of it.

6.1 Data collection process

There are four popular ways of collecting data:

6.1.1 Internal Data

Companies collect some exclusive data by its actions or through collaborations with other workers which is generally the most applicable data.

6.1.2. Searching Online

If we require a labeled set of a huge number of videos, a webpage is there to provide that. The collection is really surprising to everyone. Available datasets permit us in prototyping prior investment in exclusive data.

6.1.3. API's

API's permit us for accessing programmatic datasets gathered by other businesses. Anything can be obtained from Twitter feeds of climate change to monetary data.

6.1.4. Web Scrapping

Web scrapping and crawling is an influential means that must be used sensibly. A total new world opens up which ensures to obey terms of services. The S&P BSE Sensex missed another 581 points or 2% to 28,288 on 19.03.20 Thursday, the lowest

closing level since February 2017, after a day of volatile trade. Fears of a global recession mounted despite massive stimulus measures from central banks and governments around the world. The broader Nifty50 index slid 205 points or 2.4% to end at 8,263. India SENSEX Stock Market Index - data, predictions, historical chart was last modified on March of 2020.










Typical Data Source	Internal Data Source	Semi/Unstructured Data Source
 Market Research & Competitive Information	 Viewership	 Clickstream
 Audience Demographics	 Advertisements	 Social Media
	 Merchandise	 Content
	 Sales & Finance	

Figure 4: Various Types of Data sources

6.2. Types of data

Programming gets applied by data science for analysis of data, which is of several kinds. The essential types of data are as follows.

6.2.1. Structured Data

The data which is simple to be presented in a table, where storing and manipulating in databases and Excel files are called structured in nature. Suppose, Airbnb consists of a database of available apartments, involving variables like, size of apartment (in square feet), number of beds, number of washrooms, number of visitors it is able to accommodate, per day rent etc.

6.2.2. Unstructured Data

Data unable to simply fit into some model, is known as unstructured data. Examples can be videos, images, PDF documents, emails and so on.

6.2.3. Natural language

Data presented in the form of human languages for communicating with others like English, Chinese, French, etc. are natural languages which is a kind of unstructured data.

6.2.4. Audio, Video, Image

Audio, video and image are obtained from sensor devices like microphone and camera which are generally unstructured and extraction of information from those is a challenging task.

6.2.5. Graph-based Data

An analytical structure modeled pairwise based on the relation among two entities. Properties, edges and nodes are used here to stock data. As an example, our Facebook friend

detail can be characterized with graph, having people representation by nodes, edges connecting two nodes signify that two nodes of people are friends with each other.

6.2.6. Machine Generated

It is any information obtained from a device, several applications without interacting with humans.

7. DATA PROCESSING COMPONENTS

The components used for data processing are described as follows.

7.1. Types of Sentimental Analysis

Here the four different types of sentiment analysis are described below.

7.1.1. Fine-grained sentiment

This analysis gives you an understanding of the feedback you get from customers. You can get precise results in terms of the polarity of the input. However, the process to understand this can be more labor and cost-intensive as compared to other types.

7.1.2. Emotion Detection Sentiment Analysis

This is a more sophisticated way of identifying the emotion in a piece of text. Lexicons and machine learning are applied to determine the feeling. Lexicons are lists of words that are either positive or negative. This makes it easier to segregate the terms according to their sentiment. The advantage of using this is that a company can also understand why a customer feels a particular way. This is more algorithm-based and might be complex to understand at first.

7.1.3. Aspect-based

This type of sentiment analysis is usually for one aspect of a service or product. For example, if a company that sells televisions uses this type of sentiment analysis, it could be for one aspect of televisions – like brightness, sound, etc. So, they can understand how customers feel about specific attributes of the product.

7.1.4. Intent analysis

This is a deeper understanding of the intention of the customer. For example, a company can predict if a customer intends to use the product or not. This means that the intention of a particular customer can be tracked, forming a pattern, and then used for marketing and advertising.

7.2. Sentiment Analysis Component

In this component, we have discussed about the basic component of text processing as below.

7.2.1. Tokenization

Each news headlines or tweets or comments or monetary report is divided into expressive words named as tokens.

7.2.2. Data standardization

In this method for data constancy, every word in the articles and information regarding companies is altered in a

document into lower case. Stop-word-removal: Phrases that do not have important meaning in a sentence like the, a, of...etc. are deleted to decrease the feature number and progress the performance. Stemming: Porter stemmer gets used sometimes on the dataset for returning every word to the corresponding stem and eliminates the suffix like -Ed, -ing, -ion...etc. for reducing the document intricacy and minimizing the operational time for improving the system performance.

8. RESULT ANALYSIS OF EXISTING SENTIMENT ANALYSIS METHODS

Here we have compared various algorithms with their objective methodology used algorithm used performance and also their future scope/drawback in Table 2.

Table 2: Comparative analysis of the existing sentiment analysis methods

Ref.	Objective	Methodology	Algorithm	Performance	Future Scope/Drawback
[44]	Automatic gathering of client feedbacks for product/service & analyze the sentiments expressed about specific features	-Sentiment analysis across Flipkart E-commerce websites for filtering of irrelevant reviews -MongoDB at backend	Corpus of reviews and inference	Total sentimental score of a component, scaled on a range of -100 to 100.	Temporal analysis of reviews and enabling manufacturers to look at the sentiments as a function of time to judge the improvements or deteriorations over time
[45]	To find out the relation between university's academic success and sentiment about university in the	Sentiment analysis between sentiments of people on social network & academic success of Turkish	Naive Bayes classifier & Time Frequency and Inverse Document Frequency	Success rate of the system is 72.33%.	Applying quantitative analysis with larger sample to relate between academic success of universities and sentiment

	social media	universities			about them
[46]	-Identify opinion in the companies' official twitter account's tweet -Classify into positive, negative & neutral sentiment -Visualize their communication network.	LingPipe Library to classify the sentiment of users' opinion into positive, negative and neutral classes.	Alchemy API & NLTK applications to label 2/3rd of the data -1/3rd of the data by Naïve Bayes classification	NodeXL for visualizing the result of user's opinion	To examine & visualize users' opinion on the network level for determining sub-social systems & consideration of user emoticons & location information to understand behavior
[47]	Public opinion analysis system consisting of a crawler to retrieve online microblog content and a text classifier	-Spider to extract the microblog from the "weibo.com" according to specific time and topic	SVM to separate sentimental content to detect public opinions on certain topics	Precision of classification exceeded 90% using support vector machine	Classification result only forecast the public opinion tendency on particular topic real-time is done.
[48]	Knowledge extraction from tweets & classification based on semantics of knowledge	-Knowledge generator to classify tweets into category -Knowledge enhancer with	Domain specific classification and sentiment analysis is to extract	Overall significant improvement from 0.1% to 55%	Personalized profile management, sentiment analysis, and recommended system.

	ge	synonym binder to increase information gain	t and archive tweet		
[49]	Turkish tweet sentiment analysis on insurance, sport, finance, food, automotive, politics, real estate, Telecommunication & health.	Classification of tweets into positive and negative emotions	Collection of Turkish tweets by using Naïve Bayes, Support Vector Machines and Random Forest.	Support vector machine gives best result compared to the other classifiers.	Only traditional algorithm is used to predict the accuracy.
[50]	Sentiment analysis on twitter data for predicting movie success rate	Counting-based classifier with ML classifier	Lingpipe sentiment analyzer to test the sentiments	Prediction accuracy is 64.4 % better than traditional systems	Use of only leverage polarity shifting detection
[51]	Predicting the movie performance based on social networking sites data using sentiment analysis	Analyzing personalized sentiments via online documents, which is an important topic of opinion mining.	Prediction of movie by K-means clustering algorithm	Proposals for decision makers when crucial event happens	Data is restricted, measuring social emotions based on big data of microblog can be beneficial

9. CASE STUDY ANALYSIS FOR PROPOSED WORK

The case study for our proposed work is analyzed as follows

9.1. Description of the datasets

The experiments were performed with two data sets on two different case study: The first dataset contains news articles published on New York Times and Dow Jones stock market report in table format collected from Yahoo Finance, both gathered between September 2001 and December 2001. In the second data set, it contains news articles published at New York Times and Dow Jones stock market report in table format collected from Yahoo Finance, both gathered between January 2020 to May 2020.

9.2. Tools

The tools, which have been used, are

1) Natural Language Toolkit (NLTK) is an open-source Python package providing numerous tools to build programs and classify data. It is appropriate for developers working on textual data and text analysis [52], researchers, educators, students, engineers and linguists. It delivers a simple way for using the interfaces of over 50 lexical resources and corpora which involves a collection of text processing libraries to classify, semantic reasoning, tokenize, stem, tag, and parse [53].

2) Valence Aware Dictionary and sEntiment Reasoner (VADER) is a rule and lexicon-based sentiment analysis tool which is exactly agreed to the opinions uttered in social media. It is a free open-source which considers word sequence and degree convertors [54] [55].

9.3. Pre-processing

All data set instances were pronunciations and punctuation mark. For different machine learning classifiers, numerous pre-processing methods have been used

- Lemma extraction: every verb is derived to an infinitive form where nouns and adjectives to a singular procedure. VADER can be used in this context [56].
- Stemming reduces all the words to their radicals using NLTK tools.
- Part of Speech (PoS) Tagging: the NLPNET Python library [57] [58], proposing a PoS tagger for organized texts can be used.
- Summarization: only the primary three lines and the news heading can be considered in the classification process. This is reliable with extracting summarization methods used in Web.

9.4. Flowchart of our Sentiment Analysis Model

In this section, we describe our sentiment analysis model as shown in Figure 5 as a flowchart.

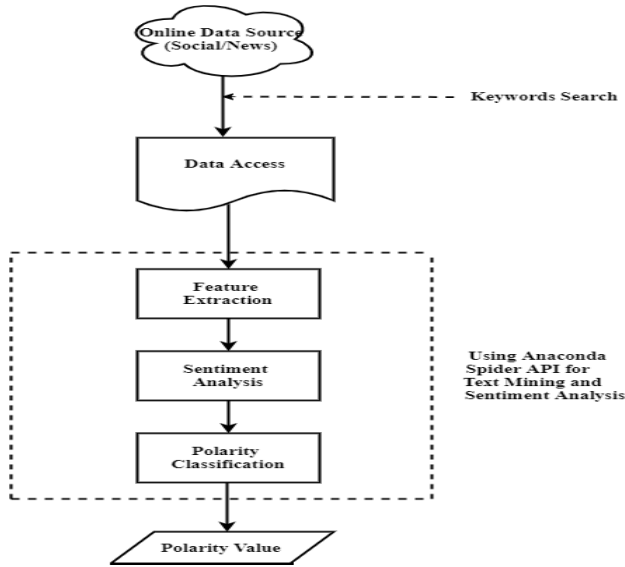


Figure 5: Algorithm Flowchart

Here in Algorithm 1, we have discussed the steps to carry on the analysis process on the news articles so that we can categorize the data about its positivity or negativity or naturalness.

ALGORITHM 1: Sentiment Classification using VADER

Input: Text File (News Articles/Social Data which include Nouns, Adjectives, Adverbs)

Output: Values > 0 (Positive), Values < 0 (Negative), Values = 0 (Neutral)

Begin:

```

1. Sentiment Analysis () ← File
2. For each row in rows
3.   if Sentiment Polarity Score(line) >= 0.05 then
4.     Sentiment ← Positive
5.   else
6.     if Sentiment Polarity Score (line) <= - 0.05 then
7.       Sentiment ← Negative
8.     else Sentiment ← Neutral
9.   end
10. end
11. end
12. end
13. end
    
```

9.5. Result Analysis

To conduct analyses of results, we proceed by carrying out computation based on data collected from news articles and stock market report. Article data are processed to identify positive or negative sentiments. On the contrary, we have considered an entire stock exchange index for the purposes of our experiments. In case of negative sentiments, it predicts that downward of stock value of shares. We have also compared the value rise-fall of gold price and crude oil to prove our guess more concrete. Here in Table 3 the polarity with their sentiment score and percentage are describe which are collected on the data set 1 by VEDAR sentiment analysis.

Table 3: Polarity value and Percentage of Polarity of Dataset1

Polarity	Sentiment score	Percentage
Positive	0.08	8.0
Neutral	0.731	73.1
Negative	0.189	18.9

Here in Table 4 the polarity with their sentiment score and percentage are describe which are collected on the data set 2 by VEDAR sentiment analysis.

Table 4: Polarity value and Percentage of Polarity of Dataset2

Polarity	Sentiment score	Percentage
Positive	0.026	2.6
Neutral	0.824	82.4
Negative	0.15	15.0

Here in Figure 6 the polarity with their percentage value are described which we get from Table 3. Here in Figure 7 the stock values are represented based on data set 1 of DOW JONES stock market, USA.

Sentiment Analysis on News article on 9/11 Terrorist Attack on USA

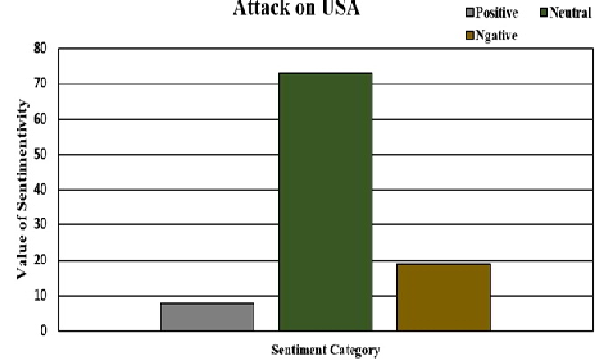


Figure 6: Sentiment Analysis on News articles on 9/11 Terrorist Attack on USA

Stock Market Close value of DOW JONES From Jan.2020 to May 2020

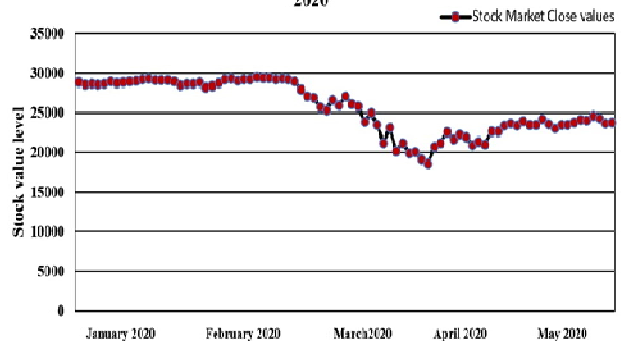


Figure 7: Stock Market Close Value of DOW JONES from September 2001 to December 2001

If we compare the above two figures (6,7) then we can conclude that most of the news articles in USA after the terrorist attack on 11th September,2001(9/11 terrorist attack) have negative polarity than positive polarity that also affect

the stock market value and can see the downward movement. Here in Figure 8 the polarity with their percentage value are described which we get from Table 4. Figure 9 represents stock values based on data set 2 of DOW JONES stock market, USA.

Figure 8: Sentiment Analysis on News articles on Covid19 in USA

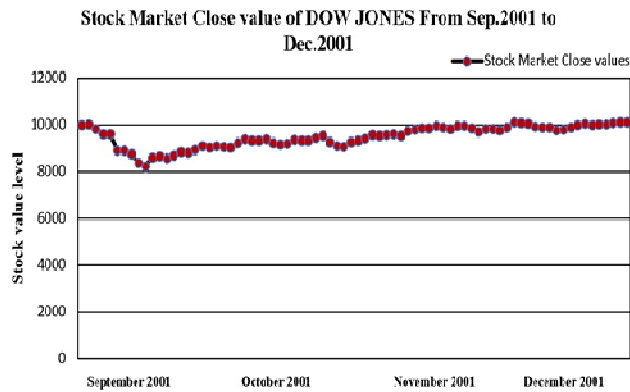


Figure 9: Stock Market Close Value Trend of DOW JONES from January 2020 to May 2020

By comparing Figure 8 Figure 9 we can conclude that most of the news articles in USA after the COVID 19 disease spread up after February, 2020 have negative polarity than positive polarity that also affect the stock market value and can see the downward movement with massive spread up of disease that cause economic downfall, layoff jobs and last of life casualty. Here in Figure 10 the crude oil price trend with up and down are represented on USD from the year 2000 to 2020.

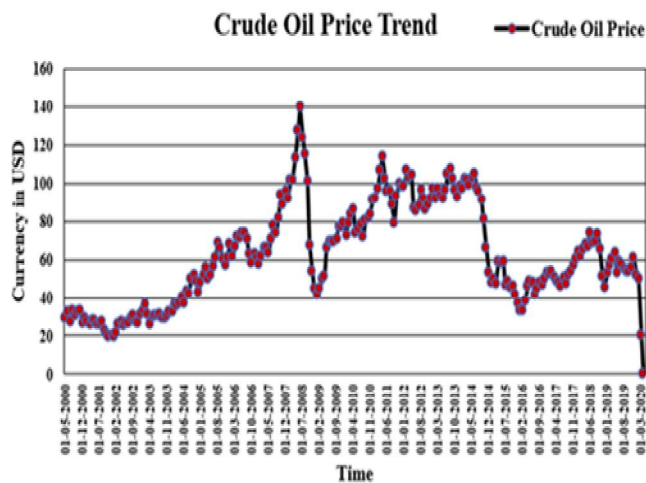


Figure 10: Crude Oil Price Trend in USD from 2000 to May 2020

Here in Figure 11 the Gold price trend with up and down are represented on USD from the year 2000 to 2020. There is strong relation between the up down of stock market with the up and down of price of Crude oil and Gold price [59] [60] [61] [62] [63] [64] [65] [66] [67]. From Figure 10 we can say that demand of crude oil in USA is so low that it almost has gone to 0 \$ per barrel after Covid 19 spread up in USA.

In opposite to Gold price from Figure 11 we can see that demand of Gold price has reach to a top position after Covid 19 spread up in USA. The news article which are published in the month of February, March 2020 which contains most news of economic related have most of negative polarity that positive polarity which inspire people to most invest in Gold rather than Company Share, Mutual Fund and Crude oil.



Figure 11: Gold Price Trend in USD 2000 to May 2020

10. CONCLUSION AND FUTURE SCOPE

In our exploratory study, investigation on the concurrent result of exploring several kinds of news with past numeric values for understanding stock market trend. The proposed model has worked on the views/ opinions of the reviewers on the shares. The views of the experts affect to the traders who want to invest into the market. Our proposed model has enhanced the forecast accuracy on the upcoming trend of stock market, by analyzing several types of everyday news with several values of numeric characteristics in a time domain. Further, research work can be applied in sequence to get more improved results in this domain. In future, the result of many other feature selection methods can be explored. Additional researches can be performed for evaluating the effect of numerous areas and domain-based factors. Enhancing the application of sentiment analysis to extended areas may result in interesting facts. In future, more mixture of n-grams and feature allowance giving a improved accuracy than the current one can be used. This research is only linked to sentiment categorization into two basic classes called binary classification. Sentiment can belong to either a positive emotion or a negative emotion. For improvement in future, more than one class for sentiment categorization can be considered like neutral, negative, positive etc. In this paper, the emphasis is on discovering features which appear clearly in the form of nouns or nominal phrases in the evaluations. The implicit feature study is kept for future work. Since the ensemble learning approaches require much computation time, the parallel computing methods can be explored to handle this issue. A main constraint of ensemble learning approaches is the absence of result depiction and the information gathered by ensembles is tough for human understanding. So, improvement of ensemble interpretability can be another vital research track. Future opinion analysis structures need wider and in depth general knowledge base

which will result in a improved consideration of natural language opinions and will be more effective to link the gap among multimodal and machine sensible data. Combination of scientific theories regarding emotion with the applied technology targets to analyze sentiments of natural language script will result in additional bio-inspired methods to design intelligent opinion analysis structures able to handle semantic information, analogy creation, learning actual information, and perceiving, identifying, and sensing emotions.

During design and development of our proposed model, we will be discussed and analyze the performance of four prominent machine learning algorithms. We have explored, how these algorithms are effective in predicting stock market trends based on user reviews and comments. We have also seen that there is a vast scope of sentiment analysis in predicting the stock market value which will reduce the risk of the share investors.

Confusion Matrix can also be used to identify classifier performance based on sample data. And also show how our proposed model can improve prediction accuracy. We will also investigate the effects of pandemic using live examples of the effects caused by corona virus infections around the world. We will analyze several types of news with numerical values to enable understand trends prevalent in stock markets based on historical patterns. In our proposed model, we will consider views and opinions posted by reviewers on shares. The views shared by reviewers may influence traders investing in stock market and also, we will prove that our proposed model can predict market trends as there is a firm co-relation between news on pandemic and spread of infectious disease with that of movement of share prices in stock markets. In future, research work can be expanded to include individual share prices and combine daily news and posts gathered from multiple financial and stock market web sites.

REFERENCES

1. Michal Skuza, Andrzej Romanowski, "Sentiment Analysis of Twitter Data within Big Data Distributed Environment for Stock Prediction", Computer Science and Information Systems pp. 1349– 1354, 2015 F230 ACSIS, Vol.5.
<https://doi.org/10.15439/2015F230>
2. Venkata Sasank Pagolu, Kamal Nayan Reddy Challa, Ganapati Panda, Babita Majhi, "Sentiment Analysis of Twitter Data for Predicting Stock Market Movements", International conference on Signal Processing, Communication, Power and Embedded System (SCOPEs), 2016.
3. Tina Ding, Vanessa Fang, Daniel Zuo, "Stock Market Prediction based on Time Series Data and Market Sentiment", 2012.
4. Phillip Tichaona Sumbureru, "Analysis of Tweets for Prediction of Indian Stock Markets", International Journal of Science and Research (IJSR), Volume 4 Issue 8, August 2015.
5. Rishabh Soni, K. James Mathai, "Improved Twitter Sentiment Prediction through „Cluster-then-Predict Model"", International Journal of Computer Science and Network, Volume 4, Issue 4, August 2015.
6. Linhao Zhang, "Sentiment Analysis on Twitter with Stock Price and Significant Keyword Correlation", April 16, 2013.
7. Manoj Kumar Danthala, "Tweet Analysis: Twitter Data processing Using Apache Hadoop", International Journal of Core Engineering & Management (IJCEM), Volume 1, Issue 11, February 2015.
8. Anuja P Jain and Padma Dandannavar 2016 .“Application of machine learning techniques to sentiment analysis”
9. kang H., Yoo S.J. 2011 . Senti-Lexicon and improved Naïve Bayes algorithms for sentiment analysis of restaurant reviews.
10. Pak, A., & Paroubek, P. 2010 . Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010).
11. Liang, P. W., & Dai, B. R. 2013 . Opinion mining on social media data. In Mobile Data Management (MDM), 2013 IEEE 14th International Conference on (Vol. 2, pp. 91-96). IEEE.
<https://doi.org/10.1109/MDM.2013.73>
12. D. Heckerman, “A Tutorial on Learning with Bayesian Networks,” in Learning in Graphical Models, Dordrecht: Springer Netherlands, 1998, pp. 301–354.
13. Hageman, R. S., Leduc, M. S., Korstanje, R., Paigen, B., Churchill, G. A. (2011). A Bayesian framework for inference of the genotype-phenotype map for segregating populations. *Genetics* 187 (4), 1163–1170. doi: 10.1534/genetics.110.123273
14. Ortigosa-Hernández, J., Rodríguez, J. D., Alzate, L., Lucania, M., Inza, I., & Lozano, J. A. 2012 . Approaching Sentiment Analysis by using semi-supervised learning of multi-dimensional classifiers. *Neurocomputing*, 92, 98115.
<https://doi.org/10.1016/j.neucom.2012.01.030>
15. Batista, F.; Ribeiro, R. “Sentiment analysis and topic classification based on binary maximum entropy classifiers”. *Proces. Del Leng. Nat.* 2013, 50, 77–84.
16. Phillips, S.J.; Anderson, R.P.; Schapire, R.E. Maximum entropy modeling of species geographic distributions. *Ecol. Model.* 2006, 190, 231–259.
17. Li, Y. M., & Li, T. Y. 2013. Deriving market intelligence from microblogs. *Decision Support Systems*, 55(1), 206217.
18. Chen, C. C., & Tseng, Y. D. 2011. Quality evaluation of product reviews using an information quality framework. *Decision Support Systems*, 50(4), 755-768.
19. T. Joachims, Text categorization with Support Vector Machines: Learning with many relevant features, *Mach. Learn. ECML-98*, 137–142.
20. J. Zhang, Y. Yang, Robustness of regularized linear classification methods in text categorization, in: Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in

- Informaion Retrieval, Toronto, Canada, 2003, pp. 190–197
21. J. Brownlee, *Master Machine Learning Algorithms: Discover how They Work and Implement Them from Scratch*, 2016.
 22. Goldberg, Y. “Neural network methods for natural language processing”. *Synth.Lectures Hum. Lang. Technol.*10,1–309 (2017).
 23. Duncan, B., & Zhang, Y. 2015 . Neural networks for sentiment analysis on twitter. In *Cognitive Informatics & Cognitive Computing (ICCI* CC)*, 2015 IEEE 14th International Conference on (pp. 275-278). IEEE. <https://doi.org/10.1109/ICCI-CC.2015.7259397>
 24. Shah, K., Patel, H., Sanghvi, D., Shah, M., 2020. A Comparative Analysis of Logistic Regression, Random Forest and KNN Models for the Text Classification. *Augment Hum Res* 5, 12 (2020).
 25. Bounabi Mariem, Karim El Moutaouakil Khalid Satori .“Text classification using Fuzzy TF-IDF and Machine Learning Models”. October 2019 Conference: BDIoT'19: The 4th International Conference On Big Data and Internet of Things.
 26. Altman, Naomi S. (1992). "An introduction to kernel and nearest-neighbor nonparametric regression" (PDF). *The American Statistician*. 46 (3): 175–185
 27. Bassel, George W.; Glaab, Enrico; Marquez, Julietta; Holdsworth, Michael J.; Bacardit, Jaume (2011-09-01). "Functional Network Construction in Arabidopsis Using Rule-Based Machine Learning on Large-Scale Data Sets". *The Plant Cell*. 23 (9): 3101–3116. doi:10.1105/tpc.111.088153. ISSN 1532-298X. PMC 3203449. PMID 21896882
 28. M., Weiss, S.; N., Indurkha (1995-01-01). "Rule-based Machine Learning Methods for Functional Prediction". *Journal of Artificial Intelligence Research*. 3 (1995): 383–403. arXiv:cs/9512107. doi:10.1613/jair.199.
 29. "GECCO 2016 | Tutorials". *GECCO 2016*. Retrieved 2016-10-14.
 30. Khoshgoftaar T.M, Bullard, L.A and Gao, K “A rule-based software quality classification model ”, *International Journal of Reliability, Quality and Safety Engineering*, vol. 15, no. 03, pp. 247-259 (2008). <https://doi.org/10.1142/S0218539308003064>
 31. I. S. Amiri, O. A. Akanbi, and E. Fazeldhkordi, “A Machine-learning Approach to Phishing Detection and Defense”. *Syngress*, 2014.
 32. Jatinder kaur 2016. Review Paper on Twitter Sentiment Analysis Techniques. In *International Journal for Research in Applied Science & Engineering Technology (IJRASET)* Volume 4 Issue X, October 2016 ISSN: 2321-965.
 33. Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. 1990. Introduction to WordNet: An on-line lexical database. *International journal of lexicography*, 3(4), 235-244.
 34. Mohammad S, Dunne C, Dorr B. 2009 Generating high-coverage semantic orientation lexicons from overly marked words and a thesaurus. In: *Proceedings of the conference on Empirical Methods in Natural Language Processing (EMNLP'09)*.
 35. Ali Harb, Michel Plantie, Gerard Dray, Mathieu Roche, Francois Troussel, and Pascal Poncelet. 2008. Web opinion mining: How to extract opinions from blogs? In *Proceedings of the 5th International Conference on Soft Computing As Transdisciplinary Science and Technology, CSTST 08*, pages 211–217, New York, NY, USA.ACM. <https://doi.org/10.1145/1456223.1456269>
 36. Peter D. Turney. 2002. “Thumbs up or thumbs down?:Semantic orientation applied to unsupervised classification of reviews”. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, ACL'02*, pages417–424, Stroudsburg, PA, USA. Association for Computational Linguistics
 37. Guangwei Wang and Kenji Araki. “Modifying SO-PMI for Japanese weblog opinion mining by using a balancing factor and detecting neutral expressions”. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Companion Volume, Short Papers, NAACL-Short'07*, pages 189–192, Stroudsburg, PA, USA. Association for Computational Linguistics.
 38. Rice, D. R., & Zorn, C. 2013. Corpus-based dictionaries for sentiment analysis of specialized vocabularies. *Proceedings of NDATAD*, 98-115.
 39. Su, K.Y., Chiang, T.H., Chang, J.S.: An Overview of Corpus-Based Statistics Oriented (CBSO) Techniques for Natural Language Processing. *Computational Linguistics and Chinese Language Processing* 1(1), 101–157 (1996)
 40. Su, K.-Y. and J.-S. Chang, "Some Key Issues in Designing MT Systems," *Machine Translation*, Vol. 5, No. 4, 1990, pp. 265-300.
 41. Su, Keh-Yih and Jing-Shin Chang, "Why Corpus-Based Statistics-Oriented Machine Translation," *Proceedings of TMI-92*, pp. 249-262, *Proceedings of 4th Int. Conf. on Theoretical and Methodological Issue in Machine Translation*, Montreal, Canada, June 25-27, 1992.
 42. Su, K.-Y, J.-S. Chang, and Yu-Ling Una Hsu, "A Corpus-based Two-Way Design for Parameterized MT Systems: Rationale, Architecture and Training Issues," *Proceedings of the Sixth International Conference on Theoretical and Methodological Issues in Machine Translation, TMI-95*, vol. 2, pp. 334-353, *Sixth Int. Conf. on Theoretical and Methodological Issues in Machine Translation*, Leuven, Belgium, July 5-7, 1995.
 43. Moss HE, Ostrin RK, Tyler LK & Marslen-Wilson WD. Accessing Different Types of Lexical Semantic Information: Evidence From Priming. *Journal of Experimental Psychology: Learning, Memory and Cognition* 1995; 21:863-883 <https://doi.org/10.1037/0278-7393.21.4.863>

44. Singh, P. K., Sachdeva, A., Mahajan, D., Pande, N., & Sharma, A, An approach towards feature specific opinion mining and sentimental analysis across e-commerce websites, IEEE International Conference on the next Generation Information Technology, 2014, pp.329-335.
45. Gunduz, S., Demirhan, F., & Sagiroglu, S, Investigating sentimental relation between social media presence and academic success of Turkish universities, IEEE 13th International conference on Machine Learning and Applications, 2014, pp.574-579.
46. Molla, A., Biadgie, Y., & Sohn, K. A. Network based visualization of opinion mining and sentiment analysis on twitter, IEEE International conference on It Convergence and Security, 2014, pp.1-4.
47. Lu, Y., & Chen, J, Public opinion analysis of microblog content, IEEE International conference on Information Science and Applications, 2014, pp.1-5.
48. Batool, R., Khattak, A. M., Maqbool, J., & Lee, S, Precise tweet classification and sentiment analysis, IEEE 12th International Conference on Computer and Information Science, 2013, pp.461-466.
49. Meral, M., & Dirir, B “Sentiment analysis on Twitter, IEEE 22nd International Conference on signal Processing and Communications Applications Conference, 2014, pp. 690-693.
50. Li, S., Wang, Z., Lee, S. Y. M., & Huang, C. R, Sentiment Classification with Polarity Shifting Detection, IEEE International Conference on Asian Language Processing, 2013, pp.129-132.
<https://doi.org/10.1109/IALP.2013.44>
51. Wang, X., & Luo, X, Sentimental Space Based Analysis of User Personalized Sentiments, IEEE 9th International Conference on Semantics, Knowledge and Grids, October 2013, pp. 151-156.
52. Natural Language Toolkit <http://www.nltk.org/> (Date Last Accessed, November 20, 2018).
53. Bird, Steven, Edward Loper and Ewan Klein (2009), Natural Language Processing with Python. O’Reilly Media Inc.
54. C. J. H. E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) <http://comp.social.gatech.edu/papers/icwsm14.vader.hutto.pdf>, 2014.
55. Elbagir, S., and Yang, J. 2019. Twitter Sentiment Analysis Using Natural Language Toolkit and VADER Sentiment. In Lecture Notes in Engineering and Computer Science: Proceedings of The International MultiConference of Engineers and Computer Scientists. Hong Kong: The International Association of Engineers.
56. Natural Language Processing with neural networks <http://nilc.icmc.usp.br/nlpnet/>
57. Fonseca, E. R. and Rosa, J.L.G. A Two-Step Convolutional Neural Network Approach for Semantic Role Labeling. Proceedings of the 2013 International Joint Conference on Neural Networks, 2013. p. 2955-2961.
<https://doi.org/10.1109/IJCNN.2013.6707118>
58. Fonseca, E. R. and Rosa, J.L.G. Mac-Morpho Revisited: Towards Robust Part-of-Speech Tagging. Proceedings of the 9th Brazilian Symposium in Information and Human Language Technology, 2013. p. 98-107.
59. Simakova, J., 2011. “Analysis of the relationship between oil and gold prices”. Journal of finance, 51(1), 651-662.
60. Smith, G. (2001). “The Price of Gold and Stock Price Indices for The United States”. The World Gold Council. Vol. 8(1): 1-16.
61. Baur, D. G., & Lucey, B. M. (2010). “Is Gold a Hedge or a Safe Haven? An Analysis of Stocks, Bonds and Gold”. Financial Review. Vol. 45(2): 217-229. doi: <https://doi.org/10.1111/j.1540.6288.2010.00244.x>.
62. A. Jain, P.C. Biswal “Dynamic linkages among oil price, gold price, exchange rate, and stock market in India” Resource.Policy, 49 (2016),pp. 179-185, 10.1016/j.resourpol.2016.06.001.
63. S. Biswas, I. Sarkar, P. Das, R. Bose, S. Roy, “Examining the Effects of Pandemics on Stock Market Trends through Sentiment Analysis”, Journal of Xidian University, Vol. 14 (6): 1163-1176.
64. D. Sarddar, R. K. Dey, R. Bose, S. Roy, “Topic Modeling as a Tool to Gauge Political Sentiments from Twitter Feeds”, International Journal of Natural Computing Research, Vol 9 (2): 1-22.
<https://doi.org/10.4018/IJNCR.2020040102>
65. K. Chanda, P. Bhattacharjee, S. Roy, S. Biswas, “Intelligent Data Prognosis of Recurrent of Depression in Medical Diagnosis”, International Conference on Reliability, Infocom Technologies And Optimization (ICRITO 2020), 2020, pp. 1-5.
66. Dey, R.K., Sarddar, D., Sarkar, I., Bose, R. and Roy, S., A Literature Survey on Sentiment Analysis Techniques involving Social Media and Online Platforms, International Journal Of Scientific & Technology Research, 1(1).
67. Adhikari, S., Ghosh, A., Das, P., Roy, S. and Bose, R., 2020. An Image Based Activity Recognition. International Journal of Emerging Trends in Engineering Research, 8(5).
<https://doi.org/10.30534/ijeter/2020/65852020>