Mapana – Journal of Sciences
2023, Vol. 22, No. 1, 89-120
ISSN 0975-3303|https://doi.org/10.12723/mjs.64.6

A Study of Stock Market Prediction through Sentiment Analysis

Sandipan Biswas*, Shivnath Ghosh†, Sandip Roy‡, Rajesh Bose‡, Sanjay Soni§

ABSTRACT

In the modern world, the current state and course of economic development and growth are determined by the fortunes and vagaries of the stock markets. In this research study, the authors provide a model that can aid in making reliable and error-free predictions of stock market trends. The research's described approach uses sentiment analytics based on financial news and past stock market patterns. The suggested model can provide more accurate results because it examines information from numerous news sources as well as the history of price development of specific equities. The proposed structure has been used to forecast stock market patterns that incorporate sentiment analysis taken from news and previous stock market patterns to provide more precise results from a variety of news data and the history of the price up and down of specific equities. The model shown here has provided a two-step process. The Naive Bayes algorithm has been utilized in the initial step to assess text polarity and determine the general mood of news data gathered and received. The next stage involves

* Department of Computer Science & Engineering, Brainware University, West Bengal, India; sandipan_diet@rediffmail.com

† Department of Computer Science & Engineering, West Bengal, India; dsg.cs@brainwareuniversity.ac.in

‡ Department of Computational Science, Brainware University, West Bengal, India

§ Department of Industrial and Production Engineering, J.E.C Jabalpur, Madhya Pradesh-482002, India; soni563@yahoo.com

forecasting future values of stocks using evaluation findings on text polarity and historical stock value movement information. A novel idea known as the KNN-LR Hybrid algorithm has been introduced to achieve better outcomes when evaluating the accuracy and efficacy of other machine learning algorithms.

Keywords: Bombay Stock Exchange (BSE), Logistic Regression, KNN-LR Hybrid Classifier, Naïve Bayes Algorithm, Particle Swarm Optimization (PSO), Sentiment Analysis, Stock Market, Support Vector Machine, social media, Twitter.

Abbreviations: Bombay Stock Exchange (BSE); Particle Swarm Optimization (PSO); K-Nearest Neighbour Algorithm-Logistic Regression (KNN-LR); Knowledge Discovery from Data (KDD); Sentiment Analysis (SA); Support Vector Machine (SVM)

I. INTRODUCTION

The changes impacting the economy and financial concerns can cause stock values to soar skyward or fall precipitously in today's age of globalisation and Internet expansion on a scale that has never before been seen and may not have even been conceived a few decades ago. Global economies depend on stock markets and fluctuations in share values across major trading in US, Europe, and Asia as well, stocks are closely watched, and share markets are meticulously studied. Wall Street in New York City, USA, is where stock and share trading first started [1].

The challenge with deterministic approaches to stock market studies is that there are numerous intricate concerns that are influenced by politics and public opinion [2]. According to research, investor sentiments may prove to be a key consideration when making decisions. Investor feelings are crucial when it comes to the stock market and the movement of share prices [3]. Although investors may be more willing to take risks in some parts of the world, they rely on their purchasing or selling [4]. News that affects investor attitudes can spread quicker than ever because of the development and widespread use of the Internet and digital social media [5]. Researchers have discovered that weblogs and Twitter feeds have a substantial impact on noise trading. The consensus among academics and industry watchers is that anticipating share prices is the key to optimising returns on stock market trading [6]. However, predicting the direction that share values will move is an important area for research that affects not just academic but also global economic interests [7].

A. Twitter: A Data Source

Twitter has rapidly risen in popularity among digital social media platforms. Every day, many users from all around the world publish and exchange enormous amounts of information on this digital social media network [8]. Twitter is an attractive platform for researchers to conduct their research because user-posted messages can include not only photographs but also geographical locational information with embedded timestamps [8]. Twitter, which had a user base of more than 300 million people globally in 2015, enables communities to form around shared or similar interests in any location. Many academics believe that Twitter can perform better than any other source of data [9]. According to studies, there is no other website like Twitter for microblogging. In one survey, researchers came to the conclusion that 74% of online adults used digital social media, such as Twitter, 2014. Because of this, a greater understanding of how to extract and infer user sentiments from Twitter feeds or messages is necessary [10]. Twitter data mining has the potential to be used for research and social experimentation. Researchers' tests using Python and Twitter's API in a study [11] have shown that SA, in relation to stock market up-down, is a useful strategy for comprehending current happenings that do influence depositor behaviour.

B. Analytical Approach For Evaluation of Tweets

Despite not being particularly novel, researchers have successfully used Twitter data to analyse sentiment. Researchers have discovered that sentiment patterns can be mapped and represented using unsupervised classification algorithms in work [12].

In this instance, the researchers assessed the emotional content and mood of tweets using the Latent Dirichlet Allocation method. In our study [8], scientists suggested a model on Nave Bayes do sentiment evaluations using data from Twitter. The results of the trials carried out using this method suggest that stock value swings based on Twitter feeds and news editorials are likely to occur.

C. Data Mining

Data mining is used as a method for forecasting stock market movements. With good reason, it has also been referred to as knowledge discovery from data [13]. Predictive and descriptive jobs can be broadly divided into two categories when using data mining. Identifying and classifying the properties of the data under analysis is the purpose of descriptive tasks. On the other hand, predictive tasks use analyses of existing data to produce conclusions[14].

D. Predictive Approach Uses

We approach the topic of predicting stock market trends using categorization analysis and predictive tasks. We have taken into account Naive Bayes classification, Logistic Regression (LR), Support Vector Machine (SVM), and K-Nearest Neighbour (KNN) classification for our proposed solution and studies. We examine these techniques, develop an ensemble strategy to develop our suggested model, and then provide the findings of our study in this publication.

II. RELATED WORK

Sentiment analysis (SA) entails examining online data and classifying the text included therein. The goal of SA is to be able to determine an author's preferences with regard to any subject, including literary, political, social, and sports topics. SA can have a positive, negative, or neutral outcome [15]. For data mining, researchers have employed a variety of methodologies [16]. Naive Bayes classification, Support Vector Machine, Logistic Regression, and K-nearest-neighbour classification are the most well-known of these methods. A hybrid technique can be used for SA. Unsupervised learning and supervised learning are two categories of machine learning. In studies, supervised learning techniques like Naive Bayes classification and Support Vector Machine (SVM) are frequently utilised [16]. There are so many data classifiers like HRC and FRC [47][48] which are strong enough in complex Fuzzy data analysers but due to small data set issue we have compared and analysed with SVM, KNN,LR, and Naïve Bayes classification.

A. Particle Swarm Optimization Technique and Support Vector Machine

The necessity of developing a suitable approach for classifying data because accuracy is a factor to be taken into account when generating predictions. Support Vector Machines (SVMs) have been demonstrated to improve prediction outcomes in a study [17]. Researchers [18] have found that SVM is a cutting-edge machine-learning technique that can aid in sentiment classification. The Particle Swarm Optimization (PSO) method and SVM were both employed by the researchers to categorise the feelings expressed in Twitter movie reviews. The researchers Kennedy and Eberhart created PSO, a type of searching algorithm, taking cues from the way that fish schools and birds swarm. The foundation of PSO was the cooperative nature of birds in their search for food or nesting [20][21]. It is difficult to manually go through the massive amounts of text that are posted in social media news feeds and tweets. Researchers [22] have noted that automating data mining is viable at the aspect, document, and sentence levels when evaluating the best approach. A proposed model by the authors of the paper [22] is based in conjunction with a syntactic approach. Their tests showed that the proposed approach produced an accuracy of 78.04%, which was an improvement of almost 6% over the Part of Speech (POS) method on the annotated data set.

B. The Hybrid Method using SVM and PSO

Researchers generally use supervised data mining techniques in their work [23] shown in their study where they suggest a Nave Bayes classifier-based method to identify moviegoers' feelings from internet reviews. In these situations, opinion mining that leads to SA or text mining of web pages becomes extremely important [24] and realised the necessity to improve SA approaches and presented an SVM algorithm paired with the PSO technique.

The authors' research has shown their recommended approach for SA produced results with a great deal more precision and accuracy than the SVM approach alone. A hybrid strategy utilising SVM and

PSO offers strengths, according to the authors of a comparable study on text mining approaches [25]. The capacity to categorise sentiments into two groups: negative sentiments and positive sentiments, according to experiments carried out utilising their proposed model.

C. Natural Language Processing

However, from the perspective of some researchers, the vast majority of SA-dependent systems [26] refer to as the "bag-of-words" concept. The authors claim that because of the reliance on traditional machine learning techniques like Naive Bayes, SVM, etc., can result in inaccurate and frequently subpar outcomes. The classic ML algorithms have a history of bias. The authors suggest a model based on Natural Language Processing (NLP) and the inclusion of semantics in their model address the shortcomings of conventional ML techniques. Their paper illustrates the results achieved using their proposed model and shows that it is possible to gain an improvement by at least 3% over the “bag-of-words” method of classification during text mining.

D. SVM by Tuning the Parameters

According to another group of researchers, the predictions of the outcome that were made were to be accurate in the majority of cases. The performance of data classification using SVM, according to the authors [27], is highly needy for fine-tuning assured sets of values. Model selection is another name for parameter adjustment. Although all linearly separable issues can be solved using linear SVMs, this is not always the case to make it possible to identify and choose the parameter set that most closely meets the requirements. The resulting parameter set is then used to create the classifier by applying it to the training set of data. To obtain generalisation precision, the created set is used for a specific set. Similar arguments have been made by researchers [28], who claims that while SVM is a classifier that has been widely applied, its main drawback is that the user must provide regular cost parameters and kernel parameters. While that is not a significant issue, the authors explain that the choice in selecting cost parameters is critically important. The authors propose an algorithm that creates a complete path of SVM solutions corresponding to each cost parameter value. The authors add that that leads to the same computational cost as would be required for just a single SVM. Researchers [29] have proposed a strategy incorporating active learning to enhance the performance of SVMs. The authors provide a concept known as "version space" in their suggested method for improving SVM performance in inductive and transductive ML scenarios [30].

III. ALGORITHM STUDY FOR PROPOSED WORK

In this segment, we discuss and depict the algorithms in our suggested technique to evaluate which of them is more optimal. We also demonstrate how the construction of our suggested model that can be aided by the induction of better-performing algorithms. Prior to examining classical algorithms,we present a flowchart representing the underlying approach of the traditional model deployed for sentiment analysis as illustrated in Figure 1.

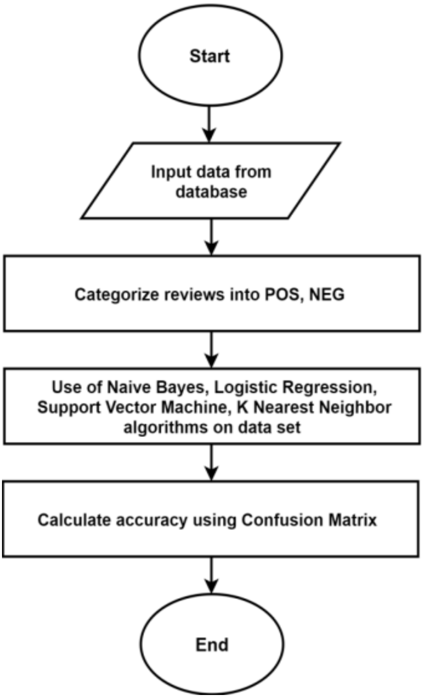

Start
Input data from database
Categorize reviews into POS, NEG
Use of Naive Bayes, Logistic Regression, Support Vector Machine, K Nearest Neighbor algorithms on data set
Calculate accuracy using Confusion Matrix
End

Fig. 1. Flow diagram of the classical paradigm used in sentiment analysis

A. Naïve Bayes Classifier

Author in [31] has shown that notwithstanding its independence assumption that is unrealistic. In situations when either independent or dependent features are present, the Naive Bayes classifier does a good job of finding solutions. While the authors believe that the former is to be expected, the latter is not in line with what people often expect from this classifier. Studies [32] show that the Nave Bayes classifier performs better than comparable ML algorithms when used in social media. In a different study, the authors [33] found that using the Naive Bayes classifier to examine market-influencing elements significantly improved their ability to predict stock market movements. The following is how a Bayes equation (1) can be created given a feature and any label:

$$P(y) = \frac{P(x)*P(y|x)}{P(y)} \quad (1)$$

Here P(x)stands for the priori of x.

P(y | x) signifies the likelihood of x.

P(y) refers to predictor prior probability.

The equation (1) can also be expressed as follows:

$$P(xb) = \frac{d\left(w,e(xb)\right)}{sl\left(d\left(w,e(x)\right)y\right)} \quad (2)$$

A two-step approach involving Naïve Bayes Classifier is proposed and explained in Figure 2. In the initial phase, inputs for training are accepted. These are then forwarded to the computation phase to obtain the expected outcome.

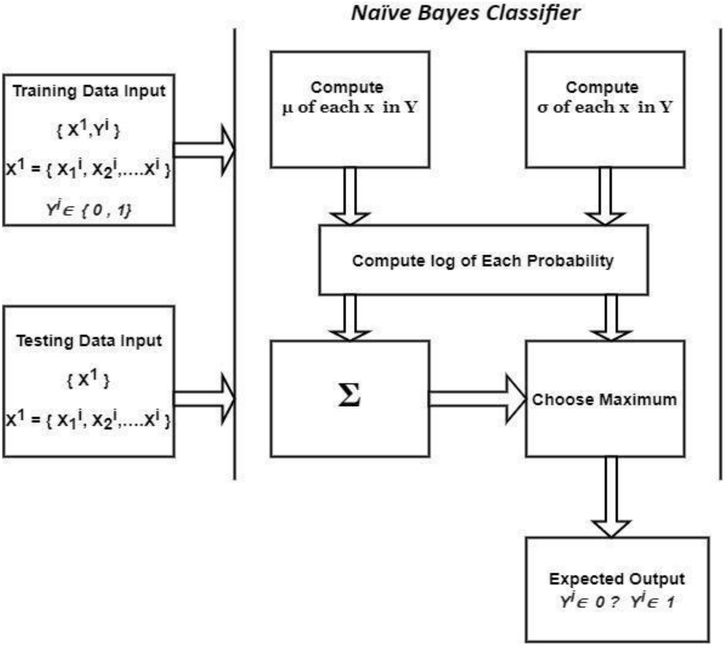


Fig. 2. Architecture of Naïve Bayes Classifier.

B. Logistic Regression

When there are typically just two possible outcomes for answers, the logistic regression (LR) model is frequently utilised. Even though it is possible to use LR models in situations where there may be more than two events, these are often less frequent [34]. The capacity of the LR algorithm to identify which of the factors being evaluated has a high probability is a key benefit. LR can also help in determining the size of a potential influence [35].

In the following section, we describe the logistic regression method. Assume *X and y* are datasets that have two possible results. The result is either y_i=1 or y_i= 0 for every occurrence x_i in *X*. An instance x_i is called positive class if outcomes y_i=1 while instances are called negative class if outcomes y_i= 0.

We proceed to construct a regression paradigm that distinguishes an occurrence x_i as a negative or positive group. The function is initiated to set up a concluded-form dominion among a set of features. The expression is denoted in equation (3).

$$P(y = 1|X, \beta) = \frac{e^{\beta X}}{1 + e^{\beta X}} \qquad (3)$$

A dominion from (3), the optimum set of factorβ is defined by extending

$$\prod_{i=1}^{n} \quad P(y_i | x_i, \beta) \qquad (4)$$

Or, in conventional representation

$$\beta^* = arg\{ \prod_{i=1}^{n} \quad P(y_i | x_i, \beta) \}$$

$$= arg\{ \sum_{i=1}^{n} \quad y_i \, log \, log \left(\frac{1}{1+e^{-\beta X}} \right) + (1 - y_i) \, log \, log \left(\frac{1}{1+e^{-\beta X}} \right) + \} \qquad (5)$$

Defining ideal factorβ* is called the training set. For an undetected instance *x*, and an optimum set of factorβ*, it can be classed after computing the (3).

If $P\,(y{=}1 \mid x, \beta^*) \geq 0.5$, the instance *x* be considered as positive, else, it is considered as negative.

Here in Fig. 3 the basic architecture of Logistic regression is described where the inputs are taken as an array of data and then it is forwarded to the 3-phase computational function stage:

-net input function,

-activation function,

-unit step function,

and the expected output is generated.

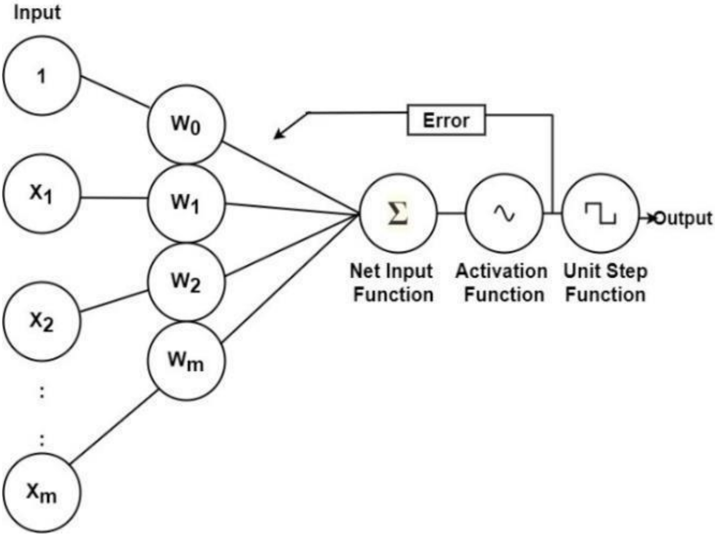


Fig. 3. Architecture of Logistic Regression.

C. Support Vector Machine

A kind of supervised machine learning model are called support vector machines, often known as support vector networks [36]. This kind of classifier is frequently used. SVM is a non-probabilistic binary sort of linear classifier, according to researchers.

Support Vector Machine is one of the simplest linear classification techniques. It is based on the segregation of several classes by the determination of a hyperplane that can best optimise margin among various classes. We have outlined the SVM approach in the part that follows. In most cases, n vectors xi will be joined to form the data. Each xi will additionally have a value yi that will indicate if the element is a member of the class (+1) or not (-1).

It is to be noted that y_i can only have two possible outcomes of -1 or +1. Moreover, most of the time vector x_i results in being associated with a lot of dimensions. It can, therefore, be suggested that x_i is a pp-dimensional vector if it has p-dimensions. In the example dataset considered, it is composed of a set of n couples of elements (x_i, y_i).

The more formal definition of an initial dataset in set theory is:

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

In the following figure Fig. 4, Support Vector Machine is composed of inputs that are forwarded into 3-layer phase stage following which output is generated.

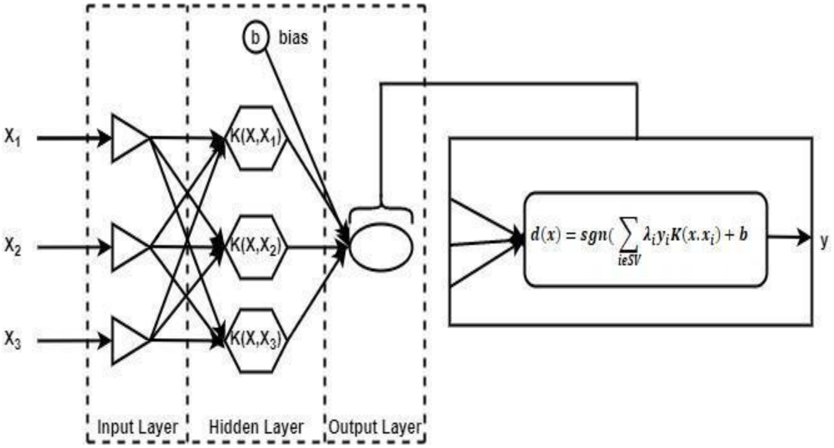


Fig. 4. Diagram of Support Vector Machine

D. k-Nearest Neighbour Algorithm (KNN)

According to the description, the KNN algorithm is a non-parametric approach used in classification and regression exercises [37]. The k closest training examples constructed in feature space serve as the algorithm's inputs.

KNN is a learning algorithm based on supervised method where the output of a new illustration is categorized and it depends on the bulk of K-NN classification. This classification ignores any prototype for adaptation and is based only on memory. For the purposes of our experiment, we assume that all occurrences with m attributes are considered like two classes such as +ve or -ve class. Specified a query casex_q,the training samples would be nearby to it, which is denoted by$N_k(x_q)$are found. By$N_k^+(x_q)$, and by$N_k^-(x_q)$, positive and negative instances sets in $N_k(x_q)$*are* designated disjointedly. If | $N_k^+(x_q)$ |>| $N_k^-(x_q)$ |, x_q is considered as +ve class; else it is considered as -ve one. At the same time, KNN can easily be adjusted to predict continual-valued objective functions.

To achieve this, we algorithm ascertains the probability of the positive class is corrected with the subsequent equation.

$$p(x_q) = \frac{|N_k^+(a_i(x))|}{|N_k^-(a_i(x))|} \qquad (7)$$

In the following Figure 5, the basic block diagram of K-NN Algorithm is described where input data are forwarded to 2-phase computational function stage (transfer function, activation function) and the output is generated.

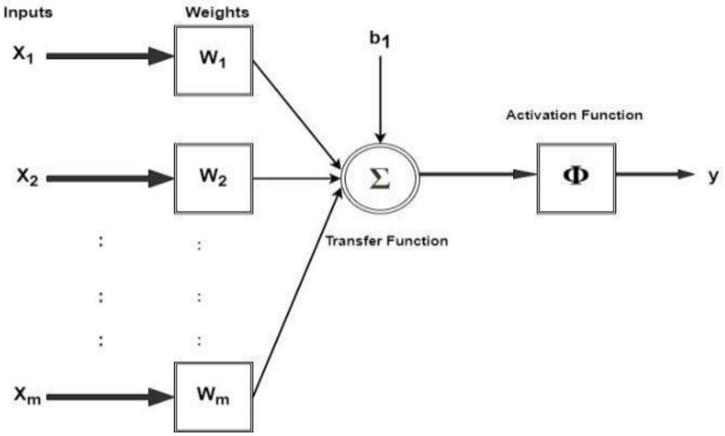


Fig. 5. Architecture of k-Nearest Neighbour Algorithm

In Table 1, we compare the characteristics of the algorithms previously described.

Algorithms	Characteristics
Naïve Bayes Classifier	In case the conditional independence of the Naïve Bayes algorithm is satisfied, the error values are comparably lower with comparable or even marginally well effective than competing discriminative models. It supports binary and multi-class classification problems. Naive Bayes is taken when the characteristics are mutually independent and have limited training data.

Algorithms	Characteristics
	Naive Bayes is dominant in text classification, spam filtering, recommender systems, etc., and also handle irrelevant features.
Logistic Regression	It is a trouble-free, speedy, and basic classification process and could be applied in favor of multiclass classifications too. The derivation function is constantly curved. It could not be operated on non-linear classification problems. The precision of LR model are tampered by Co-linearity and outliers.
Support Vector Machine	SVM uses a core approach to resolve complicated results. A convex optimization function is used by SVM, because of which global minima are forever attainable. The pivot shortfall gives greater precision. Outliers could be sound controlled by constant C, which is a soft margin.
k-Nearest Neighbour Algorithm	It is an effortless and facile machine-learning model. k should be carefully chosen. Huge computing amount in runtime if the sample size is bulk. In this method appropriate scaling should be offered for good remedy amid features.

Table 1: Comparison of characteristics of different ML algorithms.

IV. THE PROPOSED MODEL

We have obtained effective results by using the algorithms KNN and LR in our paper. However, there are certain disadvantages associated with both LR and KNN algorithms.

First, the logistic function of the parameters has been defined. The relation is supposed as repetitious. This postulation, though, is not constantly correct in actual conditions and may depreciate classification precision.

Second, LR desires definite variables pre-processing. Steady and precision logistic regression prototypes are endangered in these situations.

Third, KNN is comparatively slower than Logistic Regression.

To defeat this type of difficulty, hybrid models means a combination of logistic regression and neural networks have exhibited and named KNN-LR(Hybrid) algorithm for effective results in stock market sentiment analysis [38, 39].

A. The KNN-LR (Hybrid) Classifier

D is a mapping, which is a binary classifier such that, $D: S^m \rightarrow [0, 1]$, Here S^m is an m-dimensional value, which contains a real number. *D(x)* could be regarded by a probability function for a vector $x \in S^m$. The value of *D(x)* is the probability such that *x* is considered as positive. In this article, we will find *D which* will be found by a grouping of K-NN, and LR both. Let *K* and *L* be the mappings composed by K-NN, LR both correspondingly, A new assigned mapping D, formation of *K* and *L*, is taken by us such that *D* =*L.K.*

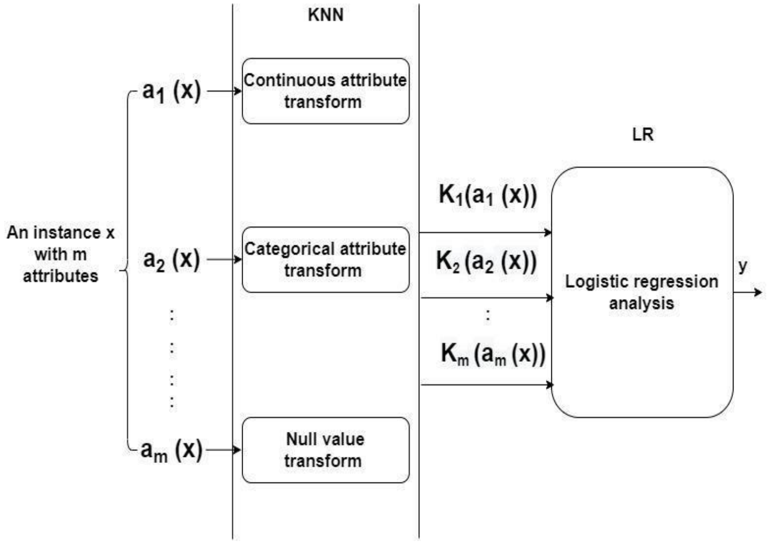
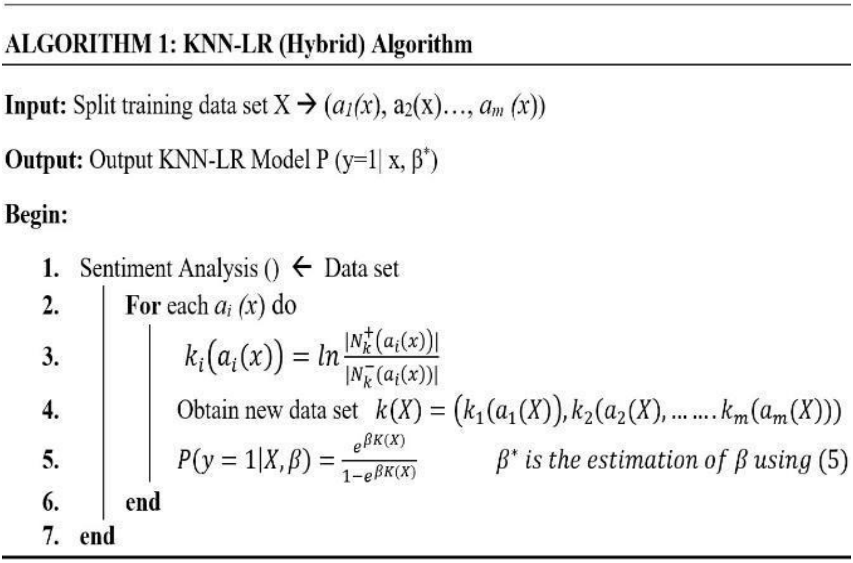

KNN
LR
An instance x with m attributes
a_1 (x)
a_2 (x)
a_m (x)
Continuous attribute transform
Categorical attribute transform
Null value transform
$K_1(a_1(x))$
$K_2(a_2(x))$
$K_m(a_m(x))$
Logistic regression analysis
y

Fig. 6. Architecture of proposed model KNN-LR (Hybrid).

A training data set$X = \{x_1, x_{12}, \dots. x_n\}$ is given in m-dimension and $Y = \{y_1, y_2, \dots. y_n\} \subseteq \{0,1\}$ be a set of labels. our aim is to form a hybrid data classifier D, made up of a KNN classifier K: Sm → ℜm and a LR classifier L: ℜm → [0,1]. The two steps that could be taken to reach this goal are depicted in Fig. 6. X is first transformed into a new data set K(X), which is made up of a collection of m-dimensional real vectors, and K is then taught by the data set X, Y. A vector in K (X) with a greater vector norm also has a better trend towards the positive class when y = 1. K (X) trains LR classifier L in the second step. The KNN-LR training algorithm is described in pseudo-code in Algorithm 1.



ALGORITHM 1: KNN-LR (Hybrid) Algorithm

Input: Split training data set X → ($a_1(x)$, $a_2(X)$…, a_m (x))

Output: Output KNN-LR Model P (y=1| x, β*)

Begin:

```
Sentiment Analysis () ← Data set
    For each $a_i$ $(x)$ do
        $k_i(a_i(x)) = ln \frac{|N_k^+(a_i(x))|}{|N_k^-(a_i(x))|}$
        Obtain new data set  $k(X) = (k_1(a_1(X)), k_2(a_2(X), \dots\dots. k_m(a_m(X)))$
        $P(y = 1|X, \beta) = \frac{e^{\beta K(X)}}{1-e^{\beta K(X)}}$        $\beta^*$ is the estimation of $\beta$ using (5)
    end
end
```

We assume that a random instance x, which is labeled by the feature vector$(a_1(x), a_2(x), \dots. a_m(x))$, where ar(x)represents the value of the feature of occurrence x. Herewith,X is a training data set which has m features could be distributed into m number of data sets $(a_1(X), a_2(X), \dots. a_m(X))$ by features. Any number of divided data sets are used with KNN, and the instances of these data sets converge to points in the one-dimensional space. Each KNN classifier is linked to a different, unrelated data collection that is utilised to implement classification. Using the training data set ar(X) and the value of the rth attribute of instances x, we can calculate the likelihood of a perfect positive class x(7). A real value objective function k r (ar (x)), log odds rate of p (ar(xr)), is in the form:

$$k_r\left(a_r(x)\right) = ln\frac{p\left(a_r(x_r)\right)}{1-p\left(a_r(x_r)\right)} = ln\frac{|N_k^+\left(a_i(x)\right)|}{|N_k^-\left(a_i(x)\right)|} \qquad (8)$$

The description is vital to pursue the nearby neighbours of an occurrence. The distance linking the identical characteristic of two occurrences xi and xj is stated to be $d(a_r(x_i), a_r(x_j)$ where equation (9) and equation (10) for continuous-valued attributes and categorical-valued attributes respectively.

$$d\left(a_r(x_i), a_r(x_j)\right) = |a_r(x_i) - a_r(x_j)| \qquad (9)$$

$$d\left(a_r(x_i), a_r(x_j)\right) = \{0 \;\; a_r(x_i) = a_r(x_j) \;\; 1 \;\; a_r(x_i) \neq a_r(x_j) \qquad (10)$$

We assume that the present value is extremely distinct from absent attribute values. If they are both absent, then $d\left(a_r(x_i), a_r(x_j)\right)$produces 0. The unspecified parameter β in (9) is assessed by the changed training data set. The subsequent KNN-LR (hybrid) classifier, which is represented by equation (11)

$$P(X, \beta^*) = \frac{e^{\beta^* K(X)}}{1- e^{\beta^* K(X)}} \qquad (11)$$

In this division, we explain in detail how our proposed model functions. We first present the flowchart as seen in Fig.7 before getting more into its characteristics below:

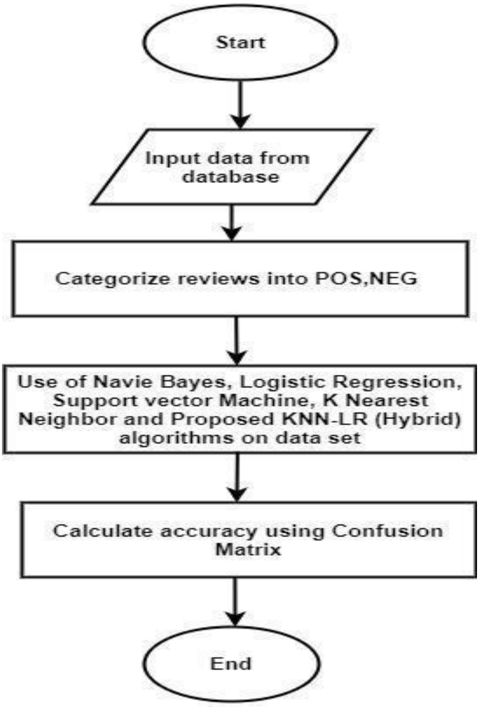

Start
Input data from database
Categorize reviews into POS,NEG
Use of Navie Bayes, Logistic Regression, Support vector Machine, K Nearest Neighbor and Proposed KNN-LR (Hybrid) algorithms on data set
Calculate accuracy using Confusion Matrix
End

Fig. 7. Flowchart of proposed model

B. Data Description

We used news from the Yahoo! Finance website and data gathered from Reddit.com, a US social news website, for the purposes of our research. The majority of the content sets were compiled from these two sources. Every day, news articles on the company and stock information were gathered. To help in forecasting future stock market rates, stock prices when opening, highest and lowest prices, as well as closing values were also gathered.

News headlines over the period of March 2019 to September, 2019 on Reddit News are extracted using the Spider Anaconda application. Spider Anaconda application is a python application that helps us to extract news headlines from reddit.com. The headlines were collected using Python and stored in a .csv file which is again transferred to a .csv file to apply the algorithms. Stock opening and closing prices of BSE (Bombay Stock Exchange) from March, 2019 to September 2019 are obtained from Yahoo! Finance.

C. Sentiment Analysis component

Our proposed model's component stage at this position includes an investigation of the partiality and bias of stock news. The classification is expressed as follows: in the case of news headlines, the goal is to categorise news as either carrying positive or negative attitudes. The Nave Bayes classification technique is then used to classify the news. The steps that make up the procedure are described in the section after this.

D. Text Pre-processing

Creating Tokens. Every news heading or financial report is now first broken down into language. Tokens are what these are.

Data uniformity and equality. All words, or tokens, are changed to lowercase and assembled in a text to facilitate uniformity and data processing.

Elimination of Non-essential Words and Punctuation. To improve performance and reduce the amount of features that need to be recorded, articles like "a," "an," and "the," as well as prepositions, are deleted. At this point, stop words and other punctuation are also eliminated.

Stemming. To reduce document complexity by processing only single phrases, we remove suffixes like "-ed," "-ing," and "-ion" using the Porter Stemmer algorithm [40, 41, 42]. This improves how well our suggested model performs.

E. Abbreviation Processing

There is made a list of abbreviations, such as "US" for the United States of America and other such terms. The related abbreviations are then replaced with full terminology or extended terms.

Filter Tokens. Words that contain two characters or less are identified and removed as part of the filtering process. Word vectors are represented as each word corresponding to a real vector in order to minimize complexity when dealing with data that contains text. In Fig. 8, we have discussed a general training and prediction process and how it works.

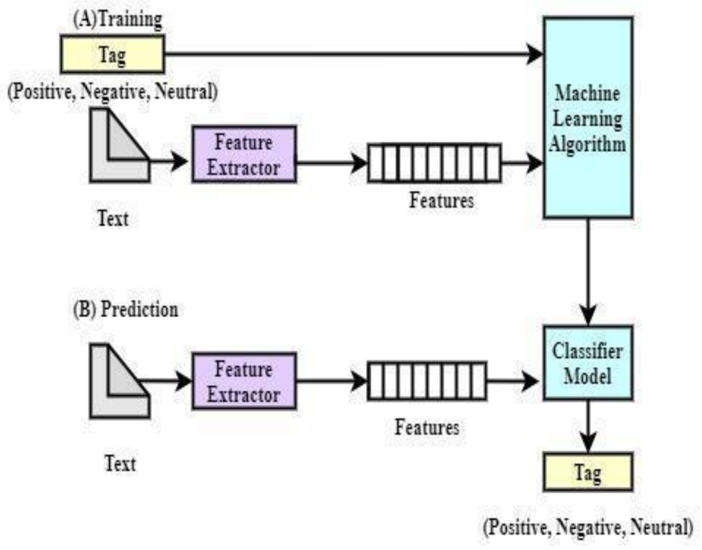


Fig. 8. Training and prediction process.

F. Data Table

We add two sample tables – one of which is Table 2 that contains samples of news headlines of data scrapped from Reddit.com and stored in a .csv file. In Table 3, sample data of value scrapped from Yahoo Finance of BSE are stored.

Table 2: Sample data collected from Reddit.com.

Sample Headlines	**Sentiment**
How can one safeguard investments?	Positive
What is happening with the NHAI's financial health?	Negative
Is the stock market going to fall tomorrow?	Negative
Ongoing "stimulus" from Finance Minister'	Positive
Indian investors that I should be following?'	Positive
Are the oil price set to rise now?	Positive
How to find growth potential in a company?	Positive

Sample Headlines	Sentiment
Do we have any Intraday/Swing traders on this sub who make regular money from trading?	Positive
Another NBFC Altico defaulting on interest payment	Negative
Is this the end of the road for Vodafone Idea?'	Negative
Will ITC ever go up?'	Positive
Will the ban on single-use plastics affect FMCG?'	
Purchase Indian stocks, which could see 10% yields in the following 6 months, says Credit Suisse'	Positive
Market Value for all debt funds from April 1, 2020	Positive
Are Real Estate ETFs/REITs liable to deemed rental income tax?	Negative
Essel group makes partial repayment to mutual fund houses	Positive

Date	Open	High	Low	Close	Adj. Close	Volume	Level value
17-09-19	37169.46	37169.56	36419.09	36481.09	36481.09	20500	0
16-09-19	37204.56	37302.06	37028.94	37123.31	37123.31	19900	0
13-09-19	37175.86	37413.5	37000.09	37384.99	37384.99	34300	1
12-09-19	37330.47	37435.15	37048.67	37104.28	37104.28	27500	0
11-09-19	37251.03	37343.46	37193.57	37270.82	37270.82	36700	1
09-09-19	36969.48	37244.08	36784.47	37145.45	37145.45	25300	1
06-09-19	36785.59	37012.98	36727.66	36981.77	36981.77	23000	1
05-09-19	36821.71	36898.99	36541.88	36644.42	36644.42	27400	0
04-09-19	36575.24	36776.31	36409.54	36724.74	36724.74	25600	1

Date	**Open**	**High**	**Low**	**Close**	**Adj. Close**	**Volume**	**Level value**
03-09-19	37181.76	37188.38	36466.01	36562.91	36562.91	23600	0
30-08-19	37222.26	37397.97	36829.81	37332.79	37332.79	33400	1
29-08-19	37381.8	37381.8	36987.35	37068.93	37068.93	31700	0
28-08-19	37655.77	37687.82	37249.19	37451.84	37451.84	28600	0
27-08-19	37658.48	37731.51	37449.69	37641.27	37641.27	36500	1
26-08-19	37363.95	37544.48	36492.65	37494.12	37494.12	35600	0
23-08-19	36387.68	36807.34	36102.35	36701.16	36701.16	38400	1

Table 3: Sample data collected of stock market obtained from yahoo finance.

V. Programming Language and Simulator

We use Python programming language, which is a general-purpose, high-level, and interpreted. In 1991, Python was created by Guido van Rossum. In terms of software development, academic tools, and scientific research, Python has advanced significantly. It places a strong emphasis on the readability of code, which makes it popular among software engineers and researchers working on projects of all shapes and sizes [43]

Vi. Experiment and Result Analysis

We choose a random sample of equities that are typically traded in high quantities for our research. Then, we gather reviews and comments made by others about the chosen stocks. We "purchase" the stock if the reviews are favorable; otherwise, we "sell" the stock. In our tests, stock purchases and sales are simulated rather than taking place in the actual market. On the data supplied in tabular form and organised by dates, we apply the Naive Bayes, LR, SVM, and k-NN algorithms. An accuracy score is created using the Confusion Matrix. We use the confusion matrix in the matrix (Table 4) to compare expected values to actual group values.

	-ve	**+ve**
-ve	FN	FP
+ve	TN	TP

Table 4: General Confusion matrix.

The labels are illustrated below:

True Positive (TP): Positive tuples labeled appropriately through classes.

True Negative (TN): Negative tuples labeled appropriately through classes.

False Positive (FP): Negative tuples wrongly labeled as positive.

False Negative (FN): Positive tuples incorrectly labeled as negative.

A. Naïve Bayes Algorithm

The confusion matrix of the Naïve Bayes algorithm is shown in Table 5.

	-ve	**+ve**
-ve	158	1
+ve	6	0

Table 5: Confusion matrix for the Naïve Bayes Algorithm.

B. Logistic Regression Algorithm

The confusion matrix of the LR algorithm is described in Table 6.

	-ve	**+ve**
-ve	159	0
+ve	6	0

Table 6: Confusion matrix for the LR Algorithm.

C. Support Vector Machine (SVM) Algorithm

The confusion matrix for the SVM is shown in Table 7.

	-ve	**+ve**
-ve	156	3
+ve	6	0

Table 7: Confusion matrix for the SVM Algorithm.

D. K-Nearest Neighbour Algorithm

The confusion matrix of the K-Nearest Neighbour algorithm is described in table 8.

	-ve	**+ve**
-ve	159	0
+ve	6	0

Table 8: Confusion matrix for the KNN Algorithm.

227 tokens in total are taken into account for the analysis, which determines each algorithm's accuracy, are shown in Table 9 below. We believe that the k-Nearest Neighbour and Logistic Regression algorithms rank equally (NN = 3), taking into account the outcomes of the algorithmic evaluation. In our opinion, additional research using various optimization techniques may be necessary to evaluate comparative viability in this research direction.

Algorithm	**Accuracy Score**
Naïve Bayes Algorithm	95.76%
Logistic Regression Algorithm	96.36%
Support Vector Machine Algorithm	94.54%
K-NN Algorithm	96.36%
KNN-LR (Hybrid)	96.89%

Table 9: Comparison between of discussed algorithms.

From analysing the algorithms we have found the results with value Naïve Bayes Algorithm with accuracy score of 95.76%, LR Algorithm with accuracy score 96.36%, SVM Algorithm with accuracy score 94.54%, K-NN Algorithm with accuracy score 96.36%, KNN-LR (Hybrid)with accuracy score96.89%. We have also compared the above algorithms by using the logarithmic trend to get the more accurate level among the above algorithms.

We also compare the four algorithms in Fig .9. Also, we have compared the above algorithms by using the logarithmic trend to get the more accurate level among the above algorithms. We have also compared the accuracy level of different Machine Learning algorithms. We have also shown that both the Logistic Regression algorithm and K-Nearest Neighbour algorithm gives better accuracy in sentiment analysis. We have also proposed a new concept KNN-LR (Hybrid) algorithm to get more accurate results.

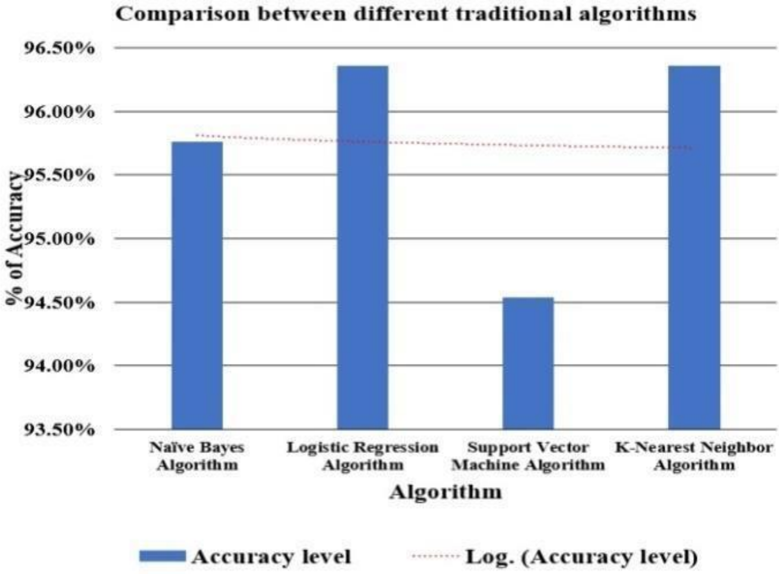


Fig. 9. Comparison of accuracy between different traditional algorithms

E. Proposed KNN-LR (Hybrid) Algorithm

We can employ a mixed method, which combines the logistic regression technique and the k-nearest neighbour algorithm. This hybrid KNN-LR quantifier can enhance the overall execution of the logistic regression in classification accuracy.

By increasing the value of k, one can improve KNN-LR. In these trials, we set k = 10. The results are shown in Table 10 and a relationship is provided in Figure 10. Here in our KNN-LR hybrid algorithm, we have used two equations for the implementation and prediction of results.

$$k_r\left(a_r(x)\right) = ln\frac{p\left(a_r(x_r)\right)}{1-p\left(a_r(x_r)\right)} = ln\frac{|N_k^+\left(a_i(x)\right)|}{|N_k^-\left(a_i(x)\right)|} \tag{12}$$

$$P(X,\beta^*) = \frac{e^{\beta^* K(X)}}{1-e^{\beta^* K(X)}} \tag{13}$$

Table 10: Confusion matrix for the Proposed KNN-LR (Hybrid) Algorithm.

	-ve	**+ve**
-ve	159	0
+ve	6	0

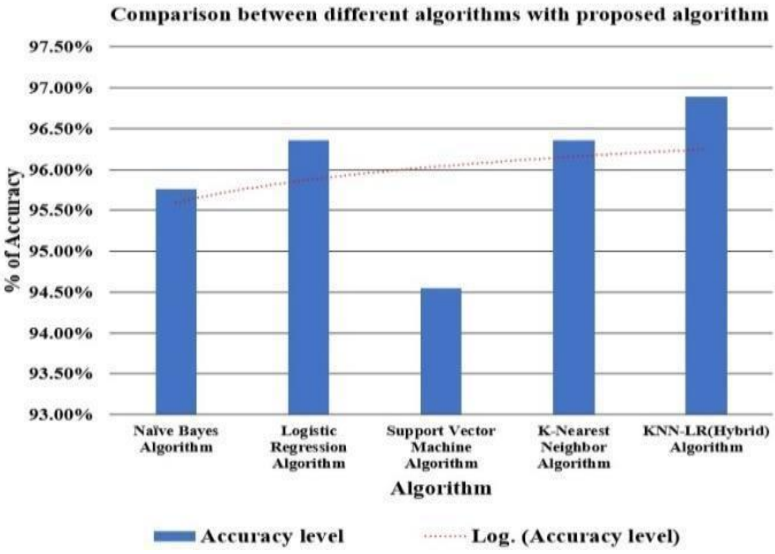


Fig. 10. Comparison of accuracy between different algorithms with our proposed algorithm

VII. Conclusion

Here in our paper four well-known machine learning algorithms' performances were examined and analysed while our proposed model was being designed and developed. We have looked at how well these algorithms foretell stock market patterns based on feedback from users. According to a comparative analysis done throughout the research, k-NN and Logistic Regression are much more effective in returning greater rates of accuracy when it comes to forecasting stock market trends. Nevertheless, our suggested KNN-LR (Hybrid) approach provides similarly high accuracy rates while keeping higher performance levels to KNN and LR algorithms. Here our data set collected is not so large, and so for the next time data analysis we will use large data set for better accuracy of prediction.

VIII. Future Scope

In our paper, we have applied Confusion Matrix [44, 45, 46] to categorize classifiers to evaluate performance from sample data. We have used limited data with column values, and so if we use large data with more data than not only positive/negative, we can also analysis data in respect of weak positive or weak negative also. Test results indicate that while Our suggested model can increase forecast accuracy; more study is advised to increase accuracy across time and across various online news sources.

IX. Acknowledgements

This research did not receive definite funding from any aid organization in the community, business-related, or charitable areas.

Conflict of Interest: Here, the authors have no discord of curiosity.

References

[1] Wyss, B. (2001). Fundamentals of the Stock Market. *McGraw Hill*, 1–245.

[2] Khan, W., Ghazanfar, M. A., Azam, M. A., Karami, A., Alyoubi, K. H., and Alfakeeh, A. S. (2020). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing.*

[3] See-To, E. W. K., and Yang, Y. (2017). Market sentiment dispersion and its effects on stock return and volatility. *Electronic Markets*, 27: 283–296.

[4] Xindan, LI., and Bing, Z. (2017). Stock market behavior and Investor sentiment: Evidence from China. *Front. Bus. Res. China*, 2(2): 277–282.

[5] Smailovic, J., Grcar, M., Lavrac, N., and Znidarsic, M. (2013). Predictive Sentiment Analysis of Tweets: A Stock Market Application. *Human-Computer Interaction and Knowledge Discovery in Complex, Unstructured, Big Data. HCI-KDD 2013. Lecture Notes in Computer Science*, 7947: 77-88.

[6] Wang, H., and Ou, P. (2009). Prediction of Stock Market Index Movement by Ten Data Mining Techniques. *Modern Applied Science*, 3(12):28-42.

[7] Shu-e, Y., Qiang, Z. (2009). Noise Trading, Investor Sentiment Volatility, and Stock Returns. *Systems Engineering – Theory & Practice*, 29(3):40-47.

[8] Pal, R., Pawar, U., Zambare, K., and Hole, V. (2020). Predicting Stock Market Movement Based on Twitter Data and News Articles Using Sentiment Analysis and Fuzzy Logic. Second *International Conference on Computer Networks and Communication Technologies. ICCNCT 2019. Lecture Notes on Data Engineering and Communications Technologies*, 44:561-571.

[9] Bose, R., Dey, R. K. Roy, S., and Sarddar, D. (2018). Analyzing Political Sentiment Using Twitter Data. *Information and Communication Technology for Intelligent Systems. Smart Innovation, Systems and Technologies*, 107:427-436.

[10] Federer, L. M., Belter, C. W., Joubert, D. J. Livinski, A., Lu, Y-L., Snyders, L. N., and Thompson, H. (2018). Data sharing in PLOS ONE: An analysis of Data Availability Statements. *PLoS ONE*, 13(5):1-12.

[11] Das, N., Ghosh, P., and Roy D. (2020). Effect of Demonetization on Stock Market Co-rrelated with Geo-Twitter Sentiment Analysis. In: Dawn S., Balas V., Esposito A., Gope S. (eds*) Intelligent Techniques and Applications in Science and Technology. ICIMSAT 2019. Learning and Analytics in Intelligent Systems*, 12:780-797.

[12] Bose, R., Dey, R. K., Chakraborty, S., Roy, S., and Sarddar, D. Examining Hidden Meaning of E-commerce Platform.

International Journal of Innovative Technology and Exploring Engineering (IJITEE), 8(12):257-261.

[13] Han, J., Kamber, M., Pei, J. (2011). Data Mining Concepts and Techniques. *Morgan Kaufmann*. 3:1-744.

[14] Namugera, F., Wesonga, R., and Jehopio, P. (2019). Text mining and determinants of sentiments: Twitter social media usage by traditional media houses in Uganda. *Computational Social Networks*, 6(3):1-21.

[15] Aqlan, A.A.Q., Manjula, B., Lakshman Naik, R. (2019). A Study of Sentiment Analysis: Concepts, Techniques, and Challenges. *International Conference on Computational Intelligence and Data Engineering. Lecture Notes on Data Engineering and Communications Technologies*, 28:14-162.

[16] Sharma, D., Sabharwal, M., Goyal, V., and Vij M. (2020). Sentiment Analysis Techniques for Social Media Data: A Review. *First International Conference on Sustainable Technologies for Computational Intelligence. Advances in Intelligent Systems and Computing*, 1045:.75-90.

[17] Kothari, A. A., and Patel, W. D. (2015). A Novel Approach Towards Context Sensitive Recommendations Based on Machine Learning Methodology. *2015 Fifth International Conference on Communication Systems and Network Technologies*, 1114-1118.

[18] Basari, A. S. H., Hussin, B., Ananta, G. P. (2012). Opinion Mining of Movie Review using Hybrid Method of Support Vector Machine and Particle Swarm Optimization. *Malaysian Technical Universities Conference on Engineering and Technology (MUCET)*. 4:545-552.

[19] Gopal, A., Sultani, M. M., Bansal, J. C. (2019). On Stability Analysis of Particle Swarm Optimization Algorithm. *Arabian Journal for Science and Engineering*, 1-10.

[20] Kennedy, J., and Eberhart, R. (1995). Particle swarm optimization. *International Conference on Neural Networks*, 4:1942-1948.

[21] Khan, W., Ghazanfar, M.A., Azam, M.A., Karami, A., Alyoubi, K. H., Alfakeeh, A. S. (2020). Stock market prediction using machine learning classifiers and social media, news. *Journal of Ambient Intelligence and Humanized Computing.*

[22] More, P., and Ghotkar, A. (2016). A Study of Different Approaches to Aspect-based Opinion Mining. *International Journal of Computer Applications*, 145(6):11-15.

[23] Smeureanu, I., and Bucur, C. (2012). Applying Supervised Opinion Mining Techniques on Online User Reviews. *Informatica Economica*, 16(2):81-91.

[24] Sharma, D., and Sabharwal, M. (2019). Sentiment Analysis for Social Media using SVM Classifier of Machine Learning. *International Journal of Innovative Technology and Exploring Engineering (IJITEE)*, 8(9):39-47.

[25] Umamaheswari, K., Rajamohana, S. P., and Aishwaryalakshmi, G. Opinion Mining using Hybrid Methods. *International Journal of Computer Applications*, 18-21.

[26] Sarddar, D., Dey, R. K., Bose, R., and Roy, S. Topic Modeling as a Tool to Gauge Political Sentiments from Twitter Feeds. *International Journal of Natural Computing Research (IJNCR)*, 9(2):1-22.

[27] Srivastava, D. K., and Bhambhu, L. (2009). Data Classification using Support Vector Machine.*Journal of Theoretical and Applied Information Technology*, 1 – 7.

[28] Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004). The Entire Regularization Path for the Support Vector Machine. *Journal of Machine Learning Research*, 5:1391-1415.

[29] Tong, S., and Koller, D. (2001). Support Vector Machine Active Learning with Applications to Text Classification. *Journal of Machine Learning Research*, 45-66.

[30] Chang, C-C., Lin, C-J. (2011). LIBSVM: A library for support vector. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1-27.

[31] Watson, T.J. (2001). An empirical study of the naive Bayes classifier. 1-6.

[32] Ahmad, S., Asghar, M.Z., Alotaibi, F.M., and Awan I. (2019). Detection and classification of social media-based extremist affiliations using sentiment analysis techniques. *Human Centric Computing and Information Science.* 9:24.

[33] Ding, G., and Qin, L. (2019). Study on the prediction of stock price based on the associated network model of LSTM. International Journal of Machine Learning and Cybernetics. 1-11.

[34] Wilson, J.R., and Lorenz, K.A. (2015). Standard Binary Logistic Regression Model. In: Modeling Binary Correlated Responses using SAS, SPSS and R. ICSA Book Series in Statistics, 9:25-54.

[35] Tolles, J. Meurer, W. J. (2016). Logistic Regression Relating Patient Characteristics to Outcomes. JAMA Guide to Statistics and Methods. 316(5):533-534.

[36] Phienthrakul, T., Kijsirikul, B., Takamura, H., and Okumura, M. (2009). Sentiment Classification with Support Vector Machines and Multiple Kernel Functions. *In: Leung C.S., Lee M., Chan J.H. (eds) Neural Information Processing. ICONIP 2009. Lecture Notes in Computer Science*, 5864:583-592.

[37] Cai,L., Hofmann, T. (2004). Hierarchical Document Categorization with Support Vector Machines. *CIKM'04, ACMI*. 78-87.

[38] Lee, T. S., Chiu, T. C. C., Lu, C. J,and Chen, I. F.(2002). Credit scoring using the hybrid neural discriminant technique.*Expert Systems with Applications*, 23(3):245–254.

[39] Bentz, Y., and Merunka, D. (2000). Neural networks and the multinomial logit for brand choice modelling: a hybrid approach. *Journal of Forecasting*, 19(3):177–200.

[40] Altman, N.S. (1992). An Introduction to Kernel and Nearest-Neighbour Nonparametric Regression. The American Statistician. 46(3):175-85.

[41] Verma, P., Om, H. (2019). A novel approach for text summarization using optimal combination of sentence scoring methods. *Sādhanā*. 44 (110).

[42] Porter, M. (1980). An algorithm for suffix stripping. *Program: electronic library and information systems.* 14(3): 130-137.

[43] Bose R., Dey, R.K., Roy, S., and Sarddar D. (2020). Sentiment Analysis on Online Product Reviews. *Advances in Intelligent Systems and Computing*, 933:559-569.

[44] Ting K.M. (2017). Confusion Matrix. *Encyclopedia of Machine Learning and Data Mining.*

[45] Singh, R., and Baidya, D. (2019). Usage of Data Science to Predict String Integrity Failures. *Society of Petroleum Engineers*.

[46] Chandra, K., Bhattacharjee, P., Roy, S., Biswas, S. (2020). Intelligent Data Prognosis of Recent of Depression in Medical Diagnosis. *ICRITO'20, IEEE.* 1-5.

[47] Narayana, G. &Kolli, Kamakshaiah.(2021). Fuzzy K-means clustering with fast density peak clustering on multivariate kernel estimator with evolutionary multimodal optimization clusters on a large dataset. Multimedia Tools and Applications. 80.1-19.10.1007/s11042-020-09718-4.

[48] Xu, Shuliang& Liu, Shenglan& Zhou, Jian & Feng, Lin. (2019). Fuzzy rough clustering for categorical data. International Journal of Machine Learning and Cybernetics. 10. 10.1007/s13042-019-01012-6.

[49] Qian ,C. , Mathur , N., Hidayati N.Z., Arora , R., Gupta ,V., Ali M.,(2022)Understanding public opinions on social media for financial sentiment analysis using AI-based techniques,Information Processing &Management,Volume 59, Issue 6,103098,ISSN 0306-4573

[50] Srijiranon K., Lertratanakham Y., TanantongT.(2022) A Hybrid Framework Using PCA, EMD and LSTM Methods for Stock Market Price Prediction with Sentiment Analysis. Applied Sciences.; 12(21):10823. https://doi.org/10.3390/app122110823

[51] Parekh R. et al., (2022)"DL-GuesS: Deep Learning and Sentiment Analysis-Based Cryptocurrency Price Prediction," in IEEE Access, vol. 10, pp. 35398-35409, 2022, doi: 10.1109/ACCESS.3163305.