ISSN : 2455-9679
SJIF Impact Factor : 6.008

IJRTSM

INTERNATIONAL JOURNAL OF RECENT TECHNOLOGY SCIENCE & MANAGEMENT

"HYBRID CLASSIFICATION METHOD FOR SENTIMENT ANALYSIS OF TWITTER DATA"

Tejaswani Jachpure [1]. Sudhir Patidar [2], Kamlesh Patidar [3]

[1] Research Scholar, Department of Computer Science and Engineering, JIT, Borawan, Madhya Pradesh, India
[2] Assistant Professor, Department of Computer Science and Engineering, JIT, Borawan, Madhya Pradesh, India
[3] HOD, Department of Computer Science and Engineering, JIT, Borawan, Madhya Pradesh, India

ABSTRACT

A variety of social media data are analyzed to determine sentiments. The data is classified using ML techniques. This research used Machine Learning (ML) algorithms to analyze attitudes in real-time Twitter data. The feelings are examined through various stages, including pre-processing of data, extracting the properties, and categorizing the data. The Support Vector Machine (SVM) technique is used at first to analyze the attitudes. The KNN method is then used to categorize the data. The integrated model is tested on five different datasets. At the end, a voting classifier that integrates K-Nearest Neighbor, DT, and Support Vector Machine is created to analyze the attitudes. The established methodology is evaluated using a variety of metrics. The results showed that the developed technique outperformed SVM because it increases accuracy while reducing execution time.

Key Words: *Twitter, Sentiment Analysis, KNN, SVM, DT, Voting.*

I. INTRODUCTION

Sentiment analysis is now more widely used by many people with a variety of interests and inspirations thanks to a significant growth in social media usage over the past few years. Obtaining knowledge from those data is an endeavor of enormous importance and significance since people worldwide may have diverse ideas about several themes linked to politics, education [1], tourism, culture, commercial items, or topics of general interest. Knowing their feelings as indicated by their words on various platforms, in addition to data on the websites people frequent, their top priority for purchases, etc., became an essential component for gauging public opinion on a particular issue. These days, one of the more frequently employed sentiment analysis techniques entails categorizing a text's polarity. [2]. In terms of labeling or number of levels, the texts can be upbeat, downbeat, or neutral, but overall, it relates to the emotions of the text varying from a cheerful to a sad mood. The term "sentiment" designates a topic that is both subjective and objective, as well as a topic that is both practical and imagined, and blurs the line between a topic that is positive or negative. An analysis based on the propagation of rumors or gossip is known as sentimental analysis. Sentiment analysis is a method of analysis based on text analysis. Finding the subjectivity of a belief, the result of a review, or the sentiment of a tweet entails SA.

Three layers make up the workflow for sentiment analysis, which is used to categorize different sentiment analysis techniques [3]. This section summarizes numerous case studies carried out at all sentiment analysis levels as featured in Figure 1.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

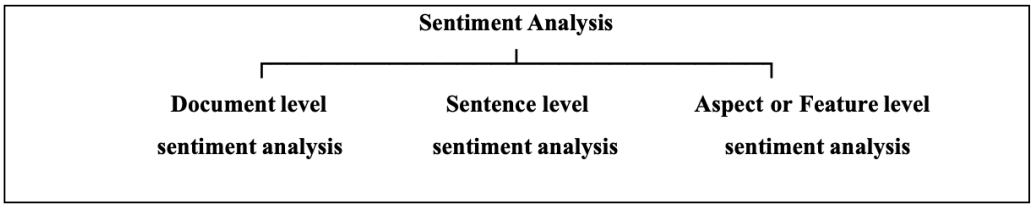

Fig. 1 Sentiment Analysis Levels

Routing indeed imposes additional strain on energy supplies, particularly in multi-hop schemes were nodes divided into Document level, Sentence level and aspect or feature level sentiment analysis.

Document-level: The goal of the document-level task is to identify whether a text's overall opinion expresses a range of emotions. For instance, the system determines whether a product review proclaims a wholly optimistic or pessimistic impression about the product once it has been reviewed. This goal is typically referred to as the method of document-level sentiment classification [4]. This level is predicated on the idea that every document expresses an opinion about a certain thing (like a product). Considering this, it is improper to document the evaluation or comparison of several things.

Sentence level: Finding out whether a sentence expresses a positive, negative, or neutral viewpoint is the goal of this level of analysis [5]. There are now two ways to perform sentence-level analysis. One strategy is to think of such an analysis as just a classification method with three stages and three different types of labels. Another strategy is to first examine subjectivity in the sentence by separating texts with opinions from texts without opinions, after which the texts are labeled with one of the two categories. This method has the drawback of linking every sentence to other parts of the text semantically and syntactically. Therefore, both local and global circumstantial knowledge is required for this task.

Aspect level: Both current levels use a set sentiment polarity that applies to the full document or sentence rather than just the subjects inside. This is inappropriate in several situations. A sub-function of sentiment analysis, aspect-level sentiment analysis aims to recognize and categorize sentiments at the aspect level [6]. Predicting the sentiment polarities for every explicit aspect word in each sentence is the goal of this level. Twitter is currently the most well-known social networking service for microblogging. As a means of information sharing, it provides its services internationally.

Therefore, it became very interesting to extract thoughts from tweets on numerous issues, assess the influence of certain events, or categorize attitudes.

To maintain and increase their impact and position, companies and people with large social media followings can now turn to Twitter, which has more than 413 million active users each month and more. These organizations can monitor numerous social media websites online thanks to sentiment analysis. Sentiment analysis is a technique for determining the polarity of a text and for automatically determining whether a section of it contains emotive or opinionated material [7]. The goal of Twitter sentiment categorization is to categorize a tweet's sentiment polarity as positive, negative, or neutral. There are several words in tweets are misspelled, non-dictionary, loud, asymmetrical, and incomplete sentences. A tweet has numerous distinctive characteristics that distinguish it from past research:

Length: A tweet can be as long as 140 words. The training set indicates that a tweet is typically 14 words or 78 characters long. This is very different from previous studies that aimed to categorize the larger volumes of work, such as movie reviews.

Data accessibility: The quantity of data already in existence is another difference. It is very simple to compile many tweets to train the algorithm thanks to the Twitter API. In earlier research, there were only thousands of training items in the assessments.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

Language model: Twitter users publish tweets using a variety of devices, including their smartphones. When compared to other fields, Tweet messages make considerably more errors and use slang.

Domain: Tweets on Twitter are brief messages. Unlike other sites, tweets can be posted on a variety of subjects that are tailored to a certain subject. This is distinct from a significant portion of earlier study, which focuses on certain fields, such as movie reviews [8].

II. Literature Review

Md. Rakibul Hasan, et.al (2019) suggested a pre-processed data model based on NLP (natural language processing) for filtering the tweets [9]. Thereafter, the motion of BoW (Bag of Words) was integrated with TF-IDF (Term Frequency-Inverse Document Frequency) to analyse the sentiments. These methods allowed for the accurate classification of tweets as favorable or negative. The accuracy to analyse the sentiments was maximized using a TF-IDF vectorizer. The outcomes of the simulation showed that the recommended approach worked well and had an accuracy of about 85.25 percent when evaluating attitudes.

ÖnderÇoban, et.al (2018) presented a text representation strategy based on W2VC (word2vec and clustering) created to evaluate Twitter sentiment [10]. The SVM algorithm (Support Vector Machine) was used to categorize the attitudes. The trials were carried out using two different types of datasets that contained Turkish Twitter feeds. The examination of the tests demonstrated the applicability of the introduced technique regarding time and performance. Moreover, this method was capable of mitigating the feature space.

Lei Wang, et.al (2019) projected an iterative algorithm recognized as SentiDiff for predicting the sentiment polarities which were expressed in Twitter messages [11]. This algorithm has taken account of inter-relationships of textual information of Twitter messages with the sentiment diffusion patterns which assisted in enhancing the Sentiment Analysis of Twitter. The experiment used a ground-truth dataset. Test findings experimentation validates, the supremacy of the projected algorithm over existing algorithms and provided the PR-AUC up to 8.38 % while analysis the sentiments on Twitter.

Paramita Ray, et.al (2017) used R software to construct a model to analyse the sentiment. The attitudes individuals had on Twitter data were examined using the Twitter API. [12]. In this model, the data was gathered from Twitter and pre-processed later. Afterward, the sentiment of the user was analysed using a lexicon-based technique. The acronym dictionary was created to replace the impactful word and the emoticons were detected in the tweet. The document level and the aspect levels were analysed to make the decision. The future work would aim to conduct a comparative analysis, make an attempt to utilize ML (machine learning) techniques, and construct a hybrid mechanism to analyse the sentiments.

Victoria Ikoro, et.al (2018) presented the outcomes attained while analysing the sentiments on Twitter that UK energy customers had posted [13]. The functions taken from two sentiment lexica were integrated with the optimization of accuracy of outcomes. The primary lexicon helped to extracting the sentiment-bearing terms and negative sentiments due to its efficacy in detecting these. The remaining data was classified using a second lexicon. The test outcomes appointed the presented approach led to enhanced accuracy of the outcomes based on the comparison with the common practice of using only one lexicon.

Rashid Kamal, et.al (2017) designed a model for visualizing the raw tweets with scalability and efficacy [14]. This model [15][16] focused on obtaining the sentiments of the people and visualizing them for better understanding. The tweets were taken in real-time using Spring XD. After that, raw tweets were converted into HDFS (Hadoop Distributed File System). The tweets were refined and labeled using HIVE (Hadoop Scripting Language) about their respective sentiments. In the end, HIVE was utilized to conduct a simulation of the designed model. This model was efficient for classifying the good, bad, and neutral emotions. The designed modelgenerated optimal outcomes while analyzing the sentiments.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

III. RESEARCH METHODOLOGY

The major social media platform for posting updates when a critical event is happening is Twitter. However, several issues related to false information occurred. Thus, the semantic texts must be categorized precisely for tackling such issues. This work focuses on deriving a dataset via Twitter for analyzing diverse versions of the Bayesian algorithm and quantifying its efficiency. When the tweets are extracted, preparing the data is another task. The online platforms consist of data that is available in far less format and this data is almost noisy. Thus, the main intent is to clean the data. The quality of data is represented in the results. The data is pre-processed in diverse stages. The first stage is executed to extricate the emoji, special characters, and any redundant information that has extra space. Consequently, the format is augmented and the imitated tweets are eliminated. Several features are obtained in the outcome. The sampling data is generated using the methods of extricating features. In addition, the applicability of this procedure is proved to evaluate the polarity in the phrase into two classes: optimistic and pessimistic to set up the objects using the duplicates. This Machine Learning approach helps in illustrating the crucial attributes of the material that are considered to distribute the data. The feature vectors are employed for quantifying the attributes that a classifier has deployed. Machine Learning is a robust scheme for classifying the data. The approach is trained for forecasting the indefinite data precisely. A voting mechanism is employed in which the K-nearest neighbor, support vector machine, and DT algorithm are integrated to analyze the sentiments. This mechanism is a hybrid of these algorithms whose training is done on a set of models. It assists in predicting a class on the principle of the extreme likelihood of the class whose selection is done as the output. This mechanism is used to combine the outcomes of each algorithm and employ them. The output class is predicted based on the notion of votes majority. The effective models are not developed and the accuracy of every model is not computed at the individual level in this work. The fundamental focus is on creating an individual framework in which these models have trained and forecasting the outputs for every output class depending upon their combined majority.

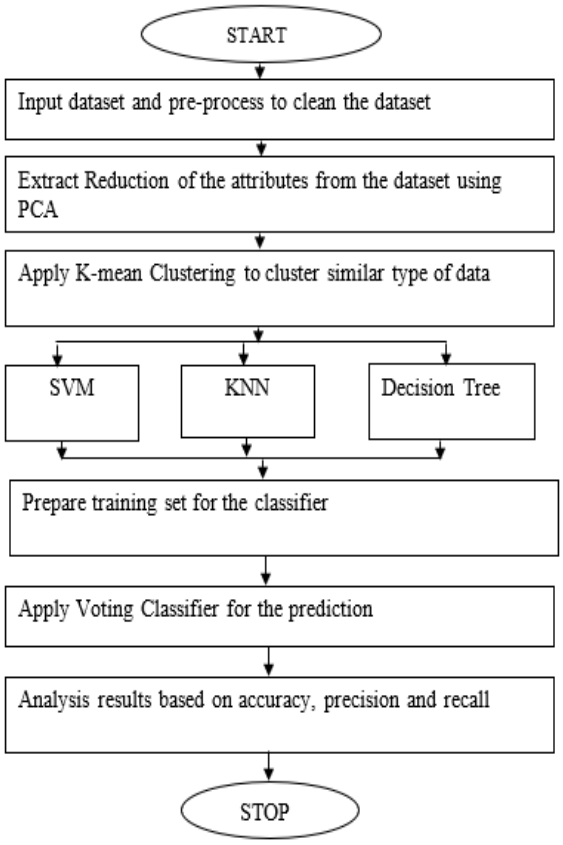

Fig. 2 Proposed Model

ISSN : 2455-9679
SJIF Impact Factor : 6.008

K-MEANS CLUSTERING

Given n data points X = $\{x_1, x_2 ,, x_n\}$ and k clusters with centroids C = $\{c_1, c_2 ,, c_k\}$, K-Means minimizes the following objective function:

$$J = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij} \backslash\backslash x_i - c_j \backslash\backslash^2$$

where:

- $w_{ij} = 1$ if x_i belongs to cluster i, otherwise 0.
- $\backslash\backslash x_i - c_j \backslash\backslash^2$ is the Euclidean distance between a data point and the cluster centroid.

CLASSIFIERS (SVM, KNN, DECISION TREE)

• Support Vector Machine (SVM):

The objective function for SVM is: $\frac{1}{2} \backslash\backslash w \backslash\backslash^2$

subject: $y_i(w^T x_i + b) \geq_1, \quad \forall_i$

where:

- *w is the weight vector,*
- *b is the bias,*
- y_i *is the class label for sample i.*

K-NEAREST NEIGHBORS (KNN):

A new data point x is classified based on the majority vote of its k nearest neighbors:

$$y = arg \sum_{i \epsilon N_k} \delta(y_i, c)$$

where $\delta(y_i, c)$ is 1 if $y_i = c$, otherwise 0.

Decision Tree:

Uses an impurity measure (e.g., Gini Index or Entropy) to split nodes. for entropy-based decision trees:

H(X) = $-\sum_{i=1}^{k} p_i p_i$

where p_i is the probability of class i.

PERFOMANCE METRICS

• Accuracy:

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

• Precision:

Precision= $\frac{TP}{TP + FP}$

• Recall:

Recall = $\frac{TP}{TP + FN}$

where:

• TP, TN, FP, FN represent True Positives, True Negatives, False Positive, and False Negatives.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

IV. Result and Discussion

Twitter functions as a micro blogging platform hosting a multitude of individual profile pages. These pages encompass the personal details of users. The platform permits users to effortlessly connect, fostering connections and facilitating the exchange of user-generated content. According to a conducted survey, the average daily tweet volume reaches an astonishing fifty billion. Each of these posts comprises myriad viewpoints and perspectives.
Various classification models exist for evaluating the precision of machine learning algorithms. Precision is a key metric in this context, focusing on the identified instances. For a specific class, accuracy refers to the proportion of correct outcomes, represented by TP (true positives), about the combined total of TP and FP (false positives) within the classification results.

Table. 1 Analysis of SVM Classifier's performance

Dataset	Precision		Recall	F1-Score	Accuracy
Dataset 1	0	83	80	82	81.38
	1	79	85	82	81.38
Dataset 2	0	94	73	82	87.73
	1	84	97	91	87.73
Dataset 3	0	82.68	68	94.60	88.10
	1	82	92	94.78	89.09
Dataset 4	0	89	40	55	88.08
	1	88	99	93	88.08
Dataset 5	0	87	92	89	86.38
	1	85	77	81	86.38

As detailed in Table 1, the evaluation of the SVM classifier's performance encompasses accuracy, precision, and recall measures. This assessment is conducted across five distinct datasets. Notably, the SVM classifier demonstrates an average accuracy of 85 percent in the context of sentiment analysis.

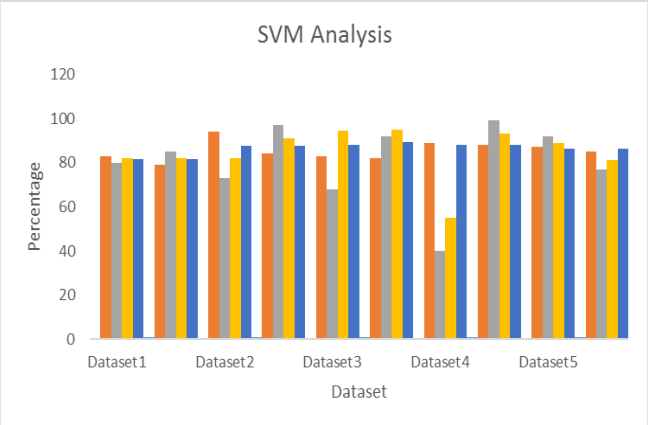

Fig 3 SVM Classifier Performance

Illustrated in Figure 3 is the utilization of the SVM classifier in assessing sentiment within Twitter data. Diverse performance metrics, including precision, recall, F1-measure, and accuracy, are employed to comprehensively evaluate the effectiveness of the SVM. Multiple datasets are employed to gauge the real-time performance of SVM classifiers.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

Table 2 Analysis of KNN Classifier's Performance

Dataset	Precision		Recall	F1-Score	Accuracy
Dataset1	0	65	72	68	65.43
	1	66	61	63	65.43
Dataset2	0	83	78	80	85.17
	1	86	90	88	85.17
Dataset3	0	82	66	93	87
	1	82	91	92	88
Dataset4	0	52	42	46	82.42
	1	88	92	89	82.42
Dataset5	0	81	85	83	77.93
	1	77	66	69	77.93

As delineated in Table 2, the performance evaluation of the KNN classifier is elaborated in relation to accuracy, precision, and recall metrics. This assessment encompasses the examination of the KNNclassifier's performance across five distinct datasets. Notably, the KNN classifier achieves an average accuracy of 75 percent when applied to sentiment analysis.

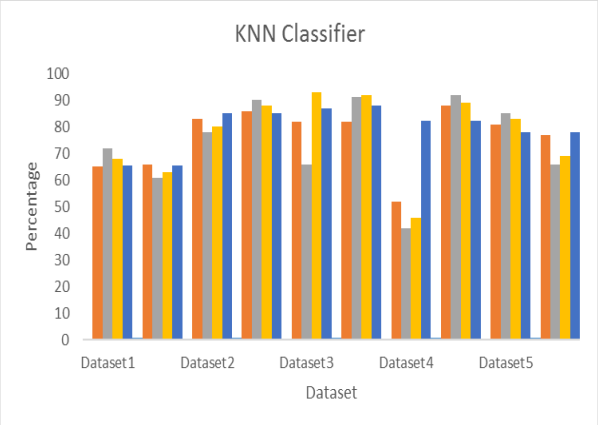

Fig. 4 KNN Classifier Performance

Displayed in Figure 4 is the practical application of the KNN classifier in conducting sentiment analysis on posted tweets. Diverse performance metrics, including precision, recall, F1-score, and accuracy, are utilized to thoroughly assess the efficacy of the KNN classifier. Multiple datasets are employed to evaluate the real-time performance of the KNN classifier.

Table 2 Analysis of Voting Classifier's performance

Dataset	Precision		Recall	F1-Score	Accuracy
Dataset1	0	86	91	88	90.69
	1	94	91	92	90.69
Dataset2	0	1	1	1	99.92
	1	1	1	1	99.92

ISSN : 2455-9679
SJIF Impact Factor : 6.008

Dataset3	0	85.68	90	95.67	98.12
	1	86	90	96.78	98.12
Dataset4	0	98	96	97	97.92
	1	98	99	98	97.92
Dataset5	0	92	97	94	94.62
	1	97	92	94	94.62

As elaborated in Table 3, the assessment of the Voting classifier's performance is detailed with regard to accuracy, precision, and recall metrics. This evaluation encompasses the analysis of the Voting classifier's performance across five distinct datasets. Notably, the Voting classifier achieves an average accuracy of 92 percent when applied to sentiment analysis.

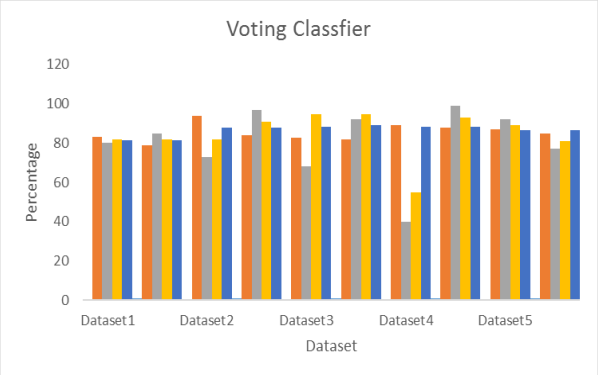

Fig 5 Voting Classifier Performance

Illustrated in Figure 5 is the application of the voting classifier in conducting sentiment analysis on posted tweets. Diverse performance metrics, including precision, recall, F1-score, and accuracy, are employed to comprehensively assess the effectiveness of the voting classifier. Multiple datasets are utilized to evaluate the real-time performance of the voting classifier.

V. CONCLUSION

The process of feature extraction establishes a crucial link between attributes and target variables. The final phase of classification involves the deployment of a classification framework capable of categorizing data into three distinct groups: positive, negative, and neutral sentiments. Although the current approach employs a hybrid classifier architecture to predict sentiment in posted tweets, it falls short in terms of achieving satisfactory accuracy and precision levels. This study introduces a novel voting classification methodology that combines Gaussian Naive Bayes and Random Forest (RF) classifiers with SVM classifiers to analyses sentiment within posted tweets. The primary objective of this project is to validate the performance of the developed architecture using four essential performance metrics: accuracy, recall, and precision. Following implementation on four datasets, the proposed architecture attains an impressive accuracy of 93.27%, surpassing the performance of the individual techniques typically employed for sentiment analysis based on tweets. Moreover, there remains untapped potential for enhancing the achieved accuracy. The proposed model could be extended in the future by integrating deep learning models, offering a promising avenue for further improvement in sentiment analysis accuracy. The voting classification approach already attains a noteworthy accuracy of up to 93%, suggesting that additional advancements are feasible. With the application of deep learning models, the potential for even higher accuracy levels becomes apparent.

ISSN : 2455-9679
SJIF Impact Factor : 6.008

REFERENCES

[1] M. Karqmibekr and A.A. ghorbani, "Sentiment Analysis of a Social Issues," International Conference on a Social Informatics, USA, pp. 215-221, 2012.
[2] M. Eirinaki, S. Pisal, J. Singh, "Feature-based opinion mining and ranking," Journal of Computer and System Sciences, vol. 78, issue 4, pp. 1175- 1184, 2012.
[3] E. Haddia, X. Liua and Y. Shib, "The Role of Text Pre-processing in Sentiment Analysis," Information Technology and Quantitative Management, vol. 17, pp. 26-32, 2013.
[4] K. Ghag and K. Shah, "Comparative Analysis of the Techniques for Sentiment Analysis," International Conference on Advances in Technology and Engineering (ICATE), pp. 1-7, 2013
[5] Q. Vuong and A. Takasu, "Transfer Learning for Emotional Polarity Classification," International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), vol. 2, pp. 94-101, 2014.
[6] V. Rout and A.D. Londhe, "Survey on Opinion Mining and Summarization of User Reviews on Web," International Journal of Computer Science and Information Technologies, vol. 5, pp. 1026-1030, 2014.
[7] Shahana P.H and Bini Omman, "Evaluation of Features on Sentimental Analysis," International Conference on Information and Communication Technologies, pp. 1585-1592, 2015.
[8] A. Tripathy, A. Agrawal and S. K. Rath, "Classification of Sentimental Reviews Using Machine Learning Techniques," International Conference on Recent Trends in Computing, pp. 821-829, 2015.
[9] Md. Rakibul Hasan, Maisha Maliha, M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data", 2019, International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2)
[10] ÖnderÇoban, GülşahTümüklüÖzyer, "Word2vec and Clustering based Twitter Sentiment Analysis", 2018, International Conference on Artificial Intelligence and Data Processing (IDAP)
[11] Lei Wang, JianweiNiu, Shui Yu, "SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis", 2019, IEEE Transactions on Knowledge and Data Engineering
[12] Paramita Ray, Amlan Chakrabarti, "Twitter sentiment analysis for product review using lexicon method", 2017, International Conference on Data Management, Analytics and Innovation (ICDMAI)
[13] Victoria Ikoro, Maria Sharmina, Khaleel Malik, Riza Batista-Navarro, "Analyzing Sentiments Expressed on Twitter by UK Energy Company Consumers", 2018, Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)
[14] Rashid Kamal, Munam Ali Shah, Asad Hanif, J Ahmad, "Real-time opinion mining of Twitter data using spring XD and Hadoop", 2017, 23rd International Conference on Automation and Computing (ICAC)
[15] R. Kamble and A. Nayak, "A hybrid deep learning improved method for share price prediction," *15th International Conference on Advances in Computing, Control, and Telecommunication Technologies (ACT 2024)*, 2024, vol. 1, pp. 693–701.
[16] A. Nayak and R. Kamble, "Artificial intelligence and machine learning techniques in power systems automation," in Artificial Intelligence Techniques in Power Systems Operations and Analysis, 2023, pp. 207–221.