Journal of Computers, Mechanical and Management
Journal homepage: www.jcmm.co.in
e-ISSN: 3009-075X
AN PUBLISHING

Volume 3 Issue 1

Article Number: 240105

Conceding Sentiment Prognosis on Twitter Data

Anshu Malhotra*[1] and Nishu Sethi[2]

[1]Department of Applied Sciences, The NorthCap University, Gurugram, India 122017
[2]Department of Computer Science and Engineering, School of Engineering, The NorthCap University, Gurugram, Haryana, India 122017

Abstract

Twitter is the biggest micro-blogging website that gives people a platform to share their opinions about any new happenings around the world. The size of tweets is generally short, which makes it very suitable for opinion mining. The key focus of the paper is to analyze the feelings and ideas. In this paper, analysis is done on the classification of tweets on a particular keyword. The tweets related to the given keyword are collected, analyzed, and the result is generated in the form of percentage of positive, neutral, and negative sentiments, which gives us a sense of the overall sentiment of the keyword. Further, Classification is done using supervised learning algorithms and the best among these will be found by calculating the accuracy of each.

Keywords: Sentiment Analysis, Logistic Regression, Support Vector Machines Opinion Mining, Supervised Machine Learning, Naïve Bayes, Decision Tress, Accuracy, Polarity Prediction.

1 Introduction

The classification of data is a continuously emerging topic among data scientists due to the continuously growing demand for classification. The main reason being that it allows analyzing the data globally, which can be used to derive a lot of information. The type of classification being discussed here is the classification of text. Text classification into already defined classes is also known as sentiment analysis [1], which recognizes the emotional aspect of the given content and assigns meaning to the sentiment, e.g., neutral, negative, and positive. Sentiment analysis [2] can be used in almost every aspect of the modern world, from products to services, e.g., online marketing, healthcare, social media. It can also be used in financial services or in political areas, and other possible domains where people leave their opinion. Organizations look to collect public or consumer opinions about their services and products. For that, opinion gathering methods such as surveys are conducted with the focus on targeted groups. So, modern solutions like sentiment analysis that work on topics such as classification using machine learning algorithms and work with collections of people's feedback or data expressed within short text messages, e.g., tweets, reviews of products prove to be very helpful. The results of this paper can be used in a variety of largescale data processing systems, finding the optimal information and their values to implement the algorithms, understanding and predicting the data to support decision making, and for knowledge gathering process. In this paper, Support Vector Machines, Naïve Bayes, Decision Tree, and Logistic Regression classifiers [3] are used for the classification task to get percentage-wise classification of sentiment on the keyword and to get the best classification accuracy using a number of tweets from Twitter. These methods are the most accurate and popular classification methods in the given domain of research.

*Corresponding author: anshumalhotra@ncuindia.edu
Received: 14 November 2023; **Revised:** 29 November 2023; **Accepted**: 04 December 2023; **Published:** 29 February 2024

DOI: 10.57159/gadl.jcmm.3.1.240105.

2 Methods

Sentiment Analysis is performed using Twitter APIs. With the help of these four keys, viz., Consumer Key, Consumer Secret, Access Token, and Access Token Secret, Twitter can be accessed. Tweedy is an open-source and easy-to-use library that allows access to the Twitter API [4]. The Tweets can now be fetched and are stored in a JSON format. A Labelled dataset is also downloaded. Natural Language Processing (NLP) is applied to both datasets to clean the data. With classification algorithms, tweets can be classified into positive, neutral, and negative polarity. Distinct classification algorithms like Support Vector Machine (SVM), Logistic Regression (LR), Naïve Bayes (NB), and Decision Trees (DT) [5] are utilized for categorization and further, the performance for each classifier can be measured in terms of accuracy. The design to build the proposed methodology is depicted in Figure 1.

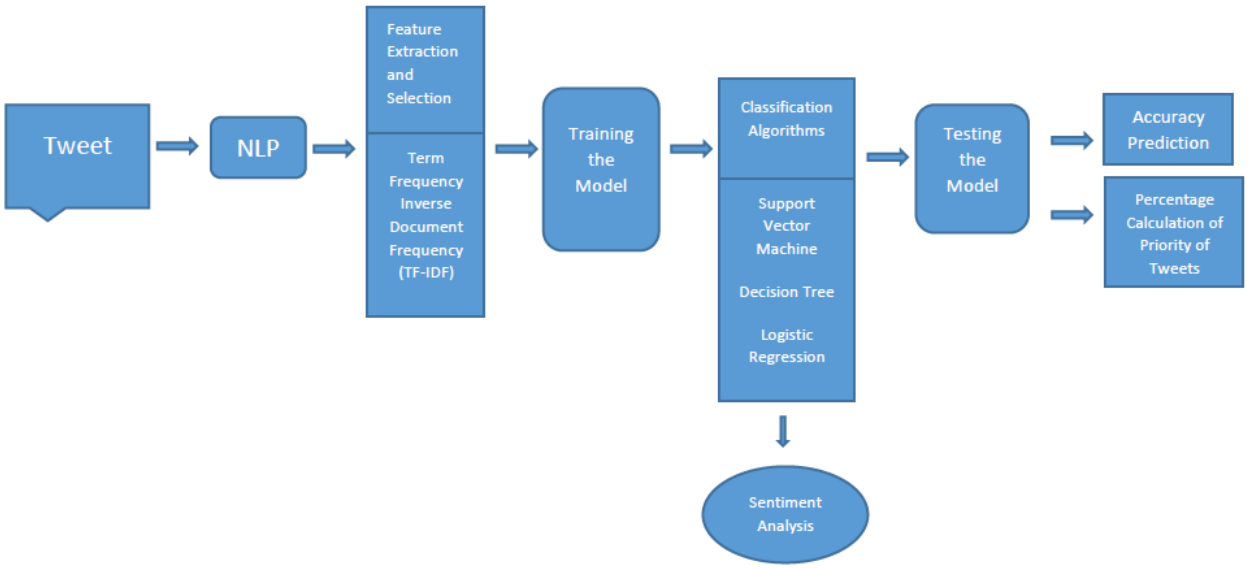

Figure 1: Workflow for the Methodology.

2.1 Extraction of Tweets

- **Getting Authentication Credentials**
  Authentication credentials such as Consumer Key, Consumer Secret, Access Token, and Access Token Secret can be generated by creating an application on Twitter. By visiting the Twitter Developer Website and providing all the necessary details, a Twitter Application can be created.

- **Authenticating Python Script**
  After establishing all the login credentials, that is APIs and Access Tokens, authentication can progress. To initialize authentication, some Python libraries need to be imported.

- **Defining Keywords**
  Keywords are the search words that search for tweets involving the specified keyword. Keywords are termed as input which is in the form of a string. A thousand tweets are fetched and stored in a JSON format.

2.2 Pre-processing the Tweets

Pre-processing is an important aspect of data mining. Pre-processing is a screening phase which means analyzing data such that the data should not result in any misleading results. This ensures quality and a good representation of data, which is extremely important to begin any analysis. In this project, Natural Language Processing (NLP) [6–8] is used by importing the Natural Language Toolkit (NLTK) library as the medium to screen and clean the data. Various techniques for processing natural language context are:

1. **Tokenization**: It is the practice of breaking a large set of contexts into smaller text or terms. Tokenization is of two types, namely Sentence Tokenization, which means breaking text into sentences, and Word Tokenization, which means splitting sentences into words.

2. **Removal of Stop Words**: Stop-words are English words that do not add meaning to a sentence and therefore can be disregarded. For example, 'have', 'the', 'are', 'was', etc.

3. **Lexicon Normalization**: It is the process of translating a non-standard text to a standard one. The goal is to eliminate various forms of words thus retaining a single and actual representation. For example, 'play', 'playing', 'playable', etc., can be reduced to "play" to retain the core meaning. Lexicon Normalization is of two types, namely:

- Stemming is the process of reducing the morphological variants of a root or base word. The aim is to remove affixes, i.e., prefixes and suffixes, for example, reducing the words “chocolates”, “chocolatey”, and “choco” to the root word “chocolate”.
- Lemmatization is the process of grouping different forms of a word so that they can be analyzed as a single item.

The workflow for pre-processing tweets is shown in Figure 2.

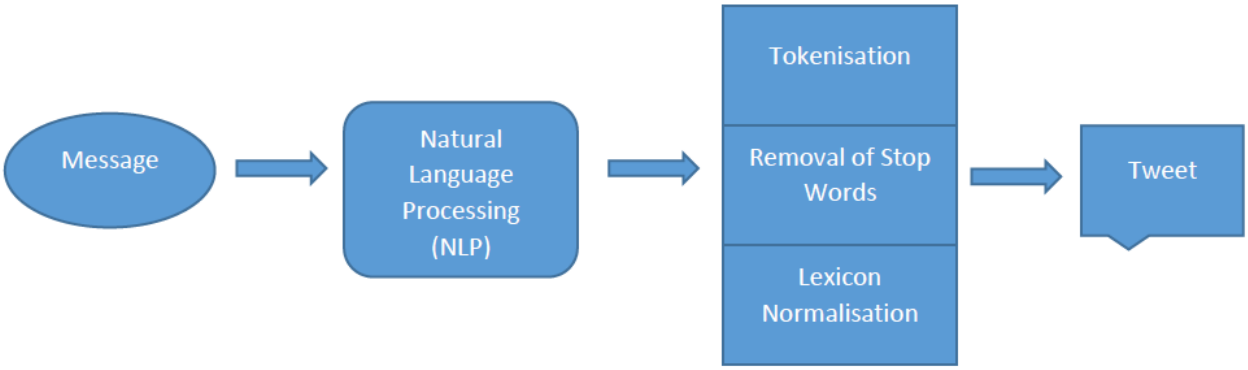

Figure 2: Workflow for Pre-processing Tweets.

2.3 Feature Extraction and Selection

Feature Extraction is the practice of selecting variables into features by efficiently reducing the amount of data that must be processed. Essentially, feature extraction involves extracting useful features from existing data. Conversely, Feature Selection refers to the process of filtering out irrelevant and unessential characteristics from the dataset, thereby choosing the most relevant attributes. In our project, the technique employed for extracting and selecting features is “Term Frequency–Inverse Document Frequency” (TF-IDF) [9, 10], a methodology that quantifies the importance of a term in a document. The weight assigned to each term indicates its significance within the corpus. This technique is widely utilized in the field of Knowledge Retrieval.

2.4 Training the Model with Machine Learning Algorithms

Supervised Learning Algorithm [11] is the machine learning task of instructing the model to learn a function that maps an input to an output based on example input-output pairs. It derives a function from a labeled training dataset that consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object (vector) and a desired outcome value (supervisory signal). This kind of learning algorithm analyzes the training data and produces an inferred function, which can be used for mapping new instances. An optimal model allows the algorithm to correctly predict the class labels for unseen instances, thus applying the learned function to new data in a ”reasonable” way. The process of mapping from input to output using the supervised learning algorithm is illustrated in Figure 3.

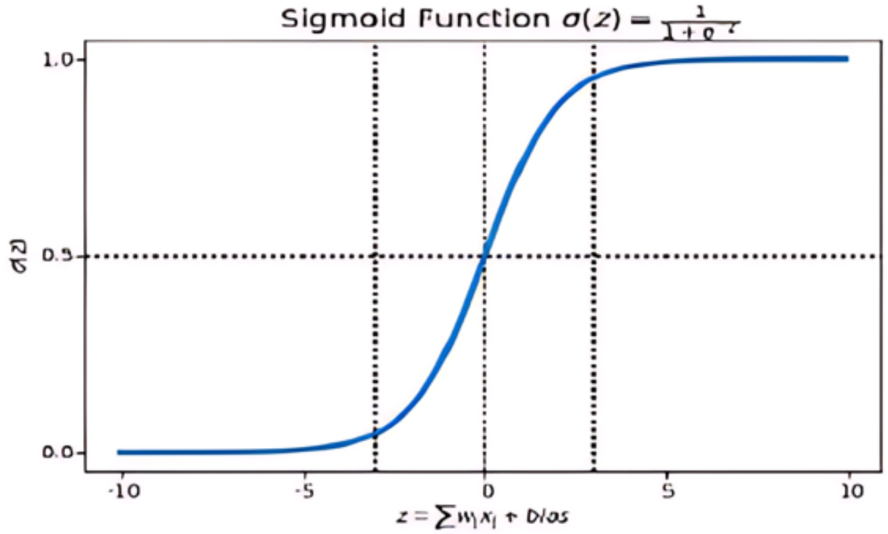

Figure 3: Illustration of the supervised learning process.

The various types of supervised learning algorithms include:

1. **Naïve Bayes (NB)**: A Classification Algorithm [12] part of a family of probabilistic algorithms based on Bayes' Theorem with a "naive" assumption of independence among each set of features. The theorem computes the probability $P(c|x)$, where c is the class of the possible outcome and x is the given instance to be categorized, with certain attributes. The formula used is as shown in Eq. (1)

$$P(c|x) = \frac{P(x|c) \cdot P(c)}{P(x)} \tag{1}$$

2. **Support Vector Machine (SVM)**: SVM [13] are Supervised Learning algorithms that analyze data for categorization. An SVM training algorithm builds a model that assigns new examples to one category or the other, making it a non-probabilistic binary linear classifier. The optimal hyperplane is identified such that it maximizes the margin between the two classes.

3. **Decision Trees (DT)**: The Decision Tree algorithm [14] is used in statistics, data mining, and machine learning as a predictive modeling tool. It moves from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

4. **Logistic Regression (LR)**: This learning algorithm [15, 16] is used for estimating the probability of a binary outcome. It is modeled by the logistic function, as represented by Eq. (2).

$$P(Y = 1|X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X)}} \tag{2}$$

2.5 Testing the Model

Testing is the phase where the model, trained in the training phase, is evaluated based on its performance. In the proposed system, the testing phase consists of two components: Accuracy Prediction and Percentage Calculation of Polarity of Tweets. Further, both Accuracy Prediction and Percentage Calculation of Polarity of Tweets are graphically depicted with the help of a Pie-Chart.

3 Results and Discussion

This section presents the details of the experiments conducted in this project along with the discussion of the outcomes. In addition to sentiment analysis, the polarity percentage is calculated [17], with the goal of determining the accuracy [18]. Various classifiers such as Decision Trees (DT), Support Vector Machines (SVM), Naïve Bayes (NB), and Logistic Regression (LR) are employed to analyze the accuracy of each classifier. The primary reason for using four different classifiers is to compare their accuracies. A total of 1000 tweets were collected from Twitter using Twitter APIs, and 48 tweets were analyzed from the downloaded dataset after pre-processing. Sentiments such as positive, neutral, and negative are validated using both datasets, namely the Downloaded and the Extracted Dataset. The following Table 1 and Figure 4 depict the polarity calculation in terms of the number of tweets with all four classifiers used on the Extracted Dataset.

Table 1: Polarity Calculation for the extracted data with different classifiers.

Classifier	Positive	Neutral	Negative
Naïve Bayes (NB)	176	225	599
Support Vector (SVM)	123	5	872
Decision Tree (DT)	23	147	830
Logistic (LR)	188	14	798

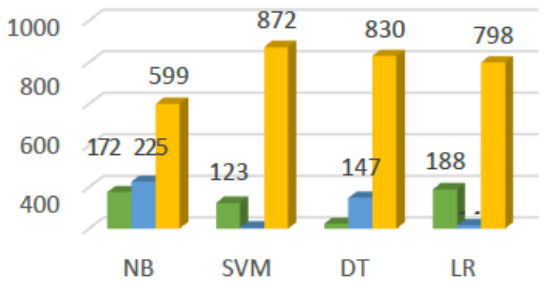

Figure 4: Graphical Representation of Tweets from Extracted Dataset.

The polarity calculation for the downloaded dataset is presented in Table 2 and the graphical representation is shown in Figure 5. This dataset was used to validate the performance of the different classifiers on a pre-existing, processed collection of data. However, after pre-processing and training the model, it was established that Logistic regression has the maximum accuracy. The following table shows the accuracy in terms of percentage of each classifier. Also, the figure depicts the graphical representation of the same. As a result, Logistic Regression shows highest accuracy.

Table 2: Polarity Calculation for the Downloaded Dataset with different classifiers..

Classifier	Positive	Neutral	Negative
Naïve Bayes (NB)	8	19	21
Support Vector Machine (SVM)	12	12	24
Decision Trees (DT)	9	17	22
Logistic Regression (LR)	13	11	24

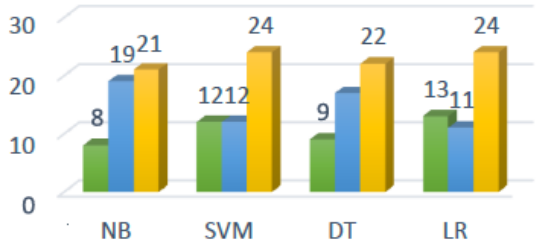

Figure 5: Graphical Representation of Tweets from Downloaded Dataset.

The analysis of the downloaded dataset allows for a direct comparison between the various machine learning algorithms in terms of their ability to classify sentiment accurately. It can be observed from Table 2 and Figure 5 that different classifiers exhibit varying levels of performance with respect to the sentiment classification task.

The accuracy of each classifier is quantified and presented in Table 3, while Figure 6 provides a visual representation of these accuracies.

Table 3: Accuracy Matrix of the classifiers.

Classifier	Accuracy (%)
Naïve Bayes (NB)	66.666
Support Vector Machine (SVM)	64.583
Decision Tree (DT)	39.583
Logistic Regression (LR)	69.416

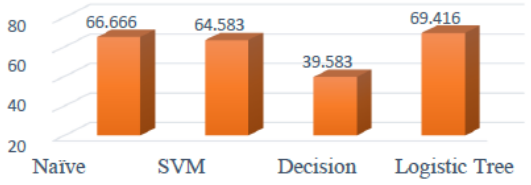

Figure 6: Graphical Representation of Accuracy Prediction.

The analysis of the accuracy of the classifiers is critical to understanding their performance. As shown in Table 3 and Figure 6, the Logistic Regression (LR) classifier achieves the highest accuracy, while the Decision Tree (DT) classifier has the lowest accuracy among the four classifiers. These results are significant in the context of sentiment analysis where the precision of prediction is paramount. The calculation of the percentage polarity of Tweets is visualized using graphical representation [19]. A Pie-Chart is drawn to depict the proportion of polarity (Negative, Positive, and Neutral) for both datasets. The accuracy with the count, i.e., the number of tweets, is also displayed in Figure 7. The graphical representation for one of the classifiers could not be shown below. The first pie-chart represents the analyzed Tweets from the Downloaded Dataset and the second pie-chart represents analyzed Tweets from the Extracted Dataset.

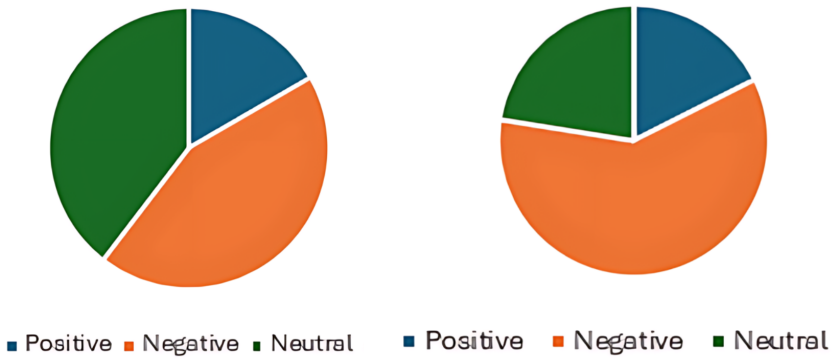

Figure 7: Graphical representation of Naïve Bayes.

4 Conclusions

In this paper, the sentiments of people on a particular keyword are analyzed in terms of percentage for negative, positive, and neutral sentiments, as well as the adoption of various machine learning algorithms for the classification of sentiments to determine the approach with the maximum accuracy. The data used in this study is a collection of tweets from the social media platform Twitter. Various classification models have been tested to reflect the best model in terms of accuracy. Further improvements to the models can be achieved by exploring different methodologies and expanding domain knowledge, including the implementation of emoticon analysis, understanding order dependence, and detecting sarcasm. Despite the challenges and complexities inherent in opinion analysis, its value to business and decision-making cannot be overlooked. Sentiment prediction relies on human-like aspects, suggesting that it will become a major influence in industry decisions in the foreseeable future. Enhancing the precision and reliability of extraction methods could help address some of the current challenges in sentiment prediction. Looking forward, a more refined and accurate representation of public opinion is envisioned, established through Sentiment Prediction. This vision includes a societal construct where every opinion is considered, and each sentiment contributes to decision-making processes, transcending the need to rely solely on a few 'experts'. It is towards this future that sentiment analysis endeavors to contribute significantly.

Declaration of Competing Interests

The authors declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding Declaration

This research did not receive any grants from governmental, private, or nonprofit funding bodies.

Author Contribution

Anshu Malhotra: Supervision, Project Administration, Writing - Review & Editing; **Nishu Sethi** Conceptualization, Methodology, Software, Writing - Original Draft

References

[1] L. Rajput and S. Gupta, "Sentiment analysis using latent dirichlet allocation for aspect term extraction," *Journal of Computers, Mechanical and Management*, vol. 1, no. 2, pp. 30–35, 2022.

[2] M. V. Mäntylä, D. Graziotin, and M. Kuutila, "The evolution of sentiment analysis—a review of research topics, venues, and top cited papers," *Computer Science Review*, vol. 27, pp. 16–32, 2018.

[3] M. Trupthi, S. Pabboju, and G. Narasimha, "Sentiment analysis on twitter using streaming api," in *Proceeding of IEEE 7th International Advance Computing Conference*, pp. 915–919, IEEE, 2017.

[4] G. Santafe, I. Inza, and J. A. Lozano, "Dealing with the evaluation of supervised classification algorithms," *Artificial Intelligence Review*, vol. 44, pp. 467–508, 2015.

[5] B. Wagh, J. Shinde, and P. Kale, "A twitter sentiment analysis using nltk and machine learning techniques," *International Journal of Emerging Research in Management and Technology*, vol. 6, no. 12, pp. 37–44, 2018.

[6] P. M. Nadkarni, L. Ohno-Machado, and W. W. Chapman, "Natural language processing: an introduction," *Journal of the American Medical Informatics Association*, vol. 18, no. 5, pp. 544–551, 2011.

[7] R. Egger and E. Gokce, "Natural language processing (nlp): An introduction: Making sense of textual data," in *Applied data science in tourism: Interdisciplinary approaches, methodologies, and applications*, pp. 307–334, Springer, 2022.

[8] J. Kaur, P. Verma, and S. Bajoria, "Sashakt: a job portal for women using text extraction and text summarization," *Journal of Computers, Mechanical and Management*, vol. 1, no. 2, pp. 22–29, 2022.

[9] F. Osisanwo, J. Akinsola, O. Awodele, J. Hinmikaiye, O. Olakanmi, J. Akinjobi, *et al.*, "Supervised machine learning algorithms: classification and comparison," *International Journal of Computer Trends and Technology*, vol. 48, no. 3, pp. 128–138, 2017.

[10] P. Kaviani and S. Dhotre, "Short survey on naive bayes algorithm," *International Journal of Advance Engineering and Research Development*, vol. 4, no. 11, pp. 607–611, 2017.

[11] Y. Yang, J. Li, and Y. Yang, "The research of the fast svm classifier method," in *Proceedings of 12th International Computer Conference on Wavelet Active Media Technology and Information Processing*, pp. 121–124, IEEE, 2015.

[12] N. Bhateja, N. Sethi, and S. Kaushal, "Machine learning and its role in diverse business systems," *Research Journal of Science and Technology*, vol. 13, no. 3, pp. 213–217, 2021.

[13] N. Sethi, N. Bhateja, N. Sethi, and S. Sinha, "Recognizing sentiment prediction on twitter data," *International Journal of Innovative Research in Computer Science & Technology*, pp. 2347–5552, 2020.

[14] L. Rokach and O. Maimon, "Decision trees," *Data Mining and Knowledge Discovery Handbook*, pp. 165–192, 2005.

[15] L. Liu, "Research on logistic regression algorithm of breast cancer diagnose data by machine learning," in *2018 International Conference on Robots & Intelligent System (ICRIS)*, pp. 157–160, IEEE, 2018.

[16] S. Deepa, "Metaheuristics for multi criteria test case prioritization for regression testing," *Journal of Computers, Mechanical and Management*, vol. 1, no. 1, pp. 42–51, 2022.

[17] I. Bala and A. Yadav, "Comprehensive learning gravitational search algorithm for global optimization of multimodal functions," *Neural Computing and Applications*, vol. 32, pp. 7347–7382, 2020.

[18] S. Jain, D. C. Bisht, and P. C. Mathpal, "Particle swarm optimised fuzzy method for prediction of water table elevation fluctuation," *International Journal of Data Analysis Techniques and Strategies*, vol. 10, no. 2, pp. 99–110, 2018.

[19] A. Hasan, S. Moin, A. Karim, and S. Shamshirband, "Machine learning-based sentiment analysis for twitter accounts," *Mathematical and Computational Applications*, vol. 23, no. 1, p. 11, 2018.