Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Optimized Ensemble Model for Sentiment Analysis incorporating Grey Wolf and Genetic Algorithm with Voting Classifier

Gaurav Pandey[1*], Narendra Kumar Gupta[2]
[1] *Ph.D. Scholar Department of Computer Science & Information Technology, SHUATS, Uttar Pradesh, India*
[2] *Assistant Professor Department of Computer Science & Information Technology, SHUATS, Uttar Pradesh, India*
Emails: gauravpandey161989@gmail.com[1], narendra.gupta@shuats.edu.in[2]

ARTICLE INFO

Received: 30 Dec 2024

Revised: 16 Feb 2025

Accepted: 25 Feb 2025

ABSTRACT

The sentiment analysis is the opinion mining approach which use NLP for the categorization. The sentiment analysis can be performed on various social media data. The machine learning is the approach which are applied for the classification. In the previous research work, tokenization technique is applied for the sentiment analysis which is less efficient as compared to machine learning.

In this research work machine learning algorithms are applied for the sentiment analysis on live twitter data. Diverse stages are executed to analyze the sentiments in which the data is pre-processed, attributes are extracted and the data is classified. The hybrid optimization algorithm is proposed in this research work for the feature extraction.

The hybrid optimization algorithm is the combination of genetic and gray wolf algorithm. The voting classification model is the used for the classification which is the combination of KNN, SVM and decision tree. The performance of proposed model is tested in terms of accuracy, precision and recall. It is analysed that proposed model achieves 98.89 percent accuracy for the sentiment analysis.

Keywords: Sentiment Analysis; Hybrid Optimization; Genetic Algorithm; Voting Classification; Gray Wolf.

INTRODUCTION

Social media platforms like Twitter serve as forums where users share thoughts, opinions, and engage with topics through short posts known as tweets, which can include text, images, and videos. Users interact via features like likes, comments, and retweets. According to Twitter, the platform boasted over 206 million daily active users in 2022, defined as logged accounts where ads can be displayed. As social media participation grows, analysing online information offers insights into changes in people's perceptions, behaviours [1][2], and psychology. Consequently, using Twitter data for sentiment analysis has gained popularity. This rising interest in social media analysis has underscored the importance of Natural Language Processing (NLP) and Artificial Intelligence (AI) technologies in text analysis, facilitating the assessment of sentiments and attitudes within specific target groups. Text analysis enables the identification of sentiments and attitudes within specific target groups.

While much of the existing literature predominantly examines texts in English, there is a growing interest in multilingual analysis. This form of analysis involves extracting subjective comments related to various topics and categorizing them into sentiments such as Positive, Negative, and Neutral [3][4]. Collecting and analysing tweets from diverse users on specific topics is a common practice for many companies and organizations. This specialized analysis is known as “sentiment analysis” which involves computationally identifying people's opinions about particular subjects from their social media posts. Sentiment analysis, especially on platforms like Twitter, has been a

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

highly researched area for years. The internet's expansion and the rapid growth of online content, coupled with widespread social media use, have made sentiment analysis crucial for companies and organizations. This surge is driven by the vast volume of user-generated data that includes personal opinions on various subjects and topics [5][6]. Companies, in particular, find value in performing sentiment analysis on such data to gauge the opinions of their average customers. This approach is effective because platforms like e-commerce stores or dedicated movie review sites tend to attract feedback from individuals who are either highly satisfied or dissatisfied, making the data polarized and insightful for business decisions.

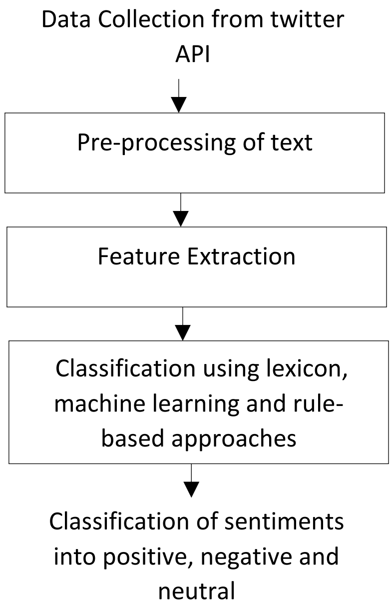

Figure 1. Sentiment Analysis/Opinion Mining from twitter data

Collecting labelled datasets is one of the crucial challenges in Twitter's sentiment analysis. The Twitter API is used to collect a collection of text posts. Then, these tweet posts are combined together to construct a dataset of three classes [9]: positive emotions, negative emotions, and a group of objective texts (no emotion). A significant drawback of using datasets from Twitter is the inherent noise in the data. Tweets can consist of plain text, user mentions (@user), and references to URLs or hashtags (#). In this step, the Twitter data undergoes pre-processing to prepare it for feature extraction and classification. Several steps are taken to remove noise from the dataset, such as eliminating retweets, duplicates, special characters [7][8], tweets in other languages, and tweets containing only a URL. These noisy elements do not enhance data classification accuracy and are thus removed. In the pre-processing step of sentiment or opinion analysis, various natural language processing (NLP) techniques can be utilized, including sentence splitting, tokenization, Part-of-speech (POS) tagging, stop word removal, stemming, and lemmatization. However, the applicability of these techniques may be limited in lower-density languages due to the requirement of pre-trained language models and linguistic resources [9][10]. In cases where linguistic resources are unavailable, researchers often rely on pre-trained models in high-density languages like English and employ translation or parallel corpora to perform cross-language sentiment classification. The Twitter language model includes several distinctive features designed to reduce the feature space. Initially, the process begins by extracting all unigrams and bigrams from the corpus that exceed a specified frequency threshold. For instance, all unigrams and bigrams occurring more than 5 times are selected as candidate features. Typically, unigrams and bigrams are utilized in sentiment analysis at the word or phrase level. Additionally, the model can be extended to include trigrams. Subsequently, for each tweet, the frequency of each identified candidate feature within it is computed. Researchers in sentiment analysis typically operate under the assumption that each document expresses a consistent sentiment polarity from a single source towards an object [11][12].

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/
Research Article

This assumption holds true in cases like product reviews where opinions are clear and singular. Given the brevity of tweets, this assumption also applies effectively to tweet data, as users generally convey straightforward sentiments due to the platform's concise nature. The effectiveness of pre-processing techniques on sentiment classification is evaluated using one of three strategies: lexicon-based approaches, machine learning methods, or rule-based approaches. Lexicon-based approach employs pre-existing dictionaries containing tagged words. The text is first tokenized to break it down into individual tokens. These tokens are then matched against entries in the dictionary. If a token matches a positive entry, its score is added to the cumulative score for the entire text. For instance, if the word "dramatic" is positively tagged in the dictionary, it contributes positively to the overall sentiment score. Conversely, if it matches a negative entry, the score is either reduced or the word is tagged as negative. Despite its straightforward nature, this approach can yield effective results. Lexicon-based classification algorithms determine the sentiment of a document based on key components such as words or phrases. Common strategies include majority voting, document scoring with thresholds, and simple word counts [13][14]. Within lexicon-based approaches, two primary categories exist: dictionary-based approach and corpus-based approaches. Dictionary-based approaches use lexicon dictionaries to identify positive and negative opinion words. They may utilize resources like WordNet to expand initial sets of words. However, words identified by these approaches tend to be contextually and domain-independent. On the other hand, Corpus-based techniques leverage large datasets (corpora) to identify opinion words based on their syntactic patterns within specific contexts. By doing so, they address challenges associated with identifying contextually nuanced opinion words.

Machine learning (ML) is a subset of artificial intelligence (AI) focused on algorithms that enable computers to learn from data. In essence, an algorithm is presented with a dataset and derives insights about its properties, which in turn allows it to predict future data encounters. Most non-random data inherently contain patterns that aid machines in generalizing from the observed information. To achieve this generalization, models are trained to discern significant features within the data. Various machine learning algorithms, whether supervised or unsupervised, are applied in sentiment analysis to extract meaningful insights from both structured and unstructured textual data, aiding decision-making processes. Supervised algorithms, known for their efficiency in determining emotional polarity, require substantial amounts of labelled data, which can be challenging to acquire. In contrast, unsupervised algorithms, while less precise, offer utility by operating without labelled data [15][16].

Commonly utilized machine learning algorithms for sentiment analysis include Naive Bayes, Support Vector Machines, and Random Forests, among others. In a rule-based classifier, the data space is structured using a set of rules. Each rule consists of a condition on the feature set, typically expressed in normal form on the left, and a corresponding class label on the right. The rules often focus on the presence of specific words, as the absence of terms is less informative, especially in sparse data scenarios. During the training phase, rules are generated based on various criteria. Two commonly used criteria are support and confidence. Support refers to the absolute number of examples in the training dataset that satisfy a rule, while confidence represents the conditional probability that the class label is correct given that the conditions of the rule are met. VADER is a prominent example of a rule-based model that incorporates multiple lexical features. It is designed for sentiment analysis in micro-blog data, such as tweets, and is known for providing robust and comprehensive results compared to other conventional models.

The sentiment analysis is the key area of research is machine learning. The various authors proposed various solutions to address issue of sentiment analysis effectively and efficiently. The authors follow various steps for the sentiment analysis which include pre-processing, feature extraction and classification. It is analysed from the previous studies that authors are unable to achieve good accuracy due to various reasons which include tokenization issue, feature extraction issue and due to misclassification. In some studies, it is analysed that authors are unable to tokenize data efficiently which directly affect classification results. When the features are not extracted effectively it is difficult to establish relation between attribute set and classes. In this research work feature extraction issue is addressed to improve accuracy for the classification. In the second section of the paper literature survey is presented which address key finding of various authors for the sentiment analysis. In the third section of the paper proposed methodology is presented along with flowchart and pseudo codes. In the fourth section of the paper results of proposed model is presented and also discussed while comparing with existing models. In the last conclusion of the study is presented with key finding and future possibilities are addressed.

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

LITERATURE SURVEY

L. Wang et al. (2020) initially examined sentiment diffusion by exploring a phenomenon named sentiment reversal, discovering fascinating characteristics of such reversals [17]. To forecast sentiment polarities in Twitter messages, they introduced a continuous approach called SentiDiff, examining inter-relationships among textual information in Twitter messages. This pioneering work was the first to use sentiment diffusion patterns to enhance sentiment analysis on Twitter, a widely acknowledged contribution. Extensive experiments on real-world datasets illustrated that the proposed algorithm achieved improvements of about 5.09% and 8.38% in PR-AUC compared to the latest sentiment analysis algorithms based on Twitter's textual information for sentiment classification tasks.

A. Mahmood et al. (2020) presented a statistical model for detecting users and social bots who disseminate distorted content via tweets on Twitter [18]. They utilized an interpreted Twitter dataset for their experiments and examined the impacts of biased users at both micro and macro levels, comparing sentiment analysis with and without biased tweets. The results demonstrated that their algorithm effectively distinguishes biased users and bots from authentic users, proving to be highly productive.

A. Feizollah et al. (2019) described the evaluation of Twitter sentiment on specific topics, focusing on two halal products: halal tourism and halal cosmetics [19]. They utilized the Twitter search function to gather data spanning a decade, followed by a preprocessing approach to clean the data. Their experiments employed deep learning algorithms for sentiment analysis of tweets, including convolutional neural networks (CNN), long short-term memory (LSTM), and recurrent neural networks (RNN). The LSTM approach combined with Word2vec feature extraction achieved the highest accuracy of 93.78%, demonstrating superior performance among the results.

M. Bibi et al. (2020) investigated the potential of hierarchical clustering for Twitter sentiment analysis [20]. They analyzed three hierarchical clustering strategies: single linkage (SL), complete linkage (CL), and average linkage (AL). They developed a collaborative structure of SL, CL, and AL to select the best clusters for tweets using majority voting. They compared hierarchical clustering methods with K-means and two classifiers (SVM and Naïve Bayes) in terms of accuracy and time efficiency. Experimental results showed that collaborative clustering provided robust clustering with a trade-off of time efficiency.

M. Bibi et al. (2019) proposed an imputation algorithm (CAARIA) based on attribute relevancy to reduce the size of Twitter data, introducing a preprocessing method named class association [21]. CAARIA achieved its goal of dimensionality reduction by grouping tweets associated with the same class, while retaining meaningful data. They compared the performance of two classifiers (Naïve Bayes and support vector machines) using three Twitter datasets, contrasting CAARIA with two widely used dimensionality reduction strategies: information gain (IG) and Pearson's correlation (PC). CAARIA outperformed IG and PC in terms of classification accuracy and time efficiency, highlighting its robustness as a data preprocessing technique for classification tasks.

S. E. Saad et al. (2019) aimed to perform detailed sentiment analysis of tweets using machine learning methodologies, focusing on ordinal regression [22]. They proposed an efficient feature extraction method and conducted preprocessing of tweets. The study employed Multinomial logistic regression (Soft Max), Support Vector Regression (SVR), Decision Trees (DTs), and Random Forest (RF) algorithms for sentiment analysis categorization. Experimental results demonstrated that their model achieved accurate ordinal regression, with Decision Trees outperforming other approaches.

L. Chaudhary et al. (2024) provided insights into quality preprocessing methods and various word embeddings for data preparation, outlining a taxonomy of DL-based algorithms [23]. They emphasized computational metrics for performance evaluation, availability of resources in the public domain for sentiment analysis tasks, and experimental applications in domain-specific contexts. The study concluded with future research directions and highlighted several research issues for further investigation.

A. Sanjay et al. (2021) collected Twitter data on farmers' protests to understand global sentiment, analyzing around 20,000 tweets on the topic [24]. They classified sentiments using Bag of Words and TF-IDF methods, finding Bag of Words to perform better than TF-IDF. They evaluated classifiers including Naive Bayes, Decision Trees, Random Forests, and Support Vector Machines, with Random Forest achieving the highest accuracy. Their study revealed that

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

29% of users expressed positive sentiments favoring farmers, while 17% used negative language in discussions about the protests.

V. Mahalakshmi et al. (2024) introduced a conditional generative adversarial network (GAN) for Twitter sentiment analysis, utilizing a convolutional neural network (CNN) for feature extraction from Twitter data [25]. Their research demonstrated superior performance in accuracy, recall, clarity, and F1 score compared to existing studies, achieving a categorization accuracy of 93.33%. The integration of LSTM in the GAN discriminator and CNN in the GAN generator was a key innovation. They highlighted the importance of setting hyperparameters, such as learning rate, to optimize sentiment analysis models based on neural networks.

S. Jabalameli et al. (2022) investigated geographic distribution of topics, recurring themes in discussions, and public sentiments during the pandemic using Twitter data [26]. They employed Natural Language Processing with the LDA method to identify 11 topics and 8 sub-topics in Twitter data. Temporal analysis of topics revealed the sensitivity of online discourse to major state news and local government responses to the pandemic. The study's findings can inform public sentiment analysis, tracking public demands, responses to local authorities' policies, and future pandemic management strategies.

METHODS

The Sentiment classification models will help us to identify type of Sentiments. The sentiment classification models have various steps which include data set pre-processing, feature extraction, classification and performance analysis. The various schemes are proposed in the past years for the efficient sentiment classification. The existing schemes has various drawbacks which we need to entertain in the research work. The dataset is very large in size due to which existing schemes are unable to establish relation of each attribute with target set. In this research work, the novel scheme will be proposed which can extract features of the dataset for the efficient classification. The hybrid classification models will be proposed which will improve performance for the sentiment classification. The motivation of this research work is to increase accuracy and methodology is described below: -

3.1. Data set input and Pre-processing

The initial stage is the dataset input in which data gathered from the genuine source named twitter is utilized for input. The dataset of twitter is collected using the twippy API. The dataset is clean and donot have any missing or redundant values. The dataset is divided into training and test which has 60 percent is utilized as training and rest as test set.

3.2. Feature Extraction

The feature extraction is the important phases in which relationship between attribute set and target set is established. The hybrid optimization algorithm is the combination of genetic and PSO algorithm. The proposed flowchart is the hybrid version of Genetic and PSO algorithm. This algorithm is useful to select the optimization attributes and encoding an effective solution for an issue into an individual. In fact, every individual is considered as an entity supporting features of chromosomes. A number of individuals collectively creates a population. The major task is to generate a population of chromosomes randomly and surround it with variables of problem prior to deploy Genetic Algorithm (GA). The next phase emphasizes on assessing the created data chromosomes. The chromosomes, which are capable of clearly demonstrating an optimal method for tackling the issue, are useful for building other chromosomes. The population is defined as the primary set of random solutions available in this algorithm. A chromosome is utilized for illustrating every member of the population in order to perform coding for a solution for dealing with the issue. The decoding formula is expressed as:

$$X = X_{min} + \frac{X_{max} - X_{min}}{2^{Nx} - 1} \sum_{n=0}^{N^x - 1} b_n^X 2^n \tag{1}$$

In which, $b_0^X, \ldots\ldots, b_{N^x-1}^X$ denote the binary representations of X's. Various iterations called generations are exploited for creating the chromosomes. In every generation, a number of fitness indicators are executed for evaluating the fitness value of the chromosomes. The pack is consisted of 4 kinds of wolves and the ranks are assigned to them from highest to lowest in the social hierarchy such as α, β, δ, and ω wolf. This algorithm is also depending upon the social hierarchy of grey wolves and their hunting behaviour. The particular mathematical models of Grey Wolf Optimization discussed below. While hunting, for surrounding the prey, computing the distance amid the current grey wolf and the prey and updating its position later on in accordance with the distance, is an essential task. The behaviour of grey wolves in hunting is expressed as:

$$X(t+1) = X_P(t) - A \times D \tag{2}$$

$$D = |C \times X_P(t) - X(t)| \tag{3}$$

The equ. 14 denotes the updating formula of the position of grey wolf, and equ. 15 defines the calculating formula. t denotes the current iteration number, $X_P(t)$ is used to represent the current position vectors of the prey and $X(t)$ illustrates the grey wolf at iteration t. The coefficient vector (CV) obtained in first formula is denoted with A, and CV of second is depicted with C [12].

$$A = 2 \times a \times r_1 - a \tag{4}$$

$$C = 2 \times r_2 \tag{5}$$

And

$$a = 2 - 2 \times \frac{t}{t_{max}} \tag{6}$$

In this, a illustrates the convergence factor, and its value is linearly diminished from 2 to 0 due to the maximization of number of iterations. r_1 and r_2 signify the random vectors in [0, 1]. Moreover, t_{max} is the maximum number of iterations.

3.2.1 Hunting

The hunting behaviour of grey wolves is simulated by assuming that α, β, and δ wolves can understand the potential location of prey more effectively. α wolf illustrates the optimal solution, β wolf is used for the suboptimal solution, and δ wolf for the third optimal solution. The α, β, and δ wolves are considered to update the position of other gray wolves, and the computation formulas are expressed as:

$$D_\alpha = |C_1 \times X_\alpha - X(t)|$$

$$D_\beta = |C_2 \times X_\beta - X(t)|$$

$$D_\delta = |C_3 \times X_\delta - X(t)| \tag{7}$$

$$X_1 = |X_\alpha \times A_1 - D_\alpha|$$
$$X_2 = |X_\beta \times A_2 - D_\beta|$$
$$X_3 = |X_\delta \times A_3 - D_\delta| \tag{8}$$

And

$$X(t+1) = (X_1 + X_2 + X_3)/3 \tag{9}$$

In this, D_α is used to show the distance amid the current grey wolf and α wolf; D_β is the distance amid the current grey wolf and β wolf; the distance amid the current grey wolf and δ wolf is illustrated with D_δ; the position vectors of α wolf is represented withX_α, β wolf with X_β, and δ wolf with X_δ. $X(t)$ is used to depict the present position of the grey wolf. The random vectors measured in Formula (6) are defined with C_1, C_2, and C_3. Formula 7 computes A_1, A_2, and A_3. The step length and direction of grey wolf individuals to α, β, and δ wolves are represented using Formula (8). Formula (9) is used to show the position-updating formula of grey wolf individuals. Figure 1 shows the algorithm flow chart of GWO as described earlier.

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

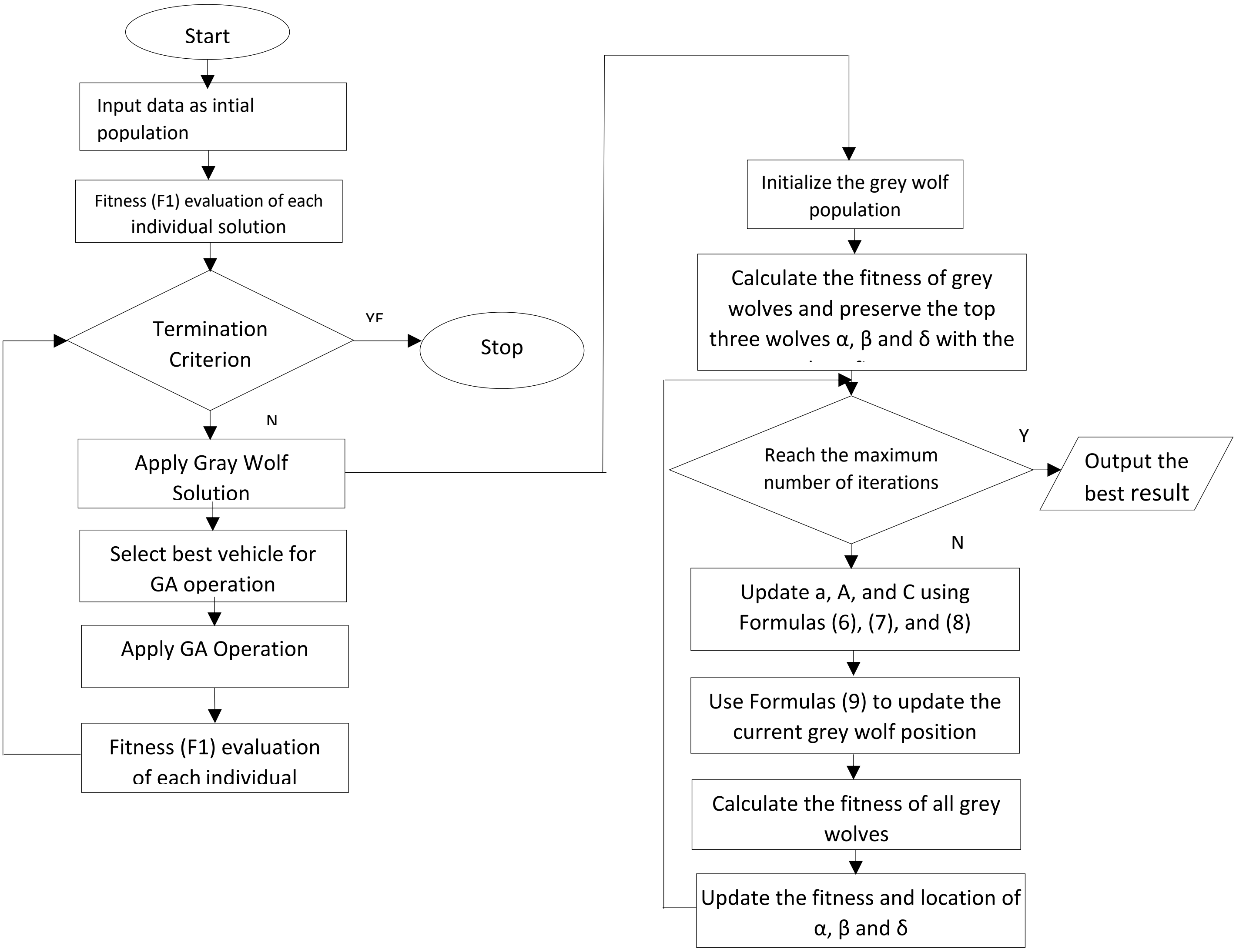

Figure 2. Proposed Feature Extraction Model

3.3. Classification

The voting classification method is applied for the sentiment classification. The input features for the classification can be computed effectively using feature vectors. A key strategy for addressing classification-related problems is controlled learning. When the classification system is trained, the anonymous data can be anticipated. KNN, SVM, and decision trees will be used in this research project's voting process for sentiment classification. A voting classification algorithm refers to an ML framework which works effectively when it is trained on ensemble of various frameworks, and assists in predicting the output with respect to the class having higher probability of being the output. It merely accumulates the outcomes obtained from every algorithm that undergoes the voting algorithm and forecasts the output class according to the vote with the highest majority. The concept is that rather than building individual, specialized frameworks and to verify the accuracy for every algorithm, we build a framework which is trained through multiple algorithms and forecasts approach that relies on their aggregate majority of votes for every output variable.

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

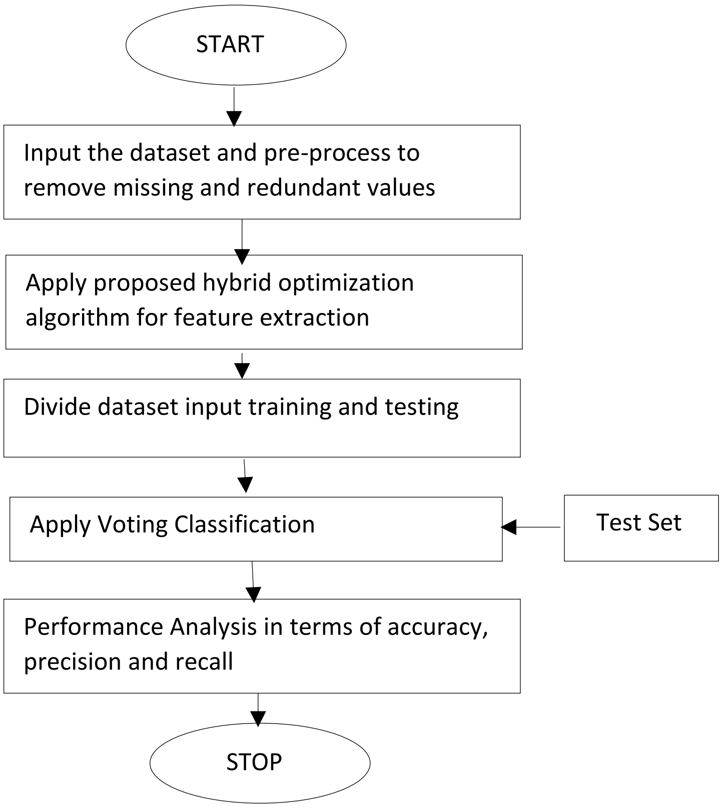

Figure 3. Proposed Calssification Model

3.3.1. Hybrid Model for Feature Extraction

Input F: Input Data

N: Size of population

D: Dimension of feature

Output S: Selected features

For each particle 1 to N

Initialize particle

End for

While (Max iteration is not reached)

Do

For each particle 1 to N in D

Calculate fitness value

if the fitness value is better than the best fitness value (gBest) in history

Update current value as the new gBest

End for

Choose the particle with best fitness value of all the global Best

For each wolf

Calculate wolf velocity

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

use gBset and velocity to update wolf data

End for

End while

3.3.2. Voting classifier algorithm

Input: A Stream of pairs (x,y),

Parameter β∈ (0,1)

Output: A Stream of prediction $\hat{y}$ for each x

1. Initialize experts $C_1 \ldots . C_n$ with weight $\omega_i = 1/N$ each

2. for each x in stream

do Collect Predictions $C_1(x) \ldots C_N(x)$

$P \leftarrow \sum_i \omega_i \cdot C_i(x)$

$\hat{y} \leftarrow Sign\ (P - \frac{1}{2})$

for $i \in 1 \ldots N$

do if $(C_i(x) \neq y)$ then

$\omega_i \leftarrow \beta \cdot \omega_i$

$S \leftarrow \sum_i \omega_i$

for $i \in 1 \ldots .. N$

do $\omega_i \leftarrow \omega_i / s$

RESULTS

In this research work, model is proposed for the sentiment analysis of twitter data. The proposed model has various phases which include pre-processing, feature extraction and classification. The hybrid optimization algorithm is proposed for the feature extraction which is the combination of genetic algorithm and gray wolf algorithm. The voting classifier model is the applied for the classification which is the combination of SVM, KNN and decision tree model. The performance of proposed model is analysed in terms of accuracy, precision and recall.

4.1. Dataset Discription

Twitter is a famous microblogging platform. The personal profile pages of the users of this microblogging site are in the millions. This page contains personal data of customers. Customers can follow each other in order to get connected and obtain the content of other users in a simple way. As per a survey, 50 Billion tweets were posted every day. Twitter posts have millions of opinions. Thus, Twitter has received a lot of attention in the research area for opinion mining as it is linked to a range of online applications.

4.2. Performance Analysis

In this section performance analysis metric are presented. The details of the metrics are presented below: -

Accuracy: - Accuracy is used to measure the performance in the evidence domain recovery and processing of the data. The fraction of the results that are successfully classified can be represented by equation as follows:

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

Journal of Information Systems Engineering and Management
2025, 10(49s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Precision: - Precision is a performance assessment that measures the ratio of correctly identified positives and the total number of identified positives. This can be seen as follows:

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

Recall: - The recall is also referred to as the sensitivity, which is the ratio of connected instances retrieved over the total number of retrieved instances and can be seen as follows:

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

4.3. Outcomes

The proposed model is implemented on the authentic dataset for the sentiment analysis. The proposed model is compared with existing models like KNN, Naïve Bayes, Decision tree and Random Forest

Table 1. Performance Analysis

Model	Accuracy	Precision	Recall
KNN	97.05	97	97
Naïve Bayes	89.21	89	89
Decision Tree	96.07	96	96
Random Forest	96	86	96
Proposed	98.89	98	98

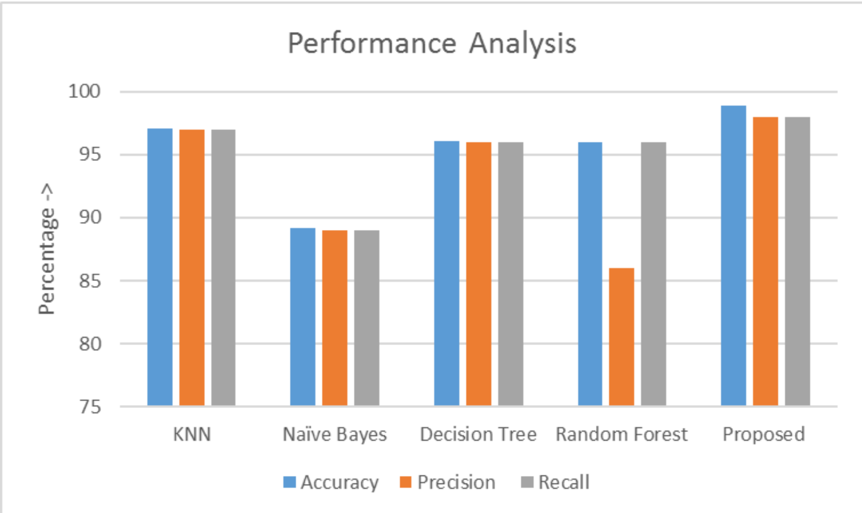

Figure 4. Performance Analysis

As shown in figure 4, the performance of proposed model is compared with other models like KNN, Naïve Bayes, Decision tree and Random Forest. It is analyzed from the results that proposed model achieves approx.99 percent accuracy for the sentiment analysis.

DISCUSSION

This research work presents novel model for the sentiment analysis which is based on the feature extraction and optimization. The various authors address issue that when features are not extracted efficiently it direct affect accuracy of the model for sentiment analysis. The hybrid optimization algorithm is presented which the combination of genetic and gray wolf optimization algorithms. The proposed model extract feature efficiently and which will be given as input to voting classifier for the classification. The proposed model is compared with various exsting classification models like KNN, Naïve bayes, Decision tree and random forest.

Due to efficent feature extraction proposed model give high accuracy. As it is analysed from the results that proposed model has maximum accuracy, precision and recall which is approximate 98 percentage as compared to KNN, Naïve bayes, Decision tree and random forest. The KNN, Naïve bayes, Decision tree, Random forest achieves accuracy, precision, recall values 97 percent, 89 percent, 96 percent and 96 percent respectively. It is analysed that proposed model achieves between 1 to 2 percentage higher accuracy as compared to KNN classifier which is the dominating classifier amoung all existing classifiers.

CONCLUSION

The feature extraction process establishes the connection between the attribute and the target suite. The last level of classification is concerned with the implementation of a classification scheme which can classify data into three groups: positive, negative and neutral. The existing strategy put on a hybrid classifier architecture for predicting the sentiment of posted tweets, but still, the accuracy and precision are not satisfactory. In this work, a voting classification methodology is developed by integrated KNN, SVM and decision tree classifiers to analyze sentiments in posted tweets. This work validates the performance of the formulated architecture with reference to three performance measures (accuracy, recall, precision). The features of the proposed model are extracted using hybrid optimization algorithm which is the combination of genetic algorithm and gray wolf algorithm. The voting classification model is applied for the classification and it is analyzed that proposed model achieves accuracy of 98.89 percent which is approximate 4 percent higher than the existing machine learning models.

REFRENCES

[1] S. Tiwari, A. Verma, P. Garg and D. Bansal, "Social Media Sentiment Analysis on Twitter Datasets," 2020 6th International Conference on Advanced Computing and Communication Systems (ICACCS), Coimbatore, India, 2020, pp. 925-927, doi: 10.1109/ICACCS48705.2020.9074208.

[2] S. A. El Rahman, F. A. AlOtaibi and W. A. AlShehri, "Sentiment Analysis of Twitter Data," 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-4, doi: 10.1109/ICCISci.2019.8716464.

[3] J. Ferdoshi, S. D. Salsabil, E. R. Rhythm, M. H. K. Mehedi and A. A. Rasel, "Unveiling Twitter Sentiments: Analyzing Emotions and Opinions through Sentiment Analysis on Twitter Dataset," 2023 Computer Applications & Technological Solutions (CATS), Mubarak Al-Abdullah, Kuwait, 2023, pp. 1-7, doi: 10.1109/CATS58046.2023.10424206.

[4] D. K. Nugroho, "US presidential election 2020 prediction based on Twitter data using lexicon-based sentiment analysis," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 136-141, doi: 10.1109/Confluence51648.2021.9377201.

[5] R. Patil, N. Gada and K. Gala, "Twitter Data Visualization and Sentiment Analysis of Article 370," 2019 International Conference on Advances in Computing, Communication and Control (ICAC3), Mumbai, India, 2019, pp. 1-4, doi: 10.1109/ICAC347590.2019.9036800.

[6] V. Ramanathan, H. Al Hajri and A. Ruth, "Conceptual Level Semantic Sentiment Analysis Using Twitter Data," 2024 International Conference on Advances in Data Engineering and Intelligent Computing Systems (ADICS), Chennai, India, 2024, pp. 1-8, doi: 10.1109/ADICS58448.2024.10533498.

[7] S. Saini, R. Punhani, R. Bathla and V. K. Shukla, "Sentiment Analysis on Twitter Data using R," 2019 International Conference on Automation, Computational and Technology Management (ICACTM), London, UK, 2019, pp. 68-72, doi: 10.1109/ICACTM.2019.8776685.

[8] M. R. Hasan, M. Maliha and M. Arifuzzaman, "Sentiment Analysis with NLP on Twitter Data," 2019 International Conference on Computer, Communication, Chemical, Materials and Electronic Engineering (IC4ME2), Rajshahi, Bangladesh, 2019, pp. 1-4, doi: 10.1109/IC4ME247184.2019.9036670.

[9] A. Madan and U. Ghose, "Sentiment Analysis for Twitter Data in the Hindi Language," 2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2021, pp. 784-789, doi: 10.1109/Confluence51648.2021.9377142.

[10] S. Dhawan, K. Singh and P. Chauhan, "Sentiment Analysis of Twitter Data in Online Social Network," 2019 5th

International Conference on Signal Processing, Computing and Control (ISPCC), Solan, India, 2019, pp. 255-259, doi: 10.1109/ISPCC48220.2019.8988450.

[11] V. Rekha, R. Raksha, P. Patil, N. Swaras and G. L. Rajat, "Sentiment Analysis on Indian Government Schemes Using Twitter data," 2019 International Conference on Data Science and Communication (IconDSC), Bangalore, India, 2019, pp. 1-5, doi: 10.1109/IconDSC.2019.8817036.

[12] P. Dauni, M. A. Ramdhani, D. R. Suryapratama, Jumadi, W. B. Zulfikar and R. S. Fuadi, "Twitter User Sentiment Analysis for RUU Omnibus Law Using Convolutional Neural Network," 2022 IEEE 8th International Conference on Computing, Engineering and Design (ICCED), Sukabumi, Indonesia, 2022, pp. 1-6, doi: 10.1109/ICCED56140.2022.10010404.

[13] N. Korivi, K. S. Naveen, G. C. Keerthi and V. M. Manikandan, "A Novel Stock Price Prediction Scheme from Twitter Data by using Weighted Sentiment Analysis," 2022 12th International Conference on Cloud Computing, Data Science & Engineering (Confluence), Noida, India, 2022, pp. 623-628, doi: 10.1109/Confluence52989.2022.9734139.

[14] S. Zahoor and R. Rohilla, "Twitter Sentiment Analysis Using Lexical or Rule Based Approach: A Case Study," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), Noida, India, 2020, pp. 537-542, doi: 10.1109/ICRITO48877.2020.9197910.

[15] S. S. Sarpal, "Gradient Boost Decision Tree Classifier Based Big Data Analysis for Twitter Sentiment Analysis Using Adaptive Neural Network," 2023 International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE), Ballar, India, 2023, pp. 1-6, doi: 10.1109/ICDCECE57866.2023.10151161.

[16] Chandra Gupta Maurya, Sudhanshu Kumar Jha, "Sentiment Analysis: A Hybrid Approach on Twitter Data", Procedia Computer Science, 2024, Vol. 235, Pages 990-999, doi.org/10.1016/j.procs.2024.04.094.

[17] L. Wang, J. Niu and S. Yu, "SentiDiff: Combining Textual Information and Sentiment Diffusion Patterns for Twitter Sentiment Analysis," in IEEE Transactions on Knowledge and Data Engineering, vol. 32, no. 10, pp. 2026-2039, 1 Oct. 2020, doi: 10.1109/TKDE.2019.2913641

[18] A. Mahmood, H. U. Khan and M. Ramzan, "On Modelling for Bias-Aware Sentiment Analysis and Its Impact in Twitter," in Journal of Web Engineering, vol. 19, no. 1, pp. 1-27, January 2020, doi: 10.13052/jwe1540-9589.1911.

[19] A. Feizollah, S. Ainin, N. B. Anuar, N. A. B. Abdullah and M. Hazim, "Halal Products on Twitter: Data Extraction and Sentiment Analysis Using Stack of Deep Learning Algorithms," in IEEE Access, vol. 7, pp. 83354-83362, 2019, doi: 10.1109/ACCESS.2019.2923275.

[20] M. Bibi, W. Aziz, M. Almaraashi, I. H. Khan, M. S. A. Nadeem and N. Habib, "A Cooperative Binary-Clustering Framework Based on Majority Voting for Twitter Sentiment Analysis," in IEEE Access, vol. 8, pp. 68580-68592, 2020, doi: 10.1109/ACCESS.2020.2983859.

[21] M. Bibi et al., "Class Association and Attribute Relevancy Based Imputation Algorithm to Reduce Twitter Data for Optimal Sentiment Analysis," in IEEE Access, vol. 7, pp. 136535-136544, 2019, doi: 10.1109/ACCESS.2019.2942112.

[22] S. E. Saad and J. Yang, "Twitter Sentiment Analysis Based on Ordinal Regression," in IEEE Access, vol. 7, pp. 163677-163685, 2019, doi: 10.1109/ACCESS.2019.2952127.

[23] L. Chaudhary, N. Girdhar, D. Sharma, J. Andreu-Perez, A. Doucet and M. Renz, "A Review of Deep Learning Models for Twitter Sentiment Analysis: Challenges and Opportunities," in IEEE Transactions on Computational Social Systems, vol. 11, no. 3, pp. 3550-3579, June 2024, doi: 10.1109/TCSS.2023.3322002.

[24] A Sanjay Neogi, Kirti Anilkumar Garg, Ram Krishn Mishra, Yogesh K Dwivedi, "Sentiment analysis and classification of Indian farmers' protest using twitter data", International Journal of Information Management Data Insights, 2021, Vol. 1, Issue 2, 100019, doi.org/10.1016/j.jjimei.2021.100019.

[25] V. Mahalakshmi, P. Shenbagavalli, S. Raguvaran, V. Rajakumareswaran, E. Sivaraman, "Twitter sentiment analysis using conditional generative adversarial network", International Journal of Cognitive Computing in Engineering, 2024, Vol. 5, Pages 161-169, doi.org/10.1016/j.ijcce.2024.03.002.

[26] S Jabalameli, Yanqing Xu, Sujata Shetty, "Spatial and sentiment analysis of public opinion toward COVID-19 pandemic using twitter data: At the early stage of vaccination", International Journal of Disaster Risk Reduction, 2022, Vol. 80,103204, doi.org/10.1016/j.ijdrr.2022.103204.