Sentiment Analysis on User Reviews based on SGD and machine learning techniques

Vidyashree K P [(1)]
Assistant Professor, Vidyavardhaka College of Engineering, Mysuru
Affiliated to Visvesvaraya Technological University, Belagavi

Dr A B Rajendra [(2)]
Professor, Vidyavardhaka College of Engineering, Mysuru
Affiliated to Visvesvaraya Technological University, Belagavi

Abstract

Opinion investigation is a method used to separate emotional data from printed information. Due to the expansion of social media platforms and the increased daily production of user-generated material, it has become more significant in recent years. Opinion examination is a well-known method for separating emotional data from text information, and it has various applications in different fields, including showcasing, client criticism investigation, and political theory. In this research work, we describe a study on sentiment analysis using ensemble machine learning techniques. Our technique combines the predictions of various base classifiers to increase the overall accuracy of sentiment categorization. We also examine how various preprocessing methods, like as tokenization, stemming, and stop-word removal, affect the precision of the models. Gradient Boosting and Random Forest are two gathering computations that are the subject of our study, along with more conventional machine learning techniques like Naive Bayes and Support Vector Machines. Additionally, we look into how feature engineering methods like sentiment lexicons and n-grams impact the performance of the models. Our results demonstrate that the Random Forest algorithm obtains the best accuracy and that ensemble approaches outperform traditional machine learning algorithms. Our results also show that feature engineering strategies are effective in improving the performance of the models. Based on these findings, it appears that sentiment analysis models for real-world applications that are more accurate and robust can be constructed using ensemble methods. Businesses and organizations that use sentiment analysis to keep an eye on market trends, customer feedback, and brand reputation should take note of these findings.

Keywords: Sentiment analysis, Machine learning, Ensemble methods, Text data, Classification, Accuracy

1. Introduction

Twitter is becoming a valuable resource for sentiment research. These platforms' enormous user-generated content provides crucial insights into the public's perceptions on numerous issues, products, and enterprises. Determining a tweet's polarity, or whether it conveys a good, negative, or neutral attitude, is required for attitude categorization on Twitter data. In recent years, sentiment categorization tasks performed by machine learning algorithms have yielded positive results. They now choose the Stochastic Gradient Descent (SGD) method due to its effectiveness and scalability. On the other hand, SGD's accuracy may be constrained by the nature of the features and the volume of training data. However, ensemble classification techniques have been created.

The following contributions are made in this paper,

1. Using the SGD algorithm and ensemble classification approaches, we offer a research on sentiment classification on Twitter data.

2. We investigate several feature engineering strategies such as n-grams, sentiment lexicons, and part-of-speech tagging.

3. Analyze the effectiveness of several classification techniques, including Naive Bayes, SVM, and Random Forest.

4. We also examine the impact of ensemble size on model accuracy and compare the effectiveness of individual classifiers to that of ensemble approaches.

Our findings indicate that ensemble classification algorithms may greatly enhance sentiment classification accuracy on Twitter data when compared to individual classifiers. Our research provides insights on the features and classifiers that should be applied for this task as well as highlights the usefulness of ensemble classification algorithms for sentiment classification on Twitter data.

2. Related works

In order to address the problems with sentiment analysis of COVID-19-related Twitter data, Usman Naseem et al. [2] conducted a research. The authors gathered datasets from Tweepy and classified them as good, negative, or neutral using the Text Blob tool. They subsequently developed a method for developing classifiers and assessing the sentiment of Twitter data. The study also highlighted how challenging it is to reduce text's inherent meaning, especially when utilising natural language processing (NLP) techniques. These methods could be unable to analyse texts that are in opposition to one another or appreciate the subtleties of language and context in sentiment analysis. This emphasises the demand for more sophisticated NLP techniques and feature engineering techniques to improve sentiment classification precision.

Using a hybrid approach, Ankit Srivastava et al. [3] recommended analysing sentiment on Twitter. The data was collected via Twitter API datasets and put through a Nave Bayes Classifier analysis. The hybrid approach involved tokenizing and data removal after removing irrelevant information from the raw data. The feature sets were then extracted using a combination of unigram and bigram methods. The researchers found that using a trigram approach on pre-processed data had an impact on the system's effectiveness. This means that both the accuracy and effectiveness of sentiment analysis algorithms for Twitter data can be significantly influenced by the feature set selection and pre-processing methods employed. To analyse and assess various feature engineering methodologies, more research may be done.

Anisha P. Rodrigues and Niranjana N. Chiplunkar [4] introduced the Hybrid Lexicon-Naive Bayesian Classifier (HL-NBC) approach for analysing sentiment in Twitter data. This technique efficiently categorises tweets and filters out undesirable ones, and it was compared to lexicon and NBC for unigram and bigram characteristics. The HL-NBC approach was used to categorise and assess a vast quantity of real-time data acquired from diverse subjects on Twitter. The scientists did highlight, however, that the technique had difficulty identifying lingual feelings because sentiment resources for many languages were unavailable in real-time. This shows a prevalent issue in sentiment analysis, where the availability and quality of resources such as lexicons can impact technique performance.

The multimode approach developed by Akshi Kumar and Geetanjali Garg [5] is effective for doing sentiment analysis on both textual and visual data. Whether incoming tweets are text or visual data, the system can discern the polarity of the sentiment present in them. While sentimental analysis of text is carried out by separating the text from the image using an optical recognizer, sentimental analysis of pictures uses senti-bank and senti-strength regions in areas of Regional Convolution Neural Network (RCNN). The authors do point out, however, that when recognizing emotion in large-scale photographs, the features of low-level visuals were constrained. This limitation highlights the challenges of performing sentiment analysis on visual data, particularly when working with large-scale images. By creating increasingly complicated models that can extract meaningful information from large-scale images utilizing advanced image processing techniques and deep learning approaches, researchers have attempted to solve these problems. Future studies in this field may focus on enhancing the effectiveness of these algorithms on visual sentiment analysis tasks.

The method developed by Lopez-Chau [6] and his colleagues is not limited to Twitter data and may be used to other social media platforms as well. The authors emphasize the importance of considering the setting in which the data was gathered since it may have a significant influence on the mood expressed in the text. This emphasizes the demand for more sophisticated algorithms capable of capturing and analyzing the context of data in order to improve sentiment analysis precision. Finally, Lopez-Chau and his colleagues' methodology provides a helpful strategy for sentiment analysis of social media data using machine learning techniques. There are still issues to work out, such as accurate data labelling and classifier design, taking into account the context of the data, and researching more sophisticated machine learning algorithms. Addressing these issues is critical for increasing the accuracy and efficacy of sentiment analysis in a variety of applications.

For Twitter sentiment analysis, Shihab Elbagir and Jing Yang [7] developed a four-module framework. The first module retrieves data from Twitter using the Twitter API. The second module preprocesses the gathered data several times to create a dataset suitable for analysis. A number of classifiers based on machine learning approaches are included in the fourth module to categorize tweets as having a negative, neutral, or positive sentiment. The third module covers feature extraction and the creation of a classification model. Bigram and trigram models are not compatible with this strategy; only the unigram model is. Furthermore, the idea can only be applied to ordinal regression using machine learning techniques; classification is not possible.

3. Problem statement

Since tweets frequently use slang, informal language, and other non-standard grammar and vocabulary, analyzing sentiment in these brief communications can be challenging. Deep learning models, machine learning algorithms, and NLP strategies are just a few of the options that have been suggested for sentiment analysis. Poor computational complexity, poor convergence rate, class imbalance, explosive gradients, overfitting, and sluggish processing speed are some of the issues that currently-used methods must deal with. These challenges can have a significant influence on sentiment analysis' accuracy and precision, making it challenging for businesses and organizations to make informed decisions using the analyzed data. By increasing convergence rates, the proposed method, SGD-OA, seeks to overcome some of these issues and may produce sentiment analysis that is more exact and accurate. Additionally, the ensemble approach of your classification model may be able to maximise precision and accuracy for user review sentiment analysis. The field of sentiment analysis needs more study, especially when it comes to analysing real-time data that has high levels of complexity and volatility. This research might result in the development of more accurate and effective algorithms, enhancing organisational and corporate decision-making and assisting in the comprehension of public opinion.

Proposed system

The recommended revised model for Twitter API tweet sentiment analysis seems to be well-organized and comprehensive. Preprocessing, feature extraction, weighted feature selection, and classification are the processes that make up the model. These steps help to improve sentiment analysis accuracy by reducing dataset complexity and selecting the most pertinent attributes.

Pre-processing is the first stage of the model and it consists of many phases, including stop word removal, empty space removal, punctuation removal, and tag removal. This stage aids in cleaning the dataset by eliminating any superfluous words or characters that might impair sentiment analysis accuracy.

The second step of the model involves feature extraction, which uses techniques like Word2vector, TF-IDF, and Bag of n-grams. These approaches aid in the extraction of relevant characteristics from a dataset that may be used to determine the attitudes contained in tweets.
Weighted feature selection is the model's third stage, which entails optimizing the weight function to enable better sentiment analysis. This process helps in the identification of the factors that have the most potential to improve sentiment analysis accuracy.

The classification portion of the methodology involves classifying tweet emotions as good, negative, or neutral. This step assists in identifying the emotions conveyed in tweets and provides illuminating data. For examining the sentiment of tweets from the Twitter API dataset, the proposed model seems to be helpful. The method may be applied for a number of tasks, such as social media analysis, brand monitoring, and customer feedback analysis. However, it is crucial to assess the model's accuracy and effectiveness using appropriate assessment metrics and datasets.

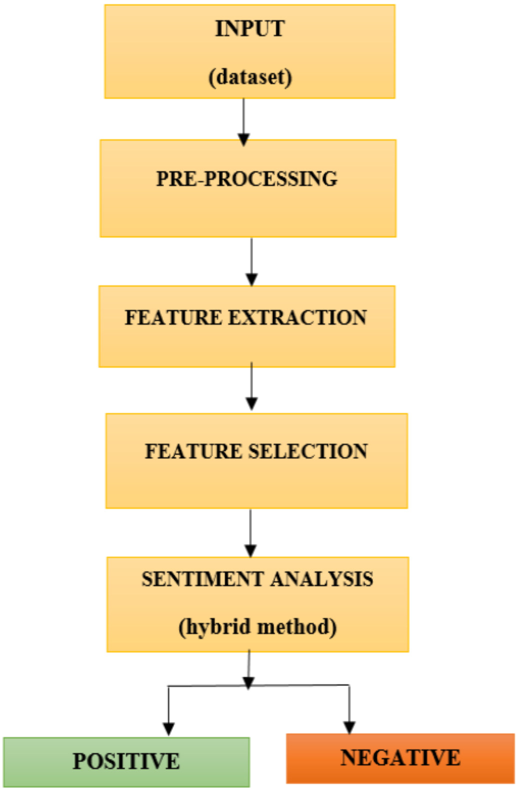

Figure.1 displays the sentiment analysis model using data from Twitter.

3.1 Dataset

Twitter delivers a tremendous amount of real-time data that reflects people's opinions and attitudes, making its usage for sentiment analysis routine practises. For the sentiment analysis to be accurate, the pre-processing methods used to clean the data are essential. The algorithm may concentrate on precisely recognising the emotion of the tweet by deleting unnecessary or repetitive information. The number of columns was manually decreased to 2 by deleting the unnecessary columns. The machine learning model used for sentiment analysis is likely to use supervised learning, which entails training the model on labelled data, to uncover patterns and relationships between characteristics and sentiments. Using new, unlabelled data, the algorithm can then be used to predict the sentiment of the tweets. Results from sentiment analysis are frequently classed as either positive or negative.

Table.1 Dataset description

Number of rows	167000
Number of columns	04
Number of classes	02

3.2 Data Pre-processing

Pre-processing techniques in the provided approach are essential for doing sentiment analysis on Twitter data. By removing frequently used words that do not significantly add to the content of the text, stop words can significantly reduce the dataset's dimensionality and improve sentiment analysis's precision. Additionally, eliminating blank spaces aids in removing any formatting errors that might have occurred during data collection, which is essential for accurate sentiment analysis. Additionally, by ensuring that only the pertinent content is studied, deleting punctuation and tags might increase the effectiveness of sentiment analysis.

It is crucial to remember that the pre-processing methods used should be customised to the particular dataset and issue at hand. Some datasets may require the inclusion of specific stop words in order to guarantee the accuracy of sentiment analysis, for example. Emoji's may also be included in certain Twitter data and should not be removed during pre-processing because they may provide important emotional information.

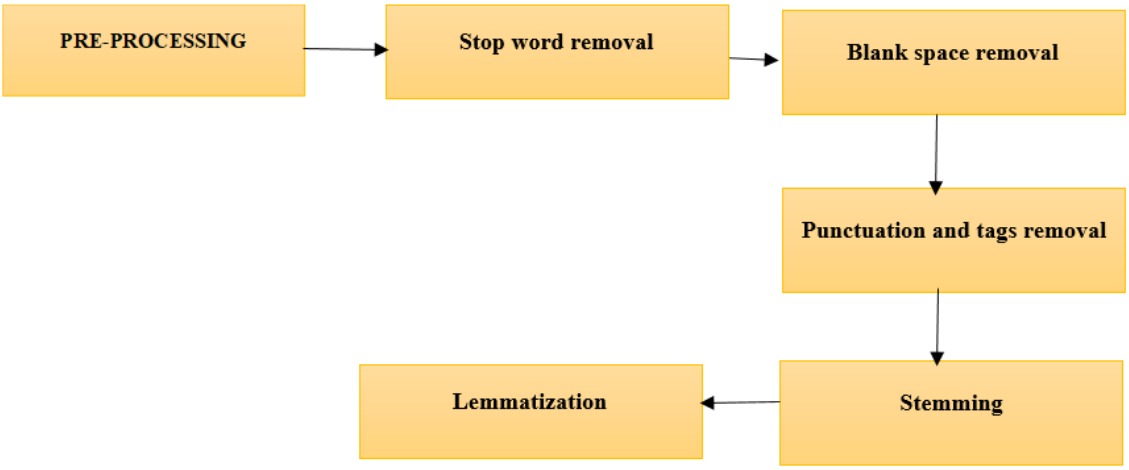

Fig.2 Stages in pre-processing

3.2 Feature extraction

(i) The Bag of n-grams model is frequently used in text analysis and natural language processing to find specific words or phrases in a text. According to the size of the unit, this model fragments the text into tiny n-grams units that can be unigrams, bigrams, or trigrams. The program can find patterns and infer a text's mood by looking at its frequency of certain n-grams in a text. The Bag of n-grams method is extremely helpful for sentiment analysis, which aims to determine the emotional tone of a text and classify it as positive, negative, or neutral. The sentences that were extracted using a bag of n-grams are referred to as T_r^{SE}

(ii) Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic that is commonly used in information retrieval and natural language processing to reflect the importance of a word in a document or corpus. TF-IDF method is based on the idea that words that look frequently in a document, but not in other corpus documents, are important in understanding the significance of that document.TF-IDF method calculates a weight for each term in a document, based on two measures: the term frequency (TF) and the inverse document frequency (IDF). The term frequency is the amount of times a term appears in a document, while the inverse document frequency is a measure of how much information a term provides, and how frequently it appears across all documents in the corpus.

$$\text{TF-IDF }(t,d) = \text{TF}(t,d) * \text{IDF}(t) \text{ ---------------------------------(1)}$$

TF(t,d) is the term frequency of term t in document d. It is calculated by dividing by the total number of words in the document divided by the frequency with which the term appears in the document.
IDF(t) is the inverse document frequency of term t. It is calculated as the logarithm of the total number of documents in the corpus divided by the number of documents that contain the term.
TF-IDF (t,d), represents the importance of the term t in the document d. The higher the TF-IDF score, the more important the term is in the document.

3.4 Feature Selection using SGD

A crucial step in machine learning is feature selection, which helps to find the most relevant characteristics and instructive for the task at hand while eliminating noisy or redundant features. Stochastic gradient descent (SGD) is a well-known optimization approach for training machine learning models, particularly on big datasets.

One technique for doing feature selection using SGD is L1 regularization, sometimes referred to as Lasso regularization. By adding a penalty term to the goal function during L1 regularization, less informative traits are effectively removed from the model by promoting their weights to zero. SGD is a method for iteratively locating the set of model parameters (also known as weights) that minimizes a given loss function. In supervised learning, the loss function assesses the discrepancy between the expected and actual productions of the model. The goal of the optimization is to discover a set of weights that reduces this discrepancy, or error, throughout the whole dataset. The L1 penalty term, which encourages sparsity in the weight vector, is added to the optimization in the context of feature selection, impartial with SGD and L1 regularization. The L1 penalty term favors models with fewer non-zero weights since it is proportional to the sum of the absolute weight values. As a result, the algorithm is prompted to choose the most advantageous features and disregard the less perceptive ones.

3.5 Classification using ensemble method

The ensemble classifier is then used to categorize the traits that the Wrapper-based technique selected. A group of algorithms are used in the ensemble based technique, which used for classification. In this learning, ML classifiers that are already in use are integrated, and classification is carried out using the adaptive boosting approach. The tweets were categorized according to their feelings using an ensemble technique. Logistic Regression (RF), Decision Tree (DT), and Support Vector Machine (SVM) classifiers are employed. to form the Ensemble technique. Generally speaking, compared to current algorithms, the ensemble-based technique has a better accuracy value.

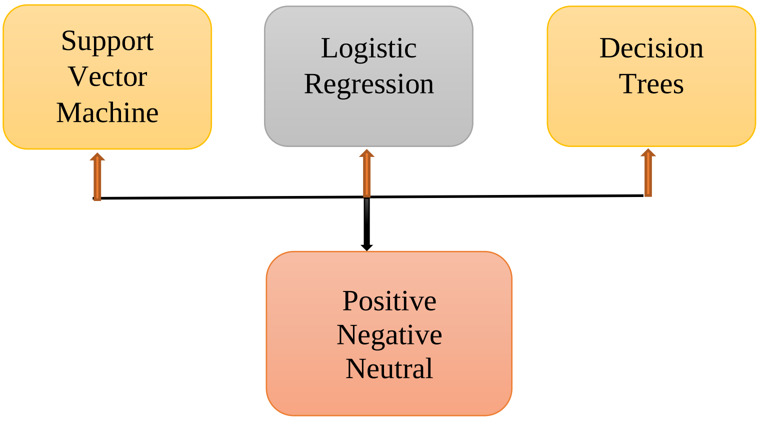

Figure.3 Classification using ensemble method

6. Results and Discussion

The below Pie chart provides a result obtained for various tweets using proposed model.

No of Tweets = 167000

No of classes = 3 ['Positive', 'Neutral, 'Negative']

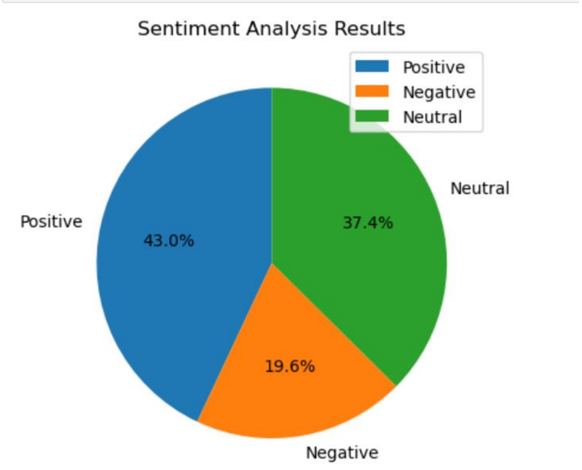

Figure 3: A pie chart depicting the total amount of tweets evaluated.

The Bag of n-grams model is commonly used in natural language processing and text analysis to detect the occurrence of certain words or phrases in a text. By analyzing the frequency of these n-grams in a text, the model can identify patterns and infer the sentiment of text. The Bag of n-grams method is very good for sentiment analysis, which aims to determine the emotional tone of a text, whether it is positive or negative or neutral. To extract phrases from textual input based on the presence of certain n-grams, the Bag of n-grams model was utilized. This approach can be helpful in summarizing the key themes or ideas present in the text, or for identifying relevant sentences for further analysis. Figure.4 shows the top 20 features extracted for the corpus of data

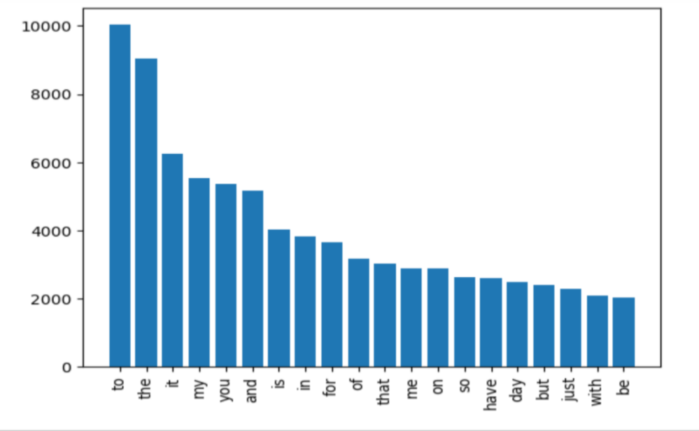

Figure. 4 Top 20 Features extracted after Feature Selection

After splitting the entire amount of tweets into training, testing, and validation, Figure 5 depicts the categorized tweets as positive, negative, and neutral. The tweets extracted into three classes after classification

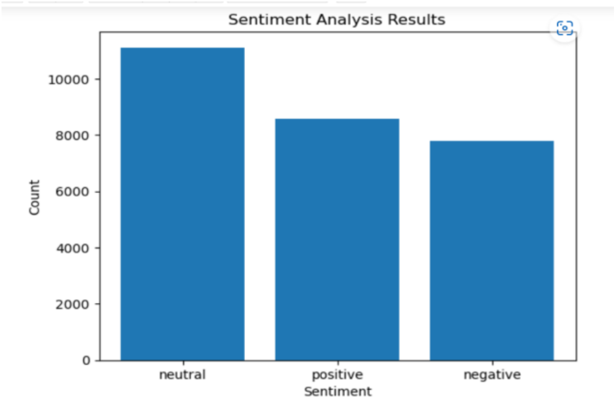

Figure.5 Classification on number of tweets as positive, negative and neutral

Figure.6 Results compared with singular classifiers (Logistic Regression, SVM, Decision Tree, Random Forest, and Naive Bayes) and four performance measures (Accuracy, Precision, Recall, and F1-Score).

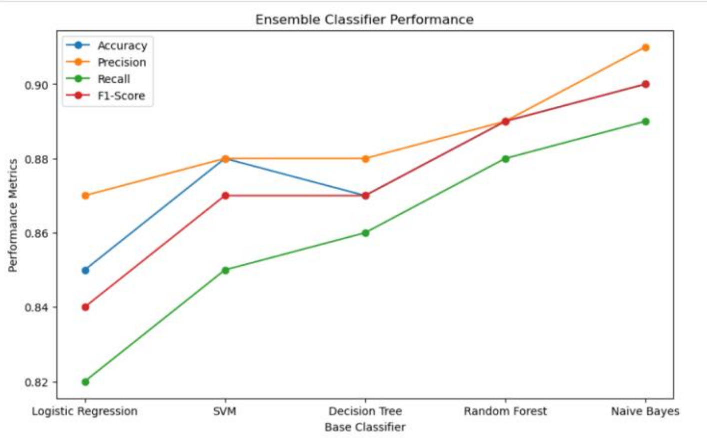

Figure.6 Results obtained from singular classifiers

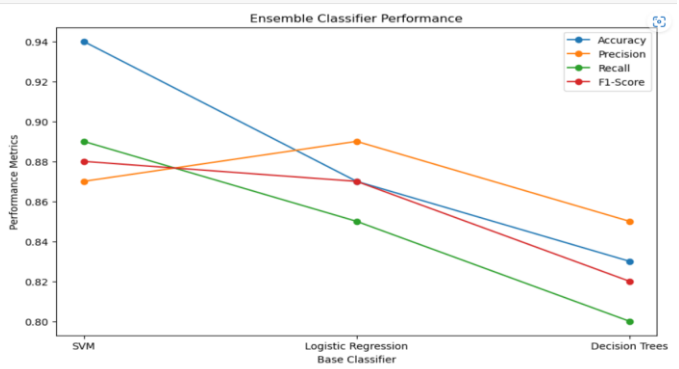

Figure.7 Results obtained from singular classifiers (SVM, Logistic Regression, Decision Trees)

Conclusion

In this paper, ML classifier is used to do sentiment categorization on Twitter tweets. The suggested technique generates pre-trained words using embedding features generated by utilizing lexical polarity based on feature sentiment and n-gram features, and then feeds the feature sets to ensemble classifiers. To capture contextual information, the suggested method constructs a text representation using a stochastic gate neural network and recurrent structure. In the future, the suggested work can be used in other social media networks and as a tool for academics working in the given topic.

References

[1] Kayıkçı, Ş, "SenDemonNet: sentiment analysis for demonetization tweets using heuristic deep neural network". Multimedia Tools and Applications volume 81, pages11341–11378 (2022)

[2] Naseem, Usman, Imran Razzak, Matloob Khushi, Peter W. Eklund, and Jinman Kim. "COVIDSenti: A large-scale benchmark Twitter data set for COVID-19 sentiment analysis." IEEE Transactions on Computational Social Systems 8, no. 4 (2021): 1003-1015.

[3] Srivastava, Ankit, Vijendra Singh, and Gurdeep Singh Drall. "Sentiment analysis of twitter data: A hybrid approach." International Journal of Healthcare Information Systems and Informatics (IJHISI) 14, no. 2 (2019): 1-16.

[4] Rodrigues, Anisha P., and Niranjan N. Chiplunkar. "A new big data approach for topic classification and sentiment analysis of Twitter data." Evolutionary Intelligence (2019): 1-11.

[5] Kumar, Akshi, and Geetanjali Garg. "Sentiment analysis of multimodal twitter data." Multimedia Tools and Applications 78, no. 17 (2019): 24103-24119.

[6] López-Chau, Asdrúbal, David Valle-Cruz, and Rodrigo Sandoval-Almazán. "Sentiment analysis of Twitter data through machine learning techniques." In Software engineering in the era of cloud computing, pp. 185-209. Springer, Cham, 2020.

[7] Saad, Shihab Elbagir, and Jing Yang. "Twitter sentiment analysis based on ordinal regression." IEEE Access 7 (2019): 163677-163685.

[8] Kokab, S. T., Asghar, S., & Naz, S. (2022). Transformer-based deep learning models for the sentiment analysis of social media data. Array, 14, 100157.

[9] Aziz, A. A., & Starkey, A. (2019). Predicting supervise machine learning performances for sentiment analysis using contextual-based approaches. IEEE Access, 8, 17722-17733.

[10] Sunitha, D., Patra, R. K., Babu, N., Suresh, A., & Gupta, S. C. (2022). Twitter sentiment analysis using ensemble based deep learning model towards COVID-19 in India and European countries. Pattern Recognition Letters, 158, 164-170.

[11] Babu, N. V., & Kanaga, E. G. M. (2022). Sentiment analysis in social media data for depression detection using artificial intelligence: a review. SN Computer Science

[12] Soumya, S., & Pramod, K. (2020). Sentiment analysis of malayalam tweets using machine learning techniques. ICT Express, 6(4), 300-305.

[13] Thavareesan, S., & Mahesan, S. (2019). Sentiment analysis in Tamil texts: A study on machine learning techniques and feature representation. Paper presented at the 2019 14th Conference on industrial and information systems (ICIIS).

[14] Vidyashree K P, A B Rajendra, "An Improvised Sentiment Analysis Model on Twitter Data Using Stochastic Gradient Descent (SGD) Optimization Algorithm in Stochastic Gate Neural Network (SGNN)." SN Computer Science (2023) 4:190.

Conflict of Interest

We the authors have no conflicts of interest to disclose.