International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 8 Issue 3, May-June 2024 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Opinion Text Analysis Using Artificial Intelligence

Rahul Sagar[1], Sumit Dalal[2], Sumiran[3]

[1]Student, ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India
[2,3]Assistant Professor, ECE, Sat Kabir Institute of Technology and Management, Bahadurgarh, Haryana, India

ABSTRACT

This paper presents a robust methodology for sentiment analysis of comments leveraging advanced techniques such as Bag of Words (BoW), K-Nearest Neighbors (K-NN), Support Vector Machine (SVM), and Discriminant Analysis. Sentiment analysis plays a crucial role in understanding user opinions, attitudes, and emotions expressed in textual data. By employing BoVW, we extract discriminative features from comments, capturing both semantic and visual cues. These features are then utilized in conjunction with machine learning algorithms including K-NN, SVM, and Discriminant Analysis to classify sentiments accurately. The proposed approach offers a comprehensive framework for sentiment classification, achieving high accuracy and reliability across diverse datasets. Experimental results demonstrate the effectiveness and scalability of the proposed methodology, showcasing its potential for real-world applications in sentiment analysis of comments across various domains.

KEYWORDS: *Opinion Analysis, K-NN, BOW, SVM, Discriminant Analysis*

How to cite this paper: Rahul Sagar | Sumit Dalal | Sumiran "Opinion Text Analysis Using Artificial Intelligence" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-8 | Issue-3, June 2024, pp.168-174, URL: www.ijtsrd.com/papers/ijtsrd64847.pdf

IJTSRD64847

INTRODUCTION

A social media platform like Facebook, provides an opportunity for users to share their views and opinions, as well as connect, communicate, and contribute to specific topics through short-character messages referred to as comments. This can be accomplished using text, images, and videos, among other things, and users can interact by clicking the like, comment, and repost icons. As more individuals utilize social media, the study of data available online can be utilized to shed light on evolving people's views, conduct, and cognition [1]. As a result, employing Twitter or Facebook data for sentiment analysis has grown more common. The growing interest in social media analysis has increased the focus on text analysis technologies such as Natural Language Processing (NLP) and Artificial Intelligence (AI)[2]. Text analysis allows you to assess the sentiments and attitudes of specific target groups. Most of the existing literature concentrates on English texts, however there is an increasing interest in multilingual analysis[3]. Text analysis can be performed by retrieving subjective comments about a specific issue utilizing various sentiments including positive, negative, and neutral [4]. Sentiment analysis can be applied to social media data to investigate variations in people's behavior, feelings, and opinions, like by categorizing the spread tendency of political campaigns. In this work, we use social media research to explore young sentiments. This article examines the feelings of tweets using multiple methodologies, including lexical and machine learning techniques. The time required is a major issue for existing machine learning approaches, posing a hurdle for all firms seeking to transition their operations to be processed by automated workflows. Deep learning techniques have been applied to a variety of real-world applications, including sentiment analysis. Techniques for deep learning use various algorithms to extract information from raw data, including texts or tweets, and express it in specific types of models. These models are used to derive information from novel datasets that have not before been represented.

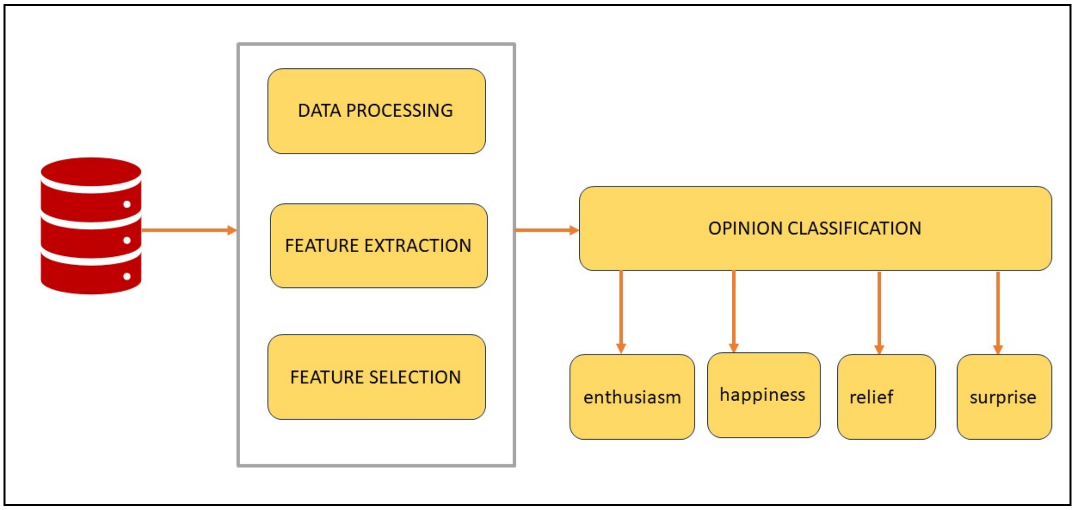


Figure 1: Opinion Classification

Literature review

The key procedure in sentiment analysis is to categorize retrieved data into sentiment polarity such as positive, neutral, and negative. A wide spectrum of emotions can also be thought of, which is fundamental to the growing fields of affective computing and sentiment analysis [5]. Sentiments can be separated into satisfied and angry categories based on the research topic. Sentiment analysis with ambivalence handling can be used to account for more fine-grained outcomes and categorize emotions into such specific areas as anxiety, sadness, anger, excitement, and happiness[6]. Sentiment evaluation is normally performed on text data, but it may also be utilized for analyzing data from devices that use audio- or audio-visual formats such as webcams to study expression, body movement, or noises, referred to as multimodal sentiment analysis[7]. Multimodal sentiment analysis enhances text-based assessment into something more complicated, allowing for the application of NLP for a variety of applications. Progress of NLP is also quickly expanding, driven by different studies, such as neural networks[8]. One instance is the use of Neurosymbolic AI, which mixes deep learning and symbolic reasoning and is seen as a promising tool in NLP for understanding reasonings (Sarker et al. 2021). This demonstrates the broad possibilities for the path of NLP research[9]. There are three major ways to detecting and classifying emotions represented in text: lexicon-based, machine-learning-based, and hybrid techniques.

The **lexicon-based methodology** relies on word polarity, whereas the machine learning approach views texts as an issue of classification and further divides as unsupervised, semi-supervised, and supervised learning. In real-world situations, machine learning and lexicon-based techniques may be utilized in conjunction. When dealing with big text datasets, such as those from Twitter, it is critical to perform data pre-processing before beginning analysis. This involves substituting upper-case letters, eliminating unnecessary words or links, extending contractions, eliminating non-alphabetical letters or symbols, eliminating stop words, and eliminating redundant datasets. Tokenization, stemming, lemmatization, and Part of Speech (POS) labeling should all be used in addition to the fundamental data cleaning process.

Tokenization breaks down texts into smaller parts and converts them into a set of tokens. This makes it easier to compute the frequency of each word in the text and determine its sentiment polarity. Stemming and lemmatization substitute words with their roots. Utilizing stemming, the words "feeling" and "felt" can be mapped to their stem word, "feel".

In contrast, **lemmatization** makes use of the words' contexts. This reduces the dimensionality and intricacy of a bag of words while simultaneously improving the effectiveness of searching for the term in the lexicon when using the lexicon-based approach. POS Tagging mechanically tags the POS of words in the text, including nouns, verbs, and adjectives, which is beneficial to choosing features and retrieval.

LEXICON-BASED APPROACH

The lexicon-based methodology works by first splitting sentences into a bag of words, then comparing them to terms in the sentiment polarity lexicon and their relevant semantic relations, and then calculating the polarity score of the entire text. These techniques can accurately assess if the text's sentiment is favorable, negative, or neutral[10]. The lexicon-based technique tags words with semantic orientation employing either dictionary-based or corpus-based methods. The former is easier and we may calculate the polarity score of words or phrases in the text employing a sentiment dictionary that includes opinion terms.

MACHINE LEARNING APPROACH

Machine learning algorithms can design classifiers that finish sentiment categorization by getting feature vectors, which mostly comprises phases such as data collection and cleaning, feature extraction, training data with the classifier, and evaluating outcomes[11]. Employing machine learning techniques, the dataset should be separated into two parts: training and testing. The training sets are designed to help the classifier learn text features, while the test dataset evaluates its efficiency. Classifiers (for example, Support Vector Machines) classify text into predefined categories. Machine learning is a common technique for text classification among scholars. Furthermore, the accuracy of the same classifier for multiple types of text can vary significantly, so the feature vectors for all kinds of text ought to be trained independently. In the following step, the tweeted data must be vectorized and divided into a training set and a test set, after which the sentiment labels can be predicted by employing various categorization models.

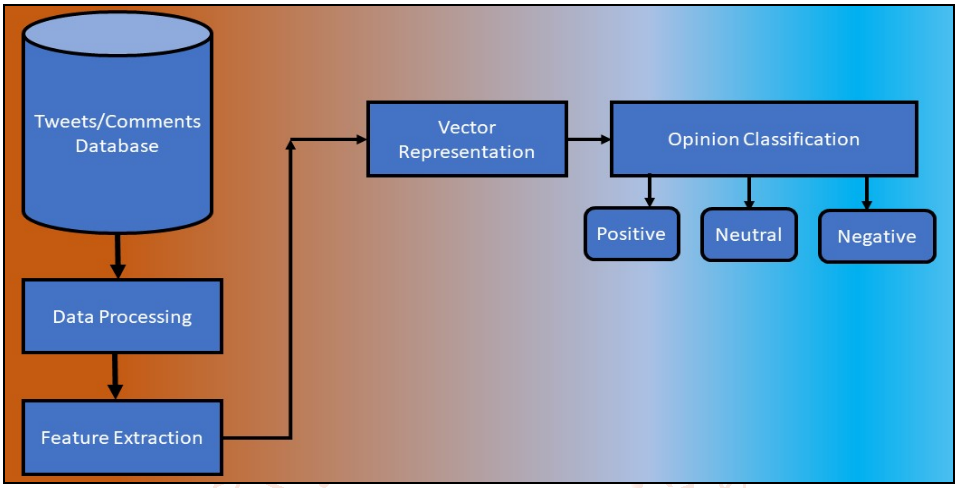


Figure 2: Machine Learning Based Classification

FEATURE REPRESENTATION MODELS

Bag of words (BoW): It translates textual input to numerical data using a fixed-length vector by calculating the frequency of every phrase in tweets/comments. Compute the frequency of terms, which will result in a sparse matrix with clean tokens. A 'Bag-of-Words (BoW)' model is being developed for this paradigm. A 'BoW' (which is additionally referred to as a term-frequency counter) keeps track of how many times a given word comes in a collection's documents. Yet, the MATLAB function 'bagOfWords()' does not separate the text into words. Every single tweet is tokenized using the 'tokenizedDocument()' function.

The provided data is kept in a file containing a table of comments and their category sentiment. Column 1 displays category sentiment, whereas column 2 displays content tweets as text. *Comments* are in sorted in an alphabetical order. The four exclusive opinions are enthusiasm, happiness, relief, and surprise. After tokenizing tweets, a filter may eliminate English punctuation, common stopwords, and rare words (less than 100 times) for counting purposes. This filter distinguishes meaningful words from irrelevant bits of speech.

***Term frequency–inverse document frequency* (*TF–IDF*) Matrix:** It assesses the relevance of a word to the overall text and the importance of the word in the tweet dataset. A TF-IDF matrix can be generated by multiplying the word frequency metric and the inverse document frequency meter for each word in clean tweets.

The final step in preparing the data is to pick the label vectors and feature matrix. There are two groups of these: one for training and another for testing and evaluation. The training set generates a feature matrix and label vectors by picking the top ***m*** rows of the TF-IDF matrix and all columns. The test set for feature matrix and label vectors includes rows following ***m*** rows.

METHODOLOGY

Once data preparation has been finished, the label vectors and features matrix can be utilized for categorizing sentiment using various Classification methods. The prepared data is sent into the categorization methods, which generate a model. The model used the training data to make prediction.

Models for classification: The classification of sentiment is the technique of estimating whether a user's tweet is positive, negative, or neutral based on its feature representation. Classifiers in supervised machine learning approaches, such as random forests, may categorize and predict unlabeled text after training on plenty of sentiment-labeled tweets. The categorization models utilized in this paper are outlined below:

Support Vector Machine Classification

The goal of this system is to find linear separators in vector space to help separate distinct types of input vector data. After the hyperplane has been retrieved, the retrieved text features can be fed into the classifier to predict the outcomes. Furthermore, the primary goal is to identify a line that is closest to the support vectors. The steps for setting up SVC involve computing the distance between the nearest support vectors, also known as the margin, maximizing the margin to find an optimal hyperplane between support vectors from given data, and utilizing this hyperplane as a decision boundary to separate the support vectors.

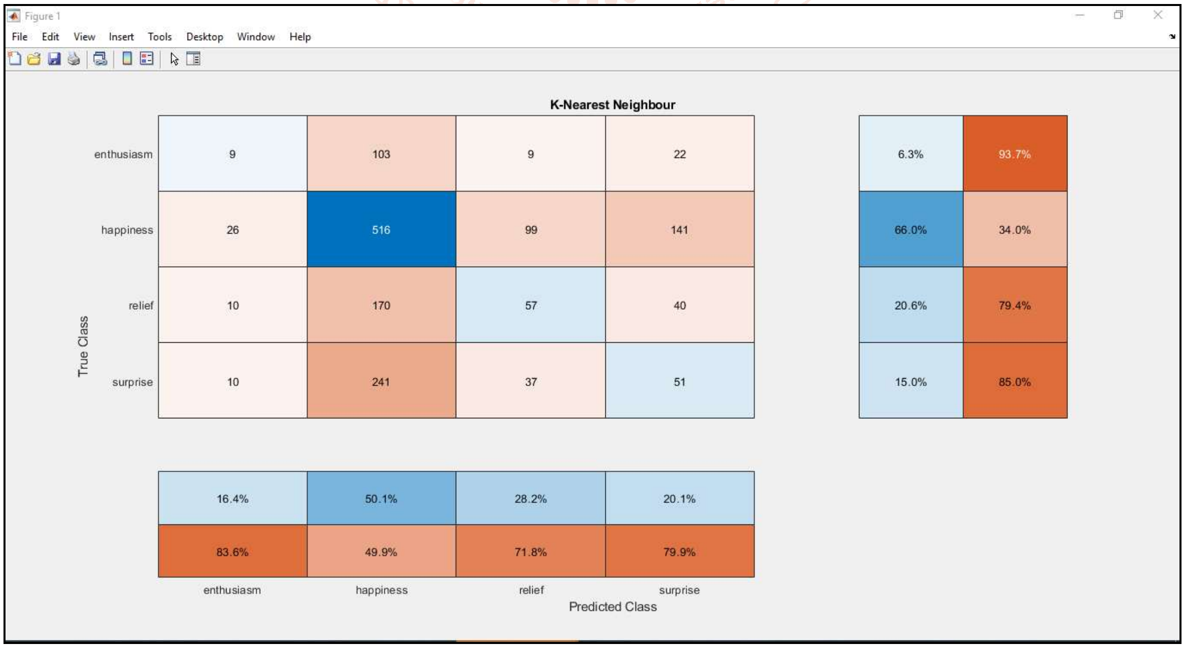
K-Nearest Neighbor (K-NN)

The K-Nearest Neighbors (KNN) method is a prominent machine learning methodology for classification and regression problems. It is based on the assumption that similar data points would have similar labels or values.Throughout the training phase, the KNN algorithm keeps the complete training dataset as a reference. When making predictions, it estimates the distance between the input data point and all of the training instances using a distance metric such as Euclidean distance. The method then identifies the K nearest neighbors of the input data point based on their distances. In the case of classification, the algorithm uses the most prevalent class label among the K neighbors to predict the label for the input data point. Regression uses the average or weighted average of the target values of the K neighbors to forecast the value of the input data point[12].

Linear discriminant analysis (LDA), as the name implies, is a linear framework for classification and reduction of dimensionality. It is a statistical approach that divides data into categories. It detects patterns in features that differentiate between classes. LDA seeks to identify a straight line or plane that best divides these groups while reducing overlap between each class. It allows for accurate classification of fresh data points by increasing the spacing between classes. Simply said, LDA helps make sense of data by determining the most effective way to split various groups, which aids tasks such as pattern detection and classification.

Data Base: we have used Sentiment140 - Automatically labelled database of tweets. We have also used Facebook comments Sentiment analysis database[13]. We combined data from these two databases and feed into our model.

In this research, strategies for text cleaning, polarity calculation, and sentiment classification models are devised and optimized utilizing two distinct sentiment analysis approaches: lexical and machine-learning-based methodologies. We then compared the results of the various approaches, including output and prediction accuracy. Machine-learning-based techniques require tweet labels, but manually annotating a significant amount of data typically takes too long. As a result, 8000 tweets/comments are picked at random in this study, with an average of roughly 1000 tweets each sentiment category.

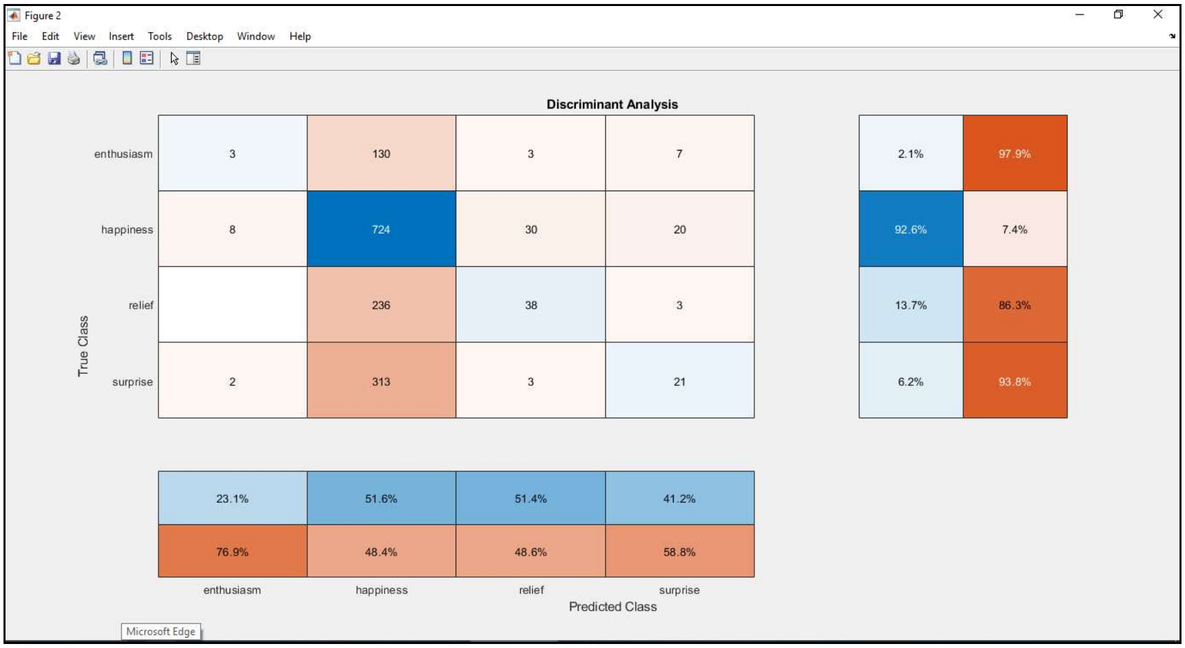


K-Nearest Neighbour

True Class	enthusiasm	happiness	relief	surprise		
enthusiasm	9	103	9	22	6.3%	93.7%
happiness	26	516	99	141	66.0%	34.0%
relief	10	170	57	40	20.6%	79.4%
surprise	10	241	37	51	15.0%	85.0%
	16.4%	50.1%	28.2%	20.1%		
	83.6%	49.9%	71.8%	79.9%		

Predicted Class

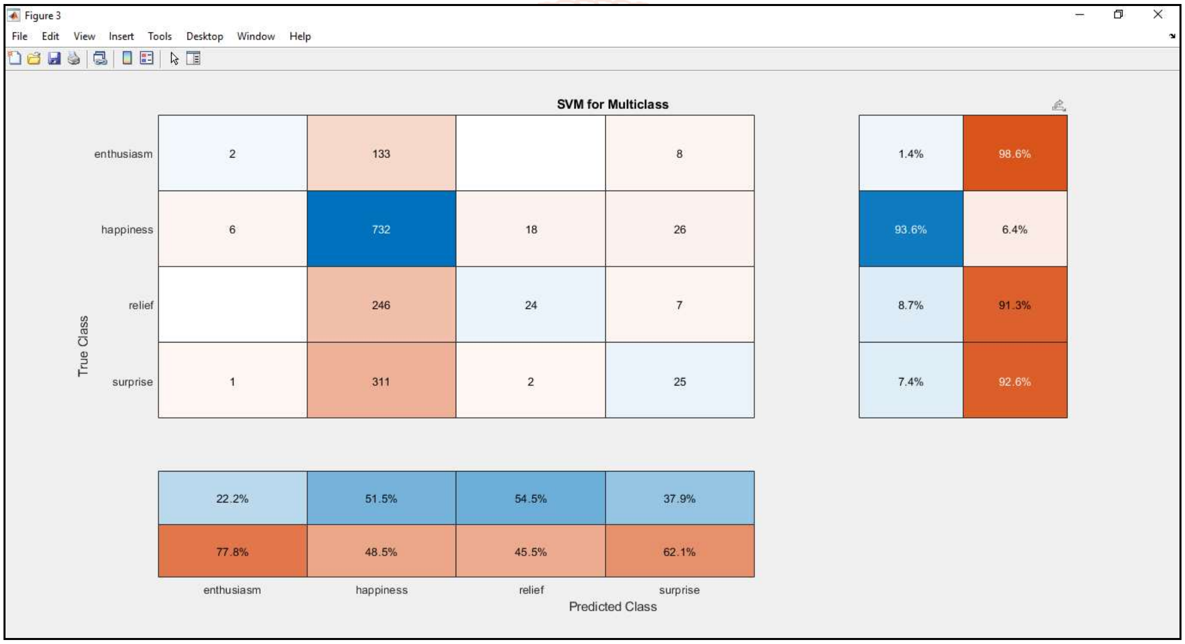
Figure 3: Confusion Matrix for K-NN



Discriminant Analysis

enthusiasm	3	130	3	7	2.1%	97.9%
happiness	8	724	30	20	92.6%	7.4%
relief		236	38	3	13.7%	86.3%
surprise	2	313	3	21	6.2%	93.8%
	23.1%	51.6%	51.4%	41.2%		
	76.9%	48.4%	48.6%	58.8%		
	enthusiasm	happiness	relief	surprise		

Figure 4: Confusion Matrix for Discriminant Analysis



SVM for Multiclass

enthusiasm	2	133		8	1.4%	98.6%
happiness	6	732	18	26	93.6%	6.4%
relief		246	24	7	8.7%	91.3%
surprise	1	311	2	25	7.4%	92.6%
	22.2%	51.5%	54.5%	37.9%		
	77.8%	48.5%	45.5%	62.1%		
	enthusiasm	happiness	relief	surprise		

Figure 5: Confusion Matrix for SVM

Accuracy, Precision, and Recall are the assessment indicators used in this research to assess the performance of each categorization model. Before computing them, the values of the confusion matrix must be known, which are TP (True Positive), TN (True Negative), FP (False Positive), and FN (False Negative). Using the formula below, accuracy is expressed as the proportion of correct observations to total occurrences.

$$Accuracy = \frac{TP+TN}{TP+FP+FN+TN} \tag{1}$$

Precision is the percentage of positive observations that accurately forecast the total number of positive forecasts using the calculation method.

$$Precision = \frac{TP}{TP+FP} \tag{2}$$

Recall is the proportion of genuine positive observations that are accurately identified, computed using:

$$Recall = \frac{TP}{TP+FN} \tag{3}$$

The F1 Score is an in-depth assessment and balancing of precision and recall values. It can be computed as follows:

$$F1 = \frac{2XRecallX\,Precision}{Recall+Precision} \qquad (4)$$

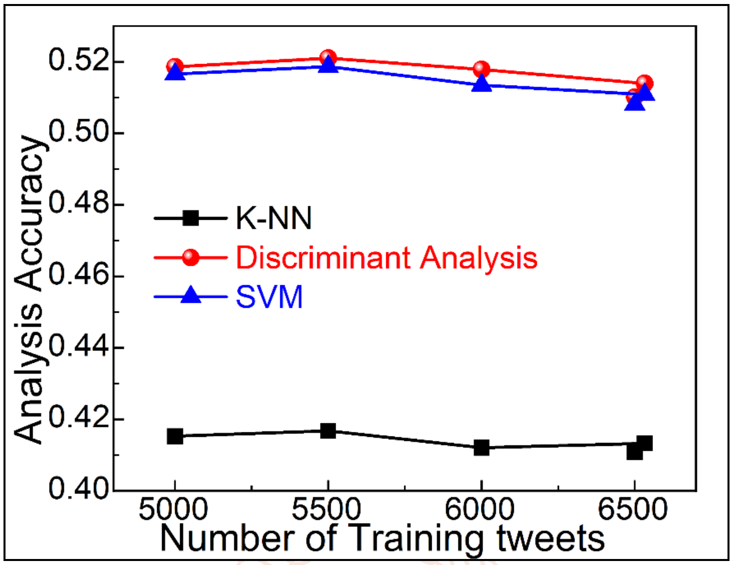


Figure 6: Comparative analysis of Classifier Accuracy

Table 1: Data retrieved After Simulations

Number of Training tweets	Number of tested Tweets	K-NN Accuracy	Discriminant analysis Accuracy	SVM Accuracy
6532	1468	0.4133	0.5140	0.5109
6500	1500	0.4108	0.5101	0.5081
6000	2000	0.4121	0.5179	0.5135
5500	7450	0.4168	0.5211	0.5187
5000	3000	0.4153	0.5186	0.5166

It has been observed that accuracy of Discriminant analysis classifier is better than other two classifiers.

CONCLUSION

Twitter sentiment analysis falls under the topic of text and opinion mining. This approach analyzes tweet sentiment and trains a machine learning model for future use based on its accuracy. The method includes data collection, text pre-processing, sentiment detection, classification, training, and testing the model. primarily we clean the data and utilize unsupervised lexicon-based methods to determine the sentiment orientations of the tweets at every stage. Then, we employ supervised machine learning algorithms utilizing a sample of annotated data to train the K-NN, SVM, and discriminant analysis.

REFERENCES

[1] A. H. Alamoodi *et al.*, "Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review," *Expert Syst. Appl.*, vol. 167, p. 114155, 2021.

[2] R. Sagar, S. Dalal, R. Sharma, and Sumiran, "Classification of Sentiment Analysis Techniques: A Comprehensive Approach," *Int. J. Adv. Res. Arts, Sci. Eng. Manag.*, vol. 11, no. 2, pp. 5200–5205, 2024, [Online]. Available: https://ijarasem.com/admin/img/62_Classification.pdf

[3] S. L. Lo, E. Cambria, R. Chiong, and D. Cornforth, "Multilingual sentiment analysis: from formal to informal and scarce resource languages," *Artif. Intell. Rev.*, vol. 48, pp. 499–527, 2017.

[4] K. Arun and A. Srinagesh, "Multi-lingual Twitter sentiment analysis using machine learning," *Int. J. Electr. Comput. Eng.*, vol. 10, no. 6, pp. 5992–6000, 2020.

[5] E. Cambria, D. Das, S. Bandyopadhyay, and A. Feraco, "Affective computing and sentiment analysis," *A Pract. Guid. to Sentim. Anal.*, pp. 1–10, 2017.

[6] Z. Wang, V. Joo, C. Tong, and D. Chan, "Issues of social data analytics with a new method for sentiment analysis of social media data," in *2014 IEEE 6th International Conference on Cloud Computing Technology*

and Science, 2014, pp. 899–904.

[7] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image Vis. Comput.*, vol. 65, pp. 3–14, 2017.

[8] Y. Kim, "Convolutional neural networks for sentence classification. arXiv [J]," *preprint*, 2014.

[9] M. K. Sarker, L. Zhou, A. Eberhart, and P. Hitzler, "Neuro-symbolic artificial intelligence," *AI Commun.*, vol. 34, no. 3, pp. 197–209, 2021.

[10] S. Zahoor and R. Rohilla, "Twitter sentiment analysis using lexical or rule based approach: a case study," in *2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, 2020, pp. 537–542.

[11] O. Adwan, M. Al-Tawil, A. Huneiti, R. Shahin, A. A. Zayed, and R. Al-Dibsi, "Twitter sentiment analysis approaches: A survey," *Int. J. Emerg. Technol. Learn.*, vol. 15, no. 15, pp. 79–93, 2020.

[12] "K-NN", [Online]. Available: https://www.analyticsvidhya.com/blog/2018/03/introduction-k-neighbours-algorithm-clustering/

[13] "Facebook comments Sentiment analysis", [Online]. Available: https://www.kaggle.com/code/mortena/facebook-comments-sentiment-analysis