Exploring ML Methods for Sentiment Analysis in Text Data

Vishu Agarwal
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
vishuagarwal183@gmail.com

Ashmit Tayal
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
ashmittayal196@gmail.com

Rajani Dixit
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
rajanidixit024@gmail.com

Bharti Chugh
Department of Computer Science and Engineering
KIET Group of Institutions
Ghaziabad, India
bharti.kathpalia@gmail.com

Shikha Jain
School of Engineering and Technology
Vivekananda Institute of Professional Studies- Technical Campus
New Delhi, India
ssjainshikha@gmail.com

***Abstract*— Data Analysis has become an important part of our life, as it helps us to draw useful information, decision-making conclusions on any particular raw data. One component of digital data analysis is sentiment analysis. Sentiment analysis evaluates the emotional tone of textual data and classifies it as neutral, negative, or positive. In business, this analysis is frequently utilized for market social media monitoring, market research, consumer feedback and many more. In order to automate the procedure, machine learning models that recognize patterns in labeled data are frequently used. The objective is to derive meaningful insights from textual data so that decisions can be made with knowledge, effectively. In order to provide important insights into public opinion and behavior, this study uses machine learning techniques to investigate the sentiments conveyed in Twitter data. A wide range of Twitter posts, including ones about trending hashtags, events, and subjects, are gathered as part of the process.**

Sentiment analysis is done using machine learning algorithms. Sentiment analysis and examples of human language are used to train machine learning models. Following this training, machine learning is used by sentiment analysis software to evaluate and rank human language according to its previous training.

The current study proposes a technique that best analyzes the sentiments of the text used in social media is gradient boosting classifier. With this method, an additive model can be built in a step-by-step, forward fashion and any differentiable loss function can be optimized. There are n-class regression trees that are fitted on the negative gradient of the loss function, which could be the multiclass log loss or the binary loss. This process is repeated in each step. In the particular case of binary classification, just one regression tree is induced.

The computational results demonstrates that gradient boosting approach with an accuracy of 93.62% which outperforms is preferable to other classifiers.

Keywords—Sentiment analysis, Twitter, Dataset, Machine learning, Gradient boosting classifier

I. INTRODUCTION

An intriguing area of machine learning, sentiment analysis extracts and understands textual emotions. Sentiment analysis analyzes human language using machine learning. Sentiment analysis uses contextual meaning to estimate a brand's social sentiment and if its product will sell in the market. The dataset is used to train a variety of ML models, including Naïve bayes, K-nearest neighbour classifier, Support vector machines, etc to find patterns and correlations between words and sentiments. The performance level is analysed with the metrics in mind like accuracy, F1 score, precision. The study tackles issues including managing sarcasm, slang, and context-specific terms in an effort to better grasp public emotion on Twitter. Market research, brand management, political analysis, and other fields can all benefit from the sentiment analysis's results. They offer businesses and scholars insightful information gleaned from the enormous volume of viewpoints expressed on the Twitter network.

Potential uses of this research include real-time public sentiment monitoring, which enables proactive response and decision-making tactics across a range of industries. Reference [1] examines a well-known microblog called Twitter and develops models to categorize "tweets" into sentiment categories that are good, negative, and neutral. Sentiment Analysis is needed to store data efficiently and cheaply. You can solve all real-time scenarios with sentiment analysis. Businesses use sentiment analysis to analyze customer opinions, build brand reputation, create better goods, and personalize content. It supports market research, policymaking, and automation. Text sentiment is measured using metrics to determine positive, negative, or neutral. Today, the internet offers several means to convey feelings. Machines are trained with text examples of emotions to identify sentiment without human input. To put it briefly, computers can learn new tasks using machine learning without the need for explicit programming. Sentiment analysis models may learn context, misapplied words and sarcasm beyond definitions. As many as methods and complicated system of equation commands and train machines to carry out sentiment analysis. However, combined they can produce great effects.

Natural language and sentiment are employed in machine learning model training. Sentiment analysis software uses machine learning to score human language after this training. Among the algorithms that are utilized the most are Linear Regression, Support vector machines (SVM), Naïve Bayes, and many more. Every new model proceeds in the direction of least prediction error within the space of possible predictions for every training example. Reference [2] states that opinion mining and sentiment analysis are two exciting new areas of research that are used to find out what people think and feel about particular subjects. As a result, opinion mining and

sentiment analysis are frequently used synonymously to convey the same concept. Reference [3] illustrates that the content of tweets has emerged as a hot research topic concerning the positive or negative polarity of sentiment. Our work with sentiment analysis of tweet contexts demonstrates that group learning, which is made up of the majority vote of the Stacking, support vector machine, SentiStrength, and naïve bayes, can enhance and be more successful in achieving accuracy performance. Reference [4] tries to illustrate why popular text classification techniques like maximum entropy, support vector machines and naïve bayes are necessary. Reference [5] states the characteristic short length and asymmetrical structure of microblogs like Twitter present various additional obstacles for sentiment analysis.

The research literature on sentiment analysis on microblogs identifies two primary study avenues. Reference [6] encompasses the examination of the content on the Internet spanning numerous domains that are seeing a rapid increase in both quantity and number as websites are devoted to particular product categories and have expertise gathering customer reviews from other websites, including Amazon, twitter, etc. Reference [7] states that the data sentiment analysis is a categorization method based on finding if a review is good, negative, or neutral based on the given opinion. Finding the review writer's position at the document, sentence, or aspect levels is SA's main goal.

These days, SA is frequently used to analyze reviews from a variety of different data domains, including reviews of products, movies, hotels, restaurants, and many more. Reference [8] Tweets on Twitter allow users to voice their thoughts on a variety of subjects. This can assist marketers target their campaigns to convey consumers' opinions about products and companies, bullying incidents, and events.

A. Key Contributions

The key contributions in this research are:

a) Creating an efficient sentiment analysis model for the twitter dataset, which can accurately identify the sentiments, moods, effects and biasness.

b) Assessing the effects of various resampling and optimization strategies on the overall performance of sentiment analysis model in order to determine the most effective classifier technique for enhancing its overall performance.

c) Examining and adjusting the model's hyperparameters for better results and accuracy.

II. Literature Review

According to [9] sentiment analysis has gained popularity as a field of study in computational linguistics due to the proliferation of sentiment data from blogs, online forums, and social media sites like Facebook and Twitter. Reference [10] states that the process of locating and categorizing viewpoints or feelings represented in the source text is known as sentiment analysis. By using sentiment analysis in a particular domain, the influence can be determined by information on sentiment classification. Reference [11] addresses several issues and obstacles in the field of sentiment analysis and presents an array of methodologies and concerns associated with the field.

Reference [12] presented that over the past 10 years, opinion study of Twitter data has received a lot of attention. This type of investigation involves examining the words that make up "tweets," or remarks. Because of this, the purpose of this study is to investigate the numerous sentiment analyses that were performed on Twitter data and the results of those analyses. Reference [13] provide a survey and an examination of the most recent developments in option mining techniques, including lexicon-based approaches and machine learning, is presented in this comparative analysis. Also provide a study on twitter data streams by employing a number of different machine learning algorithms, such as Naïve Bayes, Support Vector Machine and Max Entropy. Reference [7] proposes that using a gradient-boosted decision tree classifier, sentiment analysis and sentiment classification of Twitter data are conducted.

Reference [14] suggested a hybrid method of opinion extraction from Twitter data that combines characteristics that are both direct and indirect. According to this method supervised classifiers that consist of a network of artificial neural connections, Naïve Bayes, Maximum Entropy and Support Vector Machines (SVM).

Reference [15] states Sentiment analysis is the method of automatically determining if an entity (i.e., product, people, topic, event, etc.) is the topic of either neutral, negative, or positive opinion expression in a user-generated text. The purpose for conducting this research is to offer a thorough examination of machine learning technique for sentiment analysis utilizing data from Twitter. Reference [16] discusses the current data mining analysis of the Twitter dataset, including the sentiment analysis using machine learning algorithms. Reference [17] provides a method that employs lexicons and classifier ensembles to automatically categorize the tone expressed in tweets. Tweets on a query word are categorized as good or bad. This method has numerous applications, including helping businesses track public opinion about their brands and consumers who can use sentiment analysis to look for products. Reference [18] states that they used machine learning methods to categorize reviews as favourable or negative after preprocessing. The conclusion of the paper states that the most accurate classification results for product reviews are achieved through the application of machine learning techniques. Reference [19] presents that people's opinions about governments, events, goods and services on social media are expressed through words and phrases. Sentiment analysis in natural language processing is the technique of extracting positive and negative polarities from text that is shared on social media platforms. Reference [20] proposes that online user sentiments have a significant impact on lawmakers, goods suppliers, and readers. Sentiment analysis has garnered substantial attention as a means of analyzing and organizing the unstructured data gleaned from social media. Reference [21] focuses on the categorization of tweets' emotional content using information obtained from Twitter. In the past, researchers used machine learning techniques that were already available for sentiment analysis, but the findings indicated that these methods were not producing superior sentiment categorization results. Reference [22] states that in the foreseeable future, financial sector fraud is anticipated to have significant repercussions.

III. Preliminaries

A. Exploratory Data Analysis

In order to construct predictive models, exploratory data analysis, or EDA, is requisite to machine learning since it helps to comprehend the traits, connections, and patterns

found in the dataset. To acquire insights, spot problems with data quality, and guide preparation actions for the best possible model performance, it makes use of descriptive statistics, data visualization, and feature analysis.

Table 1 gives detailed description of the dataset, including target value i.e., "label" (binary value) and the number of data used.

TABLE I. THE DETAIULED DESCRIPTION OF THE DATASET USED

	Description	Value
0	Session ID	123
1	Target	Label
2	Target Type	Binary
3	Original Data Shape	(31962, 2)
4	Transformed Data Shape	(31962, 2)
5	Transformed Train Data Shape	(22373, 2)
6	Transformed Test Data Shape	(9589, 2)
7	Numeric Features	1
8	Preprocess	True
9	Imputation Type	Simple
10	Numeric Imputation	Mean
11	Categorial Imputation	Mode

B. Data Preprocessing

A tweet with a positive sentiment is represented by a value of 1, whereas a tweet with negative emotion is represented by a value of 0.

Table 2 describes each attribute of the dataset that is ID is used as the serial number which contains the numeric values, The label contains binary values (good or bad) and the tweets are the categorial values.

TABLE II. DESCRIPTION OF EACH ATTRIBUTR OF THE DATASET

Column Name	Attribute Description
ID (numerical)	Serial number
Label (binary)	Good or Bad
Tweet (categorial	Twitter Post

To ensure data quality, the dataset was carefully examined for missing, duplicate, and trash values before being subjected to additional analysis. In order to get the dataset ready for machine learning algorithms, the z-score method was utilized to standardize the data by taking the mean value (μ) and dividing it by the standard deviation (σ) of each feature as shown in (1). The label encoding method was then used to convert the categorial features into numerical representations.

$$z = \frac{x - \mu}{\sigma} \quad (1)$$

The data is then centered around zero and its standard deviation is measured to be one, making appropriate for ML algorithms that require normalized data. This technique makes the data suitable for machines.

IV. MACHINE LEARNING MODELS

Reference [23] states machine learning models are able to construct classifiers with extraction of feature vectors, which allows them to complete sentiment categorization. The primary components of these methods are the collection and cleaning of data, the extraction of features, the formation of the classifier through training data, and the evaluation of the results. Through the application of machine learning strategies, the dataset is partitioned into a training dataset and a test dataset. A performance evaluation of the classifier is carried out using the test dataset, whereas the training sets are designed to provide assistance to the classifier in learning the text features.

A. Logistic Regression

It ascertains the likelihood that a specific instance will belong to a class for binary classification tasks. It's a classification algorithm, not regression. It uses the logistic function to model the independent factors and dependent variable log-odds as shown in (2). A threshold is used to make binary predictions from output probabilities. LR is popular for its simplicity, interpretability and versatility.

$$y = \beta_0 + \beta_1 x \quad (2)$$

B. Linear Discriminant Analysis

It reduces dimensionality and classifies in machine learning. The combinations that are linear in characteristics that most separate two or more classes while conserving class discriminating information are its goal. Multi-class classification issues benefit from linear discriminant analysis. The maximizing of between-class variance compared to within-class variance determines class labels and discriminant functions. Linear discriminant analysis assumes Gaussian data and the same covariance matrix for all classes.

C. Naïve Bayes Classifier

They are often classified using the probabilistic machine learning method Naïve Bayes. It is assumed that characteristics given the class designation, are conditionally independent based on the Bayes theorem, as given in (3), simplifying likelihood computation regarding text categorization and spam filtering, NB often executes well despite its "naive" premise. It uses little training data and is computationally efficient.

$$P(A/B) = P(B/A) * \frac{P(A)}{P(B)} \quad (3)$$

D. K- Nearest Neighbour Classifier

It is a basic and simple categorization machine learning model. It classifies new data points using the feature space class majority of their K closest neighbors. How many neighbors there are is determined by user-defined parameter K. The non-parametric, lazy KNN doesn't make assumptions about the data distribution or develop an explicit model during training, making it adaptable for varied datasets.

E. Decision Trees

It is a regression and classification technique for supervised machine learning. Recursively partitioning the dataset by the most important feature at each node yields a tree-like model. The idea is to partition data to reduce impurity and variation. DT are interpretable, visualizable, and can capture complex data relationships.

F. Support Vector Machines

It is a sophisticated and supervised ML method for classification and regression. Data points are divided into groups by identifying a hyperplane that optimizes their distance from one another. SVM performs admirably in high-dimensional areas and handles non-linear interactions with kernel functions. Data points closest to the border make up support vectors of the decision- are selected to find the

best decision boundary. SVM's resilience and capacity to handle complex datasets make them popular.

G. Random Forest Classifier

It is a popular ensemble learning algorithm for categorization. It generates many decision trees during training and delivers class mode for categorization. RF reduces overfitting by employing subsets of characteristics and training data for each tree. It is durable, manages high-dimensional data, and has less variance than individual decision trees, making it a popular and effective machine learning alternative.

H. Gradient Boosting Classifier

The technique is known as ensemble learning, and it involves the generation of a number of unsuccessful learners in a sequential manner, typically decision trees, in order to repair errors that were produced by earlier machine learning models. Using GBC optimization, it reduces errors to a minimum while simultaneously updating forecasts at each stage. Implementations such as XGBoost, LightGBM, and CatBoost are becoming increasingly popular. A powerful and accurate prediction model is produced as a result of the combination of these weak learners, which is the final model.

I. Ridge Classifier

It is a linear grouping technique that regularizes LR to reduce overfitting. It adds L2 regularization (ridge penalty) to the cost function during training. This penalty discourages large coefficients, making the model more stable and generalizable. Ridge Classifier is effective for multicollinearity in feature space, and its regularization parameter determines regularization strength.

J. AdaBoost (Adaptive Boosting) Classifier

It uses ensemble learning for classification. It takes many weak learners, usually decision trees, and turns them into a potent prediction model. AdaBoost weights each instance based on weak learners' performance. Iteratively highlighting misclassified cases improves accuracy. AdaBoost is adaptable, outperforms models, and overfits less.

V. Resampling Methods

A. SMOTE (Synthetic Minority Over-sampling Technique)

In order to address the issue of class imbalance, the popular machine learning approach known as SMOTE generates synthetic samples of the minority class. It interpolates additional instances between samples of the minority class that already exist in order to increase the representation of the minority class in the training data. This results in the minority class being represented more accurately.

B. ADASYN (Adaptive Synthetic Sampling)

This is an addition to the SMOTE algorithm created especially to deal with the unbalanced datasets. By concentrating on the samples that are more difficult to accurately identify, it intelligently creates synthetic samples for the minority class, enhancing the classifier's overall performance.

C. TomekLinks

The TomekLinks technique finds and eliminates the pairs of samples from several classes that are most similar to one another. It seeks to strengthen the divide between the majority and minority classes by getting rid of these occurrences, which could improve algorithmic classification performance. It might not work well in situations when the classes overlap much, though, and it might result in the loss of some potentially helpful samples.

D. ALLKNN (All k- Nearest Neighbors)

This is a classification-based under-sampling strategy which finds and duplicates each minority class instance's k closest neighbors. It seeks to improve the representation of the minority class by duplicating these nearby examples, which may help classifiers perform better on unbalanced datasets. To get the best results, parameter tuning must be done carefully as close duplicates of the nearest neighbors may lead to overfitting.

VI. Performance Evaluation Metrics

The performance of each individual model is based on Accuracy, F1 score, AUC and Precision.

A. Accuracy

The frequency with which a ML model predicts the result accurately is measured by its accuracy.

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of predictions}$$

B. Recall

The percentage of data samples that a ML model correctly classifies as belonging to a particular class i.e., the "positive class", out of the total sample. It is also called as True Positive Rate (TPR).

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

C. Precision

It is a measurement of how effective a machine learning model is. The ratio of true positives to total positive predictions, which is the sum of true positives and false positives, is what we mean when we talk about this particular metric.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

D. F1 Score

The accuracy of a model can be quantifies using the F1 score. The accuracy metric is a system that determines the number of times a model has correctly predicted across the entirety of the dataset under consideration. It is the measure of predictive performance.

$$F1\ score = 2 * \frac{Precision * Recall}{Precision + Recall}$$

E. Kappa

A statistic that is utilized for the purpose of evaluating the efficiency of ML classification models is known as the Kappa Coefficient, which is also widely referred to as Cohen's Kappa Score. The traditional 2x2 confusion matrix serves as the foundation for its formula, which is utilized in the field of statistics and ML for the purpose of evaluating binary classifiers.

$$k = \frac{p_o - p_e}{1 - p_e}$$

F. MCC

An example of the þ coefficient that is a special case is the MCC. In the context of this discussion, a value of one that is positive indicates that the classification is perfect, a value that is close to zero indicates that the prediction was made by chance, and a value that is negative one indicates that the opposite prediction is perfect. Furthermore, all of the negative samples were predicted to be positive, and vice versa.

$$MCC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

VII. Proposed Methodology

ALGORITHM-

Step I: Twitter dataset has been pre-processed and then encoded.

Step II: Generate models for an unbalanced dataset.

Step III: Select the top models for the evaluation of the results.

Step VI: Perform the Re-sampling techniques to balance the dataset.

Step V: Optimization of the model by applying hyperparameters.

Step VI: Prediction based on test data.

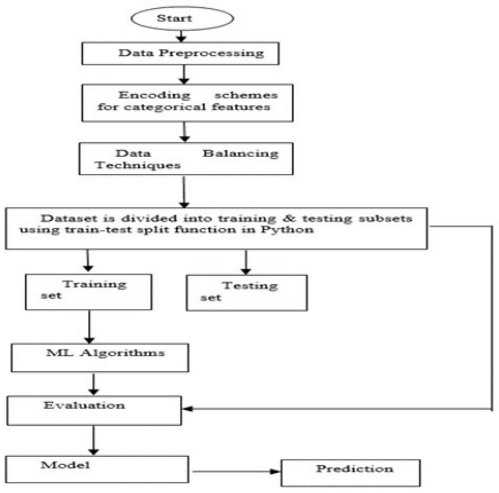

Fig. 1. Flow Chart of the proposed model.

Fig.1 gives block diagram of the model that is being suggested. These are the ways in which the model that was suggested operates: data is gathered and processed, and models are investigated to see which ones are the most appropriate for the type of data; then trains, tests, and evaluates the models. The important thing is to increase the accuracy of machine learning models. Providing labeled data to an AI system. This indicates that a label has been assigned to each piece of information in accordance with its unique significance.

VIII. Results and Discussions

Table 3 compares the various ML models on the processed unbalanced datasets various performance metrics including precision, recall, F1 score and recall for 13 different classifiers.

TABLE III. Performance analysis of all the models for imbalanced data

	Model	Accuracy	AUC	Recall	Prec.	F1	Kappa	MCC	TT (Sec)
gbc	Gradient Boosting Classifier	0.9362	0.795	0.0988	0.9289	0.1776	0.1664	0.2889	1.587
knn	K Neighbors Classifier	0.9342	0.7526	0.2587	0.5665	0.3547	0.3251	0.3533	0.184
lightgbm	Light Gradient Boosting Machine	0.9314	0.7951	0.0516	0.634	0.0944	0.085	0.1642	0.458
xgboost	Extreme Gradient Boosting	0.9309	0.7987	0.0478	0.6295	0.0879	0.0784	0.1562	0.195
lr	Logistic Regression	0.9299	0.5184	0	0	0	0	0	0.612
ridge	Ridge Classifier	0.9299	0	0	0	0	0	0	0.034
qda	Quadratic Discriminant Analysis	0.9299	0.6163	0	0	0	0	0	0.037
ada	Ada Boost Classifier	0.9299	0.762	0	0	0	0	0	0.439
lda	Linear Discriminant Analysis	0.9299	0.5184	0	0	0	0	0	0.037
dummy	Dummy Classifier	0.9299	0.5	0	0	0	0	0	0.03
svm	SVM - Linear Kernel	0.9293	0	0	0	0	-0.0011	-0.0054	0.437
et	Extra Trees Classifier	0.9143	0.7032	0.3607	0.383	0.3713	0.3254	0.3256	1.307
rf	Random Forest Classifier	0.9136	0.7567	0.3684	0.3799	0.3737	0.3273	0.3275	3.604
dt	Decision Tree Classifier	0.9134	0.662	0.3697	0.3795	0.3742	0.3277	0.3279	0.133

Table 4 represents the resultant values of the sentiment analysis for various sampling strategies. SMOTE and ADASYN are used as the oversampling techniques and ALLKNN and TomekLinks are used as the under-sampling techniques.

TABLE IV. Performance analysis of gradient boosting classifier without using hyper- parameters

Gradient Boosting Classifier		
	Sampling Technique	**Accuracy**
Over-Sampling	SMOTE	0.7735
	ADASYN	0.7193
Under-Sampling	ALLKNN	0.9338
	TomekLinks	0.9366

Table 5 shows the resultant values of the sentimental analysis of the given dataset after balancing and tuning the dataset and using the different sampling techniques. The sampling techniques used for over-sampling dataset are the SMOTE and ADASYN and the sampling technique used for fetching the resultant values for the under-sampling dataset are ALLKNN and Tomek Links.

TABLE V. Performance analysis of gradient boosting classifier with using hyper- parameters

Gradient Boosting Classifier		
	Sampling Technique	**Accuracy**
Over-Sampling	SMOTE	0.796
	ADASYN	0.7894
Under-Sampling	ALLKNN	0.9343
	TomekLinks	0.9368

Table 6 briefs us about the confusion matrix for the imbalanced dataset. In relation to the numbers of actual no and actual yes. Specifically, it displays the number of erroneous predictions as well as the number of true predictions. Both of their values are as:

True- Positives = 51

True- Negatives = 622

False- Positives = 7

False- Negatives = 8909

TABLE VI. Confusion matrix for imbalanced data

	Predicted No	**Predicted Yes**
Actual No	8909	7
Actual Yes	622	51

IX. Conclusion

Improving the accuracy of the model that has been suggested is the primary goal of the research that has been proposed. Consequently, in order to accomplish this goal, the presence of a balanced dataset is one of the primary prerequisites. This is because an imbalanced dataset can result in bias towards classes that have a greater number of samples. Every classifier has its unique set of benefits, which are determined by the kind of data that is used for training. It is seen that the results of the Gradient boosting classifier were superior to those of the other classifiers. The purpose of this endeavour is to employ a variety of optimization strategies aimed at selecting the most advantageous features, in conjunction with resampling techniques, in order to ensure that classifiers produce satisfactory outcomes. The first step is to select the five most accurate classifiers and nominate them for further consideration. Each model is assessed according to its level of accuracy. In conclusion, the Gradient boosting classifier beats the other classifiers in terms of accuracy, surpassing prior efforts that were performed with the same dataset by a margin of 93.62%. The work that will be done in the future can involve other datasets to work on.

References

[1] Agarwal, Apoorv, et al. "Sentiment analysis of twitter data." *Proceedings of the workshop on language in social media (LSM 2011)*. 2011.

[2] Wang, Yili, et al. "Sentiment analysis of Twitter data." *Applied Sciences* 12.22 (2022): 11775.

[3] Chalothom, Tawunrat, and Jeremy Ellman. "Simple approaches of sentiment analysis via ensemble learning." *information science and applications*. Springer Berlin Heidelberg, 2015.

[4] Xia, Rui, Chengqing Zong, and Shoushan Li. "Ensemble of feature sets and classification algorithms for sentiment classification." *Information sciences* 181.6 (2011): 1138-1152.

[5] Saif, Hassan, Yulan He, and Harith Alani. "Semantic sentiment analysis of twitter." *The Semantic Web–ISWC 2012: 11th International Semantic Web Conference, Boston, MA, USA, November 11-15, 2012, Proceedings, Part I 11*. Springer Berlin Heidelberg, 2012.

[6] Gautam, Geetika, and Divakar Yadav. "Sentiment analysis of twitter data using machine learning approaches and semantic analysis." *2014 Seventh international conference on contemporary computing (IC3)*. IEEE, 2014.

[7] Neelakandan, S., and D. Paulraj. "A gradient boosted decision tree-based sentiment classification of twitter data." *International Journal of Wavelets, Multiresolution and Information Processing* 18.04 (2020): 2050027.

[8] Le, Bac, and Huy Nguyen. "Twitter sentiment analysis using machine learning techniques." *Advanced Computational Methods for Knowledge Engineering: Proceedings of 3rd International Conference on Computer Science, Applied Mathematics and Applications-ICCSAMA 2015*. Springer International Publishing, 2015.

[9] Sahayak, Varsha, Vijaya Shete, and Apashabi Pathan. "Sentiment analysis on twitter data." *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* 2.1 (2015): 178-183.

[10] Neethu, M. S., and R. Rajasree. "Sentiment analysis in twitter using machine learning techniques." *2013 fourth international conference on computing, communications and networking technologies (ICCCNT)*. IEEE, 2013.

[11] Sankar, H., and V. Subramaniyaswamy. "Investigating sentiment analysis using machine learning approach." *2017 International conference on intelligent sustainable systems (ICISS)*. IEEE, 2017.

[12] Alsaeedi, Abdullah, and Mohammad Zubair Khan. "A study on sentiment analysis techniques of Twitter data." *International Journal of Advanced Computer Science and Applications* 10.2 (2019): 361-374.

[13] Kharde, Vishal, and Prof Sonawane. "Sentiment analysis of twitter data: a survey of techniques." *arXiv preprint arXiv:1601.06971* (2016).

[14] Anjaria, Malhar, and Ram Mohana Reddy Guddeti. "Influence factor based opinion mining of Twitter data using supervised learning." *2014 sixth international conference on communication systems and networks (COMSNETS)*. IEEE, 2014.

[15] Jain, Anuja P., and Padma Dandannavar. "Application of machine learning techniques to sentiment analysis." *2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*. IEEE, 2016.

[16] Sahayak, Varsha, Vijaya Shete, and Apashabi Pathan. "Sentiment analysis on twitter data." *International Journal of Innovative Research in Advanced Engineering (IJIRAE)* 2.1 (2015): 178-183.

[17] Da Silva, Nadia FF, Eduardo R. Hruschka, and Estevam R. Hruschka Jr. "Tweet sentiment analysis with classifier ensembles." *Decision support systems* 66 (2014): 170-179.

[18] Jagdale, Rajkumar S., Vishal S. Shirsat, and Sachin N. Deshmukh. "Sentiment analysis on product reviews using machine learning techniques." *Cognitive Informatics and Soft Computing: Proceeding of CISC 2017*. Springer Singapore, 2019.

[19] Singh, Jaspreet, Gurvinder Singh, and Rajinder Singh. "Optimization of sentiment analysis using machine learning classifiers." *Human-centric Computing and information Sciences* 7 (2017): 1-12.

[20] Ain, Qurat Tul, et al. "Sentiment analysis using deep learning techniques: a review." *International Journal of Advanced Computer Science and Applications* 8.6 (2017).

[21] Rathi, Megha, et al. "Sentiment analysis of tweets using machine learning approach." *2018 Eleventh international conference on contemporary computing (IC3)*. IEEE, 2018.

[22] Chugh, Bharti, and Nitin Malik. "Machine Learning Classifiers for Detecting Credit Card Fraudulent Transactions." *Information and Communication Technology for Competitive Strategies (ICTCS 2021) ICT: Applications and Social Interfaces*. Singapore: Springer Nature Singapore, 2022. 223-231.

[23] Qi, Yuxing, and Zahratu Shabrina. "Sentiment analysis using Twitter data: a comparative application of lexicon-and machine-learning-based approach." *Social Network Analysis and Mining* 13.1 (2023): 31.

[24] Pasrija, Paridhi, Utkarsh Singh, and Mehak Khurana. "Performance Analysis of Intrusion Detection System Using ML Techniques." *Applying Artificial Intelligence in Cybersecurity Analytics and Cyber Threat Detection* (2024): 135-150.