e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

SOCIAL MEDIA: COMPARATIVE ANALYSIS TOWARDS HATE SPEECH DETECTION USING SUPERVISED MACHINE LEARNING TECHNIQUES

Gurpreet Kaur*1, Er. Harjasdeep Singh*2, Er. Puja Das*3

*1Baba Farid College Of Engineering & Technology, Bathinda, Punjab, India.

*2Assistant Professor, Computer Science & Engineering MIMIT Malout, Punjab, India.

*3Assistant Professor, Baba Farid College Of Engineering And Technology, Punjab, India.

DOI : https://www.doi.org/10.56726/IRJMETS51819

ABSTRACT

People widely use social media applications to spend free time with their friends and relatives over long distances, and users generate vast amounts of data on social media applications where abusive speech has become a significant issue, so it is essential to identify the solution to hate speech detection. Researchers regularly work hard to build models that automatically detect hate speech context. In this research, we focus on abusive words posted on Twitter comment sessions by users. First, we will download the two datasets from Kaggle, "cyber-bullying "and "Twitter hate speech." Then, we will divide this dataset into training and testing and combine both datasets. Further, we performed data preprocessing and feature extraction. Moreover, we have also used different machine learning techniques such as logistic regression, KNN, naïve Byes, decision tree, and random forest for detecting abusive speech. Algorithms were shown different accuracies. We got the best accuracy from a random forest. It was 98.93%, rather than other algorithms. This research work will go ahead if we change the parameters of the datasets.

Keywords: Hate Speech Detection, Twitter, Supervised Machine Learning, User Post Comment, Automatic Detect Hate Speech Context.

I. INTRODUCTION

Hatred speech is a type of communication. It can be informed of symbols and writing speaking. When Individuals use this term to target a single group of people, religious sender disabilities are called hate speech. Every year, a wider group of people utilize social media. The most popular social media platforms used in the world are Instagram and Twitter. These applications are popular because of their services. People can share their opinions or thoughts on social media through different services such as audio, video, and writing content. However, people post vulgar posts on their accounts that hit the community's religious, gender, sexual orientation, and national origin. Hate speech is the biggest problem in the world. Every person has a personal account on social media applications. The accounts of social media app users are increasing day by day. Individuals share their through social media applications. Moreover, through Internet services and social media platforms, people can add known or unknown persons with the help of social media applications. The fact that everyone has the right to free expression is the root of this issue. To address the above-mentioned problem, all countries make laws to control hate speech communication as the constitution of India allocated laws IPC under parts that describe the spoken and written words that recommend language, community, religion, and ethnicity. The computer science branch also deals with hate speech problems by building automatic hate speech detection models with different languages, such as machine learning and deep learning. But sometimes, they misuse their freedom of speech, especially on social media platforms. Bangladesh holds the second position in Facebook accounts. People manage and create different groups of their products or ideas, apps from that, political, players, actors, and users the Facebook page for purpose, by studying many of Bangladesh's counter-terrorism and transnational crime cyber units. The report found that 2500 Facebook accounts spread hateful content [20]. At present, hate speech is a significant topic because everyone is using social media platforms. The authors tried to solve the issue of hate speech on social media with automatic detection models using natural processing language and took the dataset from the Crowd Flower site. They then applied feature extraction to the dataset to eliminate the information chatter. Afterward, they applied feature engineering to fetch the critical attribute of a dataset.[10] Moreover, most of the researchers use text information. However, this research work illustrates information on video-based material of hate speech. The authors deal with videos

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

on Twitter and classify them as hate speech and non-hate speech. They extracted audio from video and converted it into text format with the assistance of a speech-to-text converter [1].

The following is a description of the research's primary contributions:

1. Datasets information has collected from the Kaggle website that is based on Twitter commands.
2. Employing feature extraction and data preprocessing to increase the accuracy of the model.
3. The model has manufactured with text information. Then comparing the results of random forest, Naïve bays, decision tree, logistic regression and k-nearest neighbor.

II. LITERATURE REVIEW

1. Plaza-Del-Arco et al.[2021] The author proposed an approach to solve the problem of bad mouthing detection on mass media by using NP (natural processing) methods. They took a public dataset that was available on Crowd Flower and filtered it with the help of a text pre-processing method. After that, they applied the crucial attribute of the machine learning algorithm. They compared the accuracy of different algorithms on each feature. The authors applied the ensemble learned method to discover. Verbal attack material from internet community with the Ad boost attribute algorithm, they got the highest result 90.30% rather than another model.
2. Yuan, L.et al.[2023] The author represented a transfer learning technique to join two independent databases or make a single presentation of hateful content. They developed an accountable two-dimensional conception tool to build hate speech portrayal in which different databases can be compared and projected. Therefore, the similar description of the foul language class of the dataset was successfully projected into the same place and assisted in finding uncover to make fault. He reveals that joint walk boosts forecast performance when a small number is available. They proposed an article on verbal satisfaction that holds ample bad writing content. Further mane the second problem was represented through material-dependent tasks that made the difficult work for a multitude. He suggested a deep language system. It is a combination of recurrent layers or convolutions layer that automatically detects hateful content from social media. He tested this model on HASOC 0.63% in verbal communication from the HASOC corpus. The risk of over fitting is rising for this coherent. He innovated various approaches for increasing in size of recourse assist and showing different opportunists like labeled corpora, leveraging unlabeled data for further work. He expected their work to experiment with assists of the above models.
3. Boishakhi, F.T et al.[2021] The author repressed a multi-model that helped to detect the hate speech word from video or audio. The researcher supported various algorithms of machine learning equally, random forest, sum, logistic regression Ada book, naïve byes, k-NN, and decision tree. Afterwards, they applied above mention algorithm to datasets to identify the attributes of datasets such as audio, image and text.
4. Florio, K., et al.[2020] The writer demonstrated the BERT model that detects hate speech comments from the Italian language. They tested the original data “control 1 odio”. They implemented an experiment with a database. Firstly, they collected a single month of information from Twitter and examined it. Secondly, they took data from Twitter's modern part along the growing dimensional of ALBERT and SUM and the training set was in the same pattern. Furthermore, they identify we did not need to increase the size of the dataset to get similar accuracy. They tried to get more information about verbal change among monthly tests. Then, they applied temporal learning to statistics analysis to improve the performance of validity over time of prediction systems for abhorrent communication. They was seen information not continuous as we known that machine learning approaches were affected by different objects and they investigated on balancing method as well as adding new information to the model. So, they can observe to analyses the model.
5. Gangurde, et al.[2022] The goal of a writer was to issue a bibliometric analysis system or plot published work for verbal communication. The writer used the Scopus dataset by using VOS viewer, and science scope. They checked different parameters, Such as top journalists, trending topics, and document type, but bibliomaniac analysis observed the present published information on fault communication was intense on a certain topic. They got unexpected results from the bibliometric analysis system caused by recent happing in offensive speech in cybernetic. This model aims to show the current scenario of hate speech detection field finding on the internet.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

6. Nikhilraj Gadekar, et al.[2019] The fault language detection is the same as text classification jobs. They implemented the five baseline elements of model evaluation, classifiers selection and training, dimensionality reduction, data collection, exploration and feature extraction. The grants of trash talks are threefold. Firstly, they were introduced to require information of an automatic abusive talk's detection system. Secondly, they discussed the sickness and hardness of the model. Finally, some pause and open issues were recognized.
7. P urnama Sari Br Ginting et al.[2019] The writer proposed the multinomial logistic regression method of machine learning. With the support of the multinomial logistic model, the researcher endeavors to group the object of text data that influences the life of social media customers. The multinomial logistic regression approach helps to detect the fault content from the Nepal language. Moreover, they trained the different models of ML to tackle the issues of bad speech detection. They made metrics of attributes that showed the overall performance of the model and it gave advice to new investigators. The multinomial logistic regression gave the best performance attribute extraction model for future work; it will be a better model for verbal attacks.
8. Hang Thi-Thuy Do et al.[2020] The writer illustrated a model that works for hate speech detection issues on social media platforms. Especially, they evaluated the VLSP operation in 2019. He trained the model site and developed a system that forecasts label information into a comment session. In his documentation work, He developed the method in response to hateful, abusive remarks and postings by utilizing the assistance of bidirectional long short-term memory. He got 71.43% accuracy from the VLSP 2019 dataset. He examined how they can address hate speech detection with many ways to improve the performance of verbal communication by comparing ritual machine learning algorithms.
9. Rini et al.[2023] The author attempted a literature review of fault language via text mining. The writer classified it into different attributes such as sexism, racism, and offense, not only hate speech but also non-hate speech. The feature of hate speech was influenced by various data sets and classes of data sets. Furthermore, it was a case history, not a specific topic. The author experimented on Facebook, YouTube, Yahoo, Instagram, Wikipedia and other websites. We know that SVM is a much more useful algorithm for detecting bad comments from social media.
10. Ahlam Alrehili et al.[2023]Providing the natural language process method that assisted in detecting vulgar mouthing on operating system network approach, Likewise, bag-word, n-gram, by social media. Multitudes have the freedom to post any opinion, through mass media. But, in some situations, the social media user posts trash talks on social media accounts. That creates a bad environment for the public, due to this action of users. The crime cases of hate speech are increasing in the world. The investors recommended rules-based techniques, sentiment analysis, and dictionary template-based approaches to handle the problem of fault communication.

III. METHODOLOGY

3.1 Data collection:

In this research work, we have assisted the two datasets. We got these datasets from the Kaggle website namely "Twitter hate speech" and "cyber-bullying". Both datasets have two main attributes text or label. The labels are attributed in the form of zero and one. After that, we trained and tested both datasets. Moreover, above both training and testing datasets are merged.

3.2 Preprocessing of data:

It is a crucial process in DM (data mining). The primary goal of data processing is to enhance the quality of information, so we can use this data for suitable data-mining tasks.

3.4 Feature extraction:

We know that feature extraction is a very useful term in machine learning algorithms. Because ML approaches support the tabular form of a dataset. It is a process that develops modern variables by fetching from the nature data. The major goal it is to decrease the size of data so that we can use this reduced data for data modeling. It has many methods of principal components analysis, text analysis, and edge detection algorithms.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

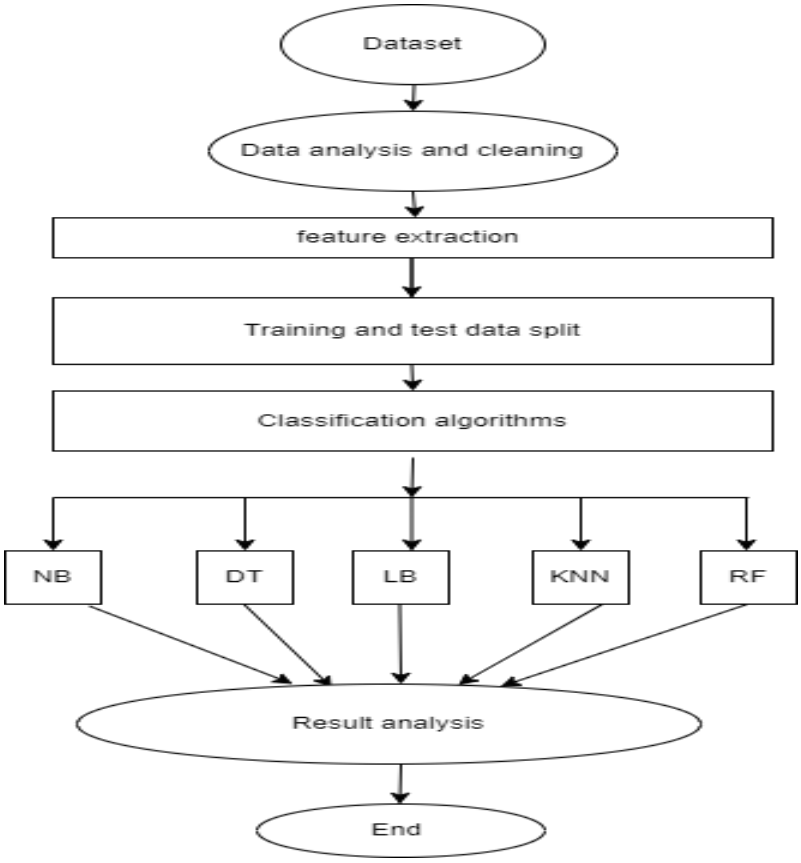

Figure 1: Methodology of proposed work

IV. ALGORITHMS USED IN THIS RESEARCH WORK

4.1 Naïve Bayes

Use the Bayes theorem to solve the classification problems. High dimensional training dataset is used in text classification which is mainly preferred with the help of fast machine learning is built to make forecasts quickly. This model is based on the probability of an object predicting spam filtering, classifying articles, and sentiment analysis are some naïve bays algorithms. It is converted dataset into a frequency table. The posterior probability is calculated by using Bay's theorem. It can be created like a hood table by finding probabilities of given features. It does better in multi-class predictions than other algorithms. If all aspects are regarded as distinct or unconnected, it is unable to understand the link between them. It is separated into three types similarly, multinomial Differential and Bernoulli.

$$P(A/B) = \frac{P(B/A).P(A)}{P(B)} \quad \ldots\ldots (1)$$

Posterior is the probability and probability (A/B) Probability (B/A), Probability of likelihood. It is a prior probability, the likelihood of A. Margin probability, or the likelihood of B, is represented by (B).

4.2 The Decision tree:

It is a flowchart-based supervised learning methodology. As the name represents, it works as a tree where nodes describe features of the dataset, branches as decision rules, and leaves the solving classification challenges in this decision tree; two types of nodes have many branches. They are used to make decisions, while the other is a leaf node resulting from the decision and does not divide further. The solution to the problem is in the form of a graphical representation. The CART algorithm is used to build the tree. It classifies the tree after the answer to the question and is easy to understand as it can copy the human thinking style while making decisions. Tree shape makes it easy to understand the logic behind it. After that, the selection measure is a technique to select the best attribute for the root model and the root. It has two types of information gain: the Gini index. The only drawback is complexity due to various layers, and over fitting can be solved using random forest algorithms.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

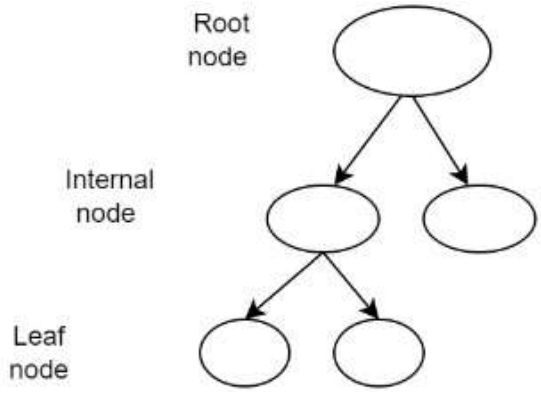

Figure 2: Decision tree

4.3 logistic regressions:

The most popular artificial intelligence technique is logistic Regression, which combines supervised training with independent variables. It anticipates a categorical dependent variable. For the most part, Logistic Regression is comparable to linear Regression; it yields a number between 0 and 1, sometimes referred to as a probabilistic value. Logistic Regression is meant to solve classification problems. However, Regression issues are dealt with by linear Regression. The logistic regression "S" shaped logistic function predicts the highest values. It can categorize fresh data as discrete and continuous datasets and provide probability. There are three logistic Regressions: nominal, ordinal, and Bernoulli. Using the linear regression equation We can get the logistic regression equation using various steps.

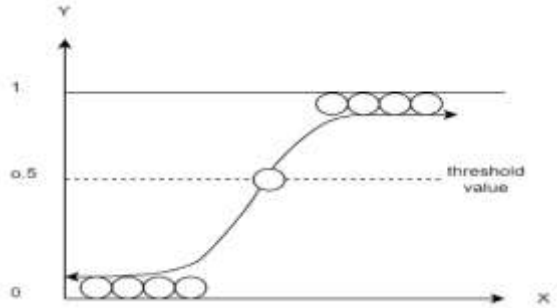

Figure 3: Logistic regression

4.4 K- K-Nearest Neighbor:

The most basic machine learning algorithm is the k-nearest Neighbor, a supervised learning method. It placed the latest instance in a group of cases that are pretty similar to each other. It can quickly classify the new data using the K-NN algorithm. It is mainly used for classification problems but can also do regression. It cannot predict underlying data, as K-NN is a non-parametric algorithm. It does not learn from the training set instant but uses old classification to sort new data, so it is also known as a lazy learner algorithm. The most challenging task or demerit is the high price of computation as the distance is significant between the data points for the entire training sample.

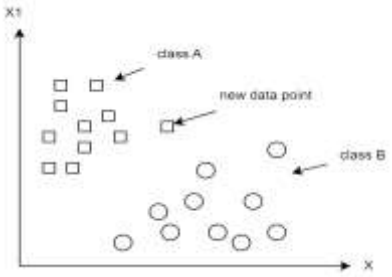

Figure 4: K-nearest neighbor

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

4.5 Random Forest:

Its algorithm is a supervised training method based on teaching difficult issue-solving through the combination of several classifiers and increasing the model's performance. This process is known as ensemble learning. Random forest takes overage of all numbers of decision trees to improve the prediction accuracy of the dataset. It is independent of a decision tree, with the help of a better understanding of decision tree algorithms. We can learn random forest algorithms. Random forest algorithms contain many trees, so only some may provide the correct result. It takes less training than other algorithms. Even for large datasets, the output is accurate, and any large ratio of data missing could not affect its result.

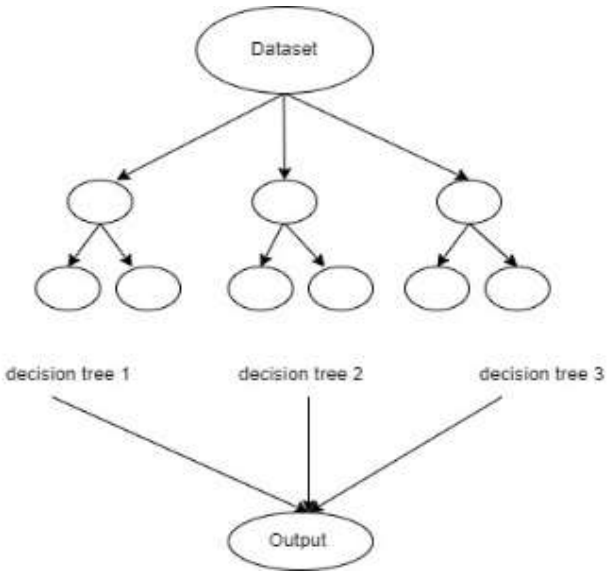

Figure 5: Random forest

Confusion matrix It's a matrix that shows how well the machine learning system performed when test data was gathered. It displays a model's performance as a 2*2 matrix with the four model values—TP (true positive), TN (true negative), FP (false positive), and FN (false negative) in it.

Predicted values	Actual values	
	TP	FN
	FP	TN

Figure 6: Confusion matrix

Accuracy: The most fundamental performance statistic is accuracy, which is calculated as the number of forecasts that are accurate divided by the total number of forecasts.

ROC CURVE: It illustrates the connection between the rate of genuine positives and the number of false positives for different attribute cut-off points**.** ROC curve is a pictorial plot that reveals the characteristics of the binary classifiers model.

True positive rate: It is the proportion of monitoring that, based on all positive tests, was accurately expected to be positive.

False negative rate: It is the proportion of monitoring that is erroneously expected to be positive tests.

$$\text{TPR} = \frac{\text{TR}}{\text{TR+FN}} \qquad \text{OR} \quad \text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$$

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

V. RESULTS AND DISCUSSION

We have implemented Decision trees, Logistic regression and Random Forest, K-NN, and Naïve Byes machine learning algorithms. These algorithms were utilized to develop the hate speech detection model and got different results. Firstly, we merged both datasets after training and testing. Secondly, data preprocessing was completed to enhance the performance of the model. Finally, feature extraction was done to reduce the volume of datasets.

5.1 Environment and parameters:

In this experiment, 70% of the samples were used for the training of different algorithms rest 30% was used for testing. In decision tree creation was set to entropy and the random state was 0. In the KNN algorithm used was five and metric use was Minkowski and p was 2. In the random forest algorithm n-estimators were 10, criterion = 'entropy' and the random state was zero. The experiment was performed in jupyter notebook. Sklearn library was used to implement the algorithm in the jupyter notebook.

Table 1: Result Analyses

ML Algorithms	Accuracy (%)
Random Forest	98.93
Decision Tree	97.26
Logistic regression	96.21
K-nearest neighbor	92.84
Naïve Bayes	80.19

Table 2: Training time of applied algorithms

Algorithms	Training Time (Sec)
Naïve Bayes	14.77 s
Decision Tree	414.80 s
K-nearest neighbor	03 s
Logistic regression	42.53 s
Random Forest	87.90 s

In this research work, we are comparing the diversified ML approaches to build the hate speech detection model and we got different accuracy of each approach. We have used the random forest, naïve bays, logistic regression, decision tree and k-nearest neighbors. These approaches have different accuracies of 98.93%, 97.26%, 96.21%, 92.84%, and 80.19%, respectively.

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

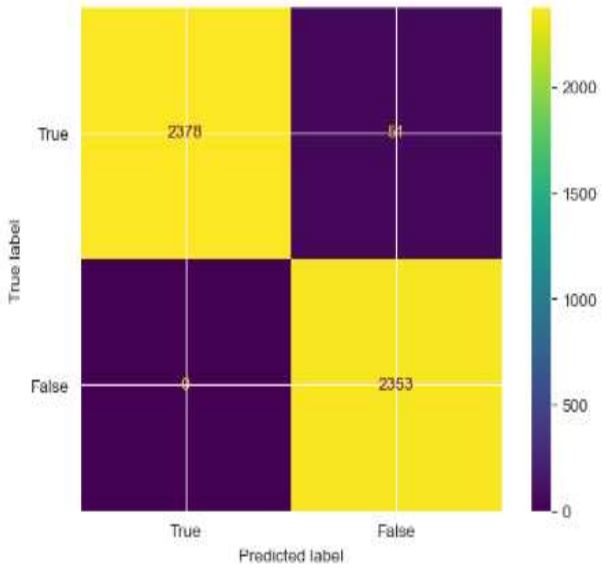

Figure 7: Random forest

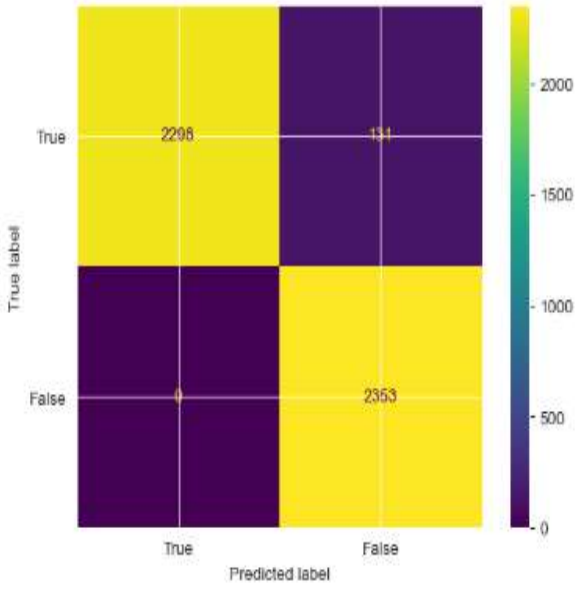

Figure 8: Decision tree

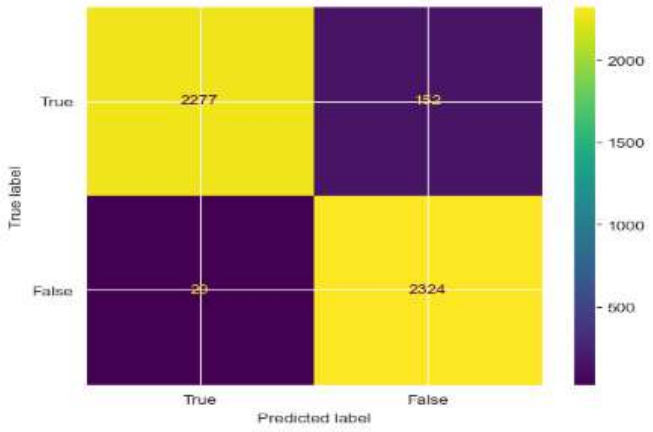

Figure 9: logistic regression

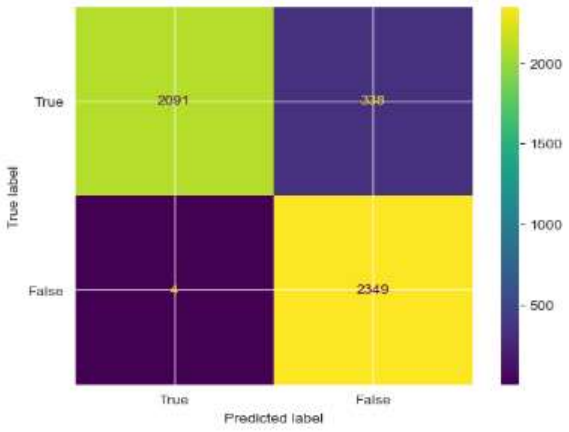

Figure 10: K-NN

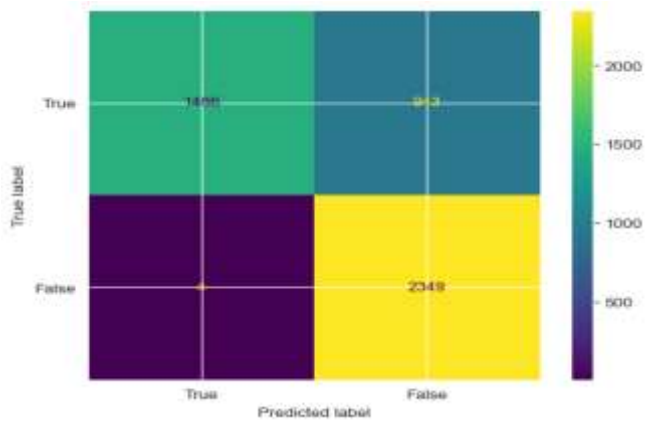

Figure 11: Naïve bayes

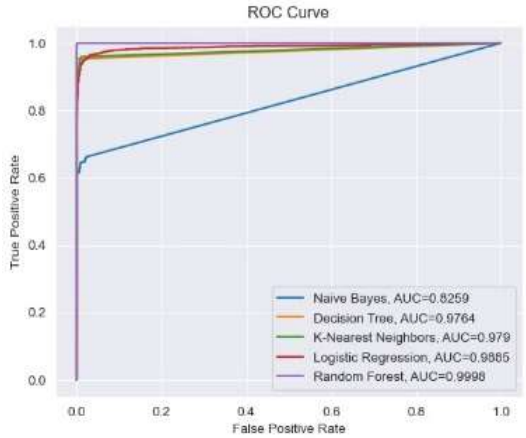

Figure 12: Roc curve

e-ISSN: 2582-5208
International Research Journal of Modernization in Engineering Technology and Science
(Peer-Reviewed, Open Access, Fully Refereed International Journal)
Volume:06/Issue:04/April-2024 Impact Factor- 7.868 www.irjmets.com

Figure 13: Non-Hate Comments

Figure 14: Hate Comments

VI. CONCLUSION

This research illustrated that machine learning algorithms help detect nasty comments on Twitter. These proposed approaches are implemented on two datasets: Twitter hate–speech and cyber-bullying, and around 1500 rows, two columns. By decreasing the dataset volume, we got the highest accuracy with the random forest at 98.93%, so we implemented these algorithms with the entire dataset. It takes a lot of time. These algorithms are pretty better than existing machine learning algorithms. The percentage of social media users is increasing every minute. Scientists are identifying the various methods to handle the problem of hate speech on internet media. This model will help deal with Twitter's problems. By using a hyper-tuning technique, we can increase the model accuracy.

VII. REFERENCE

[1] In 2021, Plaza-Del-Arco, F.M., Ureña-López, L.A., Molina-González, M.D., and Martín- Valdivia, M.T. a sentiment analysis-based multitask learning method for detecting hate speech. IEEE Access, 9, 112478–112489 pages.

[2] Yuan, L., Wang, T., Ferraro, G., Suominen, H. and Rizoiu, M.A., 2023. Transfer learning for hate speech detection in social media. Journal of Computational Social Science, 6(2), pp.1081-1101.

[3] Boishakhi, F.T., Shill, P.C. and Alam, M.G.R., 2021, December. Multi-modal hate speech detection using machine learning. In 2021 IEEE International Conference on Big Data (Big Data) (pp. 4496-4499). IEEE.

[4] Florio, K., Basile, V., Polignano, M., Basile, P. and Patti, V., 2020. Time of your hate: The challenge of time in hate speech detection on social media. Applied Sciences, 10(12), p.4180.

[5] Gangurde, A., Mankar, P., Chaudhari, D. and Pawar, A., 2022. A systematic bibliometric analysis of hate speech detection on social media sites. Journal of Scientometric Research, 11(1), pp.100-111.

[6] Nikhilraj Gadekaret ,2022,june, hate speech identification using machine learning, International Research Journal of Engineering and Technology, 9 ,(pp 3398 -34001).

[7] Ginting, P.S.B., Irawan, B. and Setianingsih, C., 2019, November. Hate speech detection on twitter using multinomial logistic regression classification method. In 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS) (pp. 105-111). IEEE.

[8] Do, H.T.T., Huynh, H.D., Van Nguyen, K., Nguyen, N.L.T. and Nguyen, A.G.T., 2019. Hate speech detection on vietnamese social media text using the bidirectional-lstm model. arXiv preprint arXiv:1911.03648.

[9] Rini, R., Utami, E. and Hartanto, A.D., 2020, October. Systematic literature review of hate speech detection with text mining. In 2020 2nd International Conference on Cybernetics and Intelligent System (ICORIS) (pp. 1-6). IEEE.

[10] Alrehili, A., 2019, November. Automatic hate speech detection on social media: A brief survey. In 2019 IEEE/ACS 16th International Conference on Computer Systems and Applications (AICCSA) (pp. 1-6). IEEE.