International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 3, pp 24-28, March-2023

https://doi.org/10.47001/IRJIET/2023.703004

Performance Analysis of Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest and SVM on Hate Speech Detection from Twitter

[1]*Subhajeet Das, [2]Koushikk Bhattacharyya, [3]Sonali Sarkar

[1,2]Department of Computer Science & Engineering, Swami Vivekananda Institute of Science & Technology, Kolkata, India
[3]Department of Chemical Engineering, Swami Vivekananda Institute of Science & Technology, Kolkata, India

***Abstract* - Hate speech specially racism, gender and religion discrimination, defaming comments are becoming one of the biggest problems in Twitter these days, that are making people to switch to other social media. Its effect is long-standing and unpreventable. To stop hateful activities from happening, Machine Learning approaches are needed to be applied. This research article focuses on the performance analysis and effectiveness of Logistic Regression, Gaussian Naive Bayes, K-Nearest Neighbor, Decision Tree, Random Forest and Support Vector Machine on detection of hate speech from Twitter. SVM, Decision Tree and Random Forest outperformed all the other models, achieving state-of-art 95.5%, 96.2% and 98.2% accuracy respectively on comments gather over a stretch.**

Keywords: Hate Speech, Logistic Regression, Decision Tree, Random Forest, K-Nearest Neighbor, Gaussian Naïve Bayes, Support Vector Machine, Count Vectorizer, One Hot Encoder, Precision, Recall, Accuracy.

I. INTRODUCTION

With the increasing amount of internet users, the usage of social media is also uprising. There are many popular social media that are used throughout the world very extensively, like – Twitter, LinkedIn, Facebook, Reddit, Quora, Instagram, Snapchat, Pinterest etc. This has both good and bad effects on the users, moreover on a large community of people or society. It helps in terms of connecting with people, sharing thoughts, keeping updated with new market trends, knowing recent changes in any community etc. But the bad effect is move prominent than that of the wise side. It includes posting uncensored contents like- images, videos, audios etc., posting hateful contents and replying with hateful comments, targeting any particular community like- LGBTQ, racial discrimination, targeting specific region or country, gender discrimination, spreading terrorist activities, defaming any people, organization and so on. Among all the social media, Twitter is mostly affected by these events. Hate comments cannot get removed from Twitter totally, but we can detect whether there is any hate comment are posted or not and after that any necessary action can be taken. To stop spreading hate in Twitter, many machine learning models are used to find out the extent of their efficiency.

In this paper, we are experimenting with many profound machine learning algorithms named Decision Tree, Gaussian Naïve Bayes, K-Nearest Neighbor, Logistic Regression, Random Forest and Support Vector Machine on a very large Twitter hate comments dataset which includes both positive and negative data. After that, we are measuring the efficiency and performance of these trained models to determine which machine learning algorithm is most effective in this task.

II. LITERATURE STUDY

Parisa Hajibabaee et al. proposed a text classification model comprising of a tokenizer, modular cleaner, three embedding methods and eight classifiers. They found out that Ada Boost, Multi-Layer Perceptron and SVM performed well only on Term Frequency-Inverse Document Frequency (TF-IDF) embedding [1]. Ruilin Xu et al. discovered most of the TF-IDF algorithm uses verbs and nouns as the elementary block and adverbs and adjectives as the secondary keyword which does not perform that well to define a query. So, they proposed a Part of Speech (POS) weighted TF-IDF algorithm to assign every term query frequency with a different value [2]. Chikashi Nobata et al. developed a machine learning approach which uses two domains to detect hate speech from online user comments over time [3]. Turki et al. used ensemble learning algorithms coupled with count vectorizer to detect hate speech and created a labeled dataset [4]. Zeerak Waseem et al. used critical race theory to annotate hate comment dataset and used character n-gram to detect hate speech [5]. Irene Kwok et al. suggested supervised machine learning approach to detect racist comments from Twitter [6]. Resham Ahluwalia et al. developed a machine learning model to find misogyny against women which they named Automatic Misogyny Identification (AMI) [7]. Zhi Xu et al. proposed a sentence level filter to remove offensive language from social media posted text messages [8]. Shofianina Dwi Ananda Putri

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 7, Issue 3, pp 24-28, March-2023
https://doi.org/10.47001/IRJIET/2023.703004

et al. used Naive Bayes, SVM and other machine learning approach to detect hate speech and offensive language from Sundanese and Javanese Indonesian local language [9]. Mila Putri Kartika Dewi et al. implemented a feature expansion system from comments using Word2Vec, Bag of Word (BOW) and TF-IDF[10]. Nurtheri Cahyana et al. proposed an automatic annotation system using K-Nearest Neighbor to annotate comments datasets easily [11]. William Warner et al. developed an approach to detect hate speech from online texts targeting religion, gender and sexual orientation [12]. Nikhil Chakravartula used multinomial naïve bayes algorithm to find the hate speech against women and immigrants [13]. Rong-En Fan et al. developed a library named LinLinear to perform large scaled linear classification tasks easily [14]. Nupur Khond et al. used Naïve Bayes algorithm to detect and hiding mechanism to hide hate comments from users [15]. Asogwa et al. used SVM and Naïve Bayes for detecting hate speech [16]. Purnama Sari Br Ginting et al. proposed the usage of multinomial logistic regression algorithm to detection [17]. Sean Mac Avaney et al. used Multiview SVM for detecting hate comments, minimizing the challenges faced by other machine learning algorithms [18].

III. PROPOSED METHOD

Step 1: Collection of appropriate datasets.

Step 2: Encoding the text data using Column Transformer and One Hot Encoder.

Step 3: Vectorizing the data using Count Vectorizer.

Step 4: Splitting the data into train and test dataset.

Step 5: Training the Gaussian Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor, Random Forestand Support Vector Machine with the train dataset.

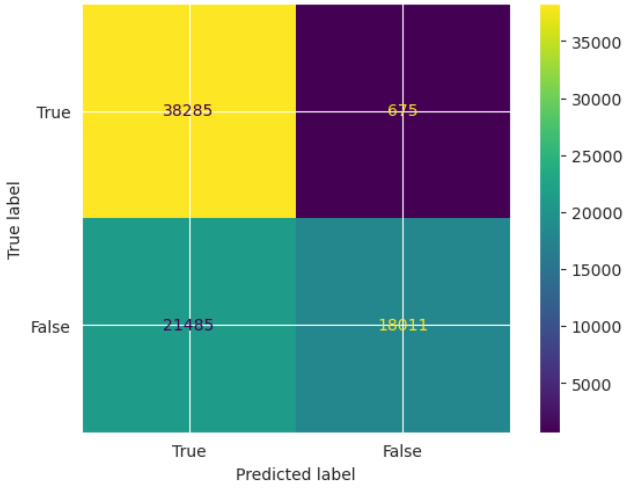
Step 6: Calculating the time taken to train the models.

Step 7: Testing the models with test datasets and generating the confusion matrix and performance measurement metrices.

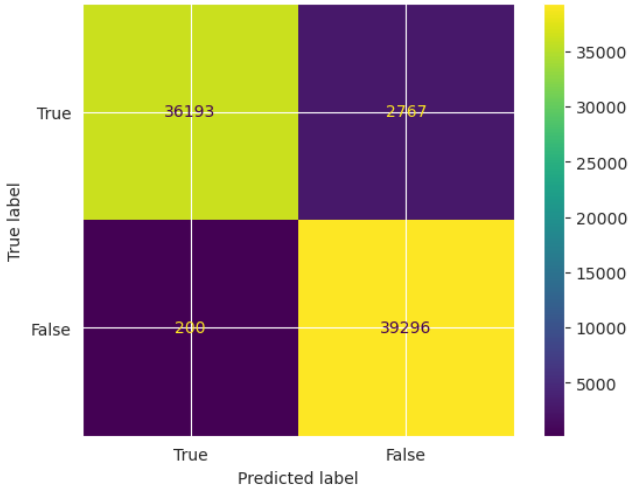
Step 8: Generating the Precision-Recall Curve and ROC Curve for further evaluation.

IV. PERFORMANCE MEASUREMENT

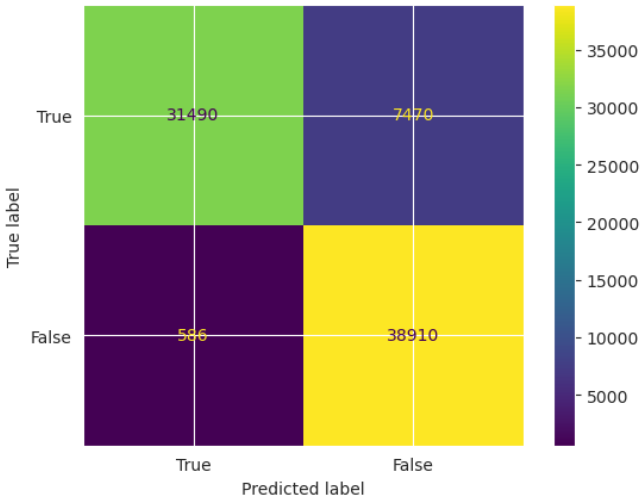
Confusion Matrix is a very useful metric to evaluate the performance of the trained models. It summarizes the model's performance by calculating the relation between the actual data with predicted values returning the True Positive (TP), False Positive (FP), True Negative (TN) and False Negative (FN) values in its four cells. It is also used to calculate Precision, Recall, Accuracy, F1 Score of the models.



(a)



(b)



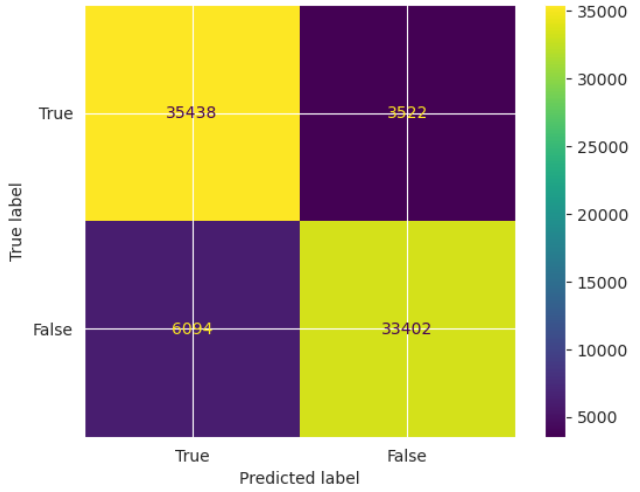
(c)

International Research Journal of Innovations in Engineering and Technology (IRJIET)

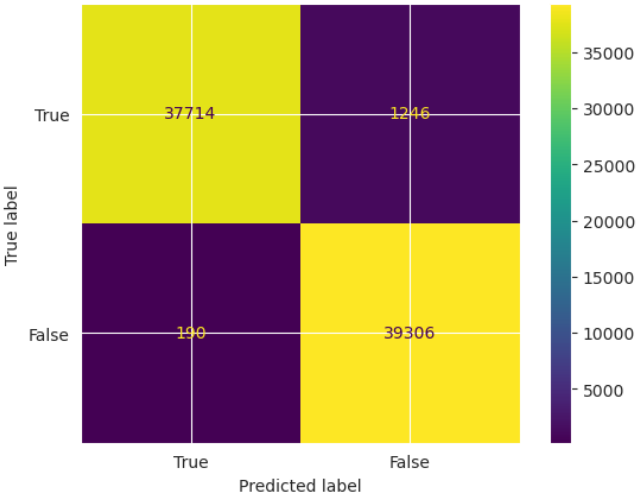
ISSN (online): 2581-3048

Volume 7, Issue 3, pp 24-28, March-2023

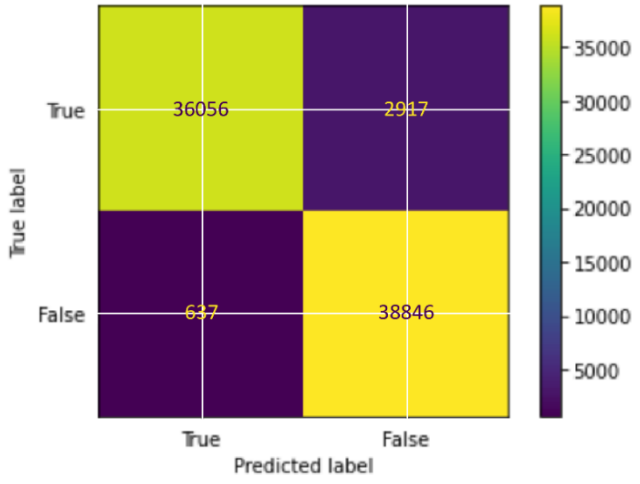
https://doi.org/10.47001/IRJIET/2023.703004



(d)



(e)



(f)

Figure 1:Confusion Matrix of the models trained with (a) Naive Bayes, (b) Decision Tree, (c) K-Nearest Neighbors, (d) Logistic Regression, (e) Random Forest and (f) Support Vector Machine

Precision can be defined as the ratio between True Positive (TP) values out of all predicted positive values i.e., True Positive (TP) and False Positive (FP) values.

$$Precision = \frac{True\ \ Positive}{(True\ \ Positive + False\ \ Positive\ )} \quad (1)$$

Recall or, True Positive Rate (TPR) can be said as the fraction of the True Positive (TP) values out of all the actual positive values i.e., True Positive (TP) and False Negative (FN) values.

$$Recall\ or\ TPR = \frac{True\ \ Positive}{(True\ \ Positive + False\ \ Negative\ )} \quad (2)$$

Precision-Recall curve is the graphical representation of the relation between the Precision and Recall of the models. The Precision of the model is plotted in the Y-axis and the Recall in the X-axis.

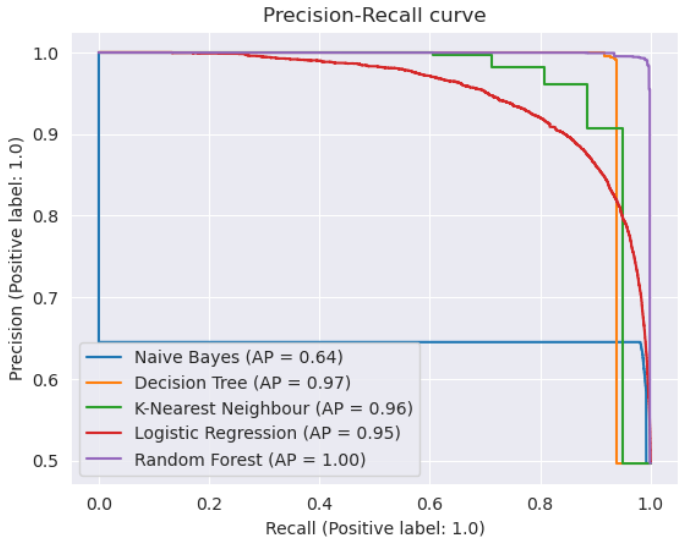


Figure 2: Precision-Recall Curve of the trained models

False Positive Rate (FPR) is defined as the ration of the negative values that are incorrectly predicted positive values i.e., False Positive (FP) to all the actual negative values i.e., True Negative (TN) and False Positive (FP).

$$False\ Positive\ Rate = \frac{False\ \ Positive}{(False\ \ Positive + True\ \ Negative\ )} \quad (3)$$

Receiver Operating Characteristic (ROC) curve graphically shows the relation between the Recall or True Positive Rate (TPR) with the False Positive Rate (FPR). TPR is plotted along Y-axis and FPR is plotted along X-axis.

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 7, Issue 3, pp 24-28, March-2023
https://doi.org/10.47001/IRJIET/2023.703004

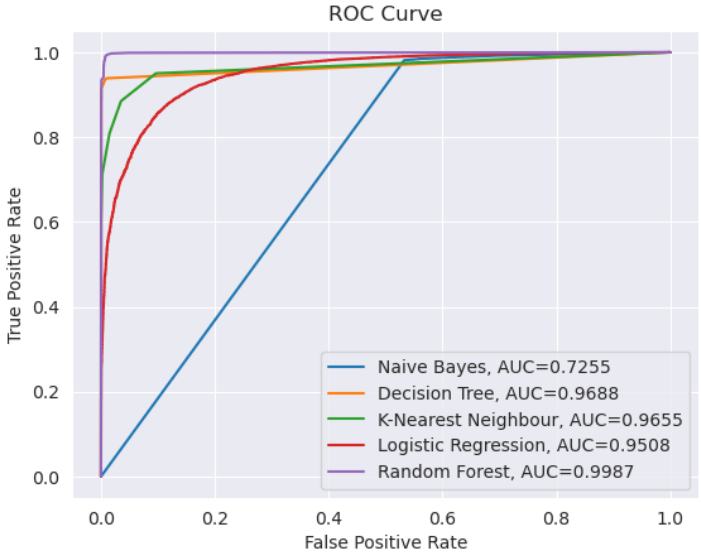


Figure 3: ROC Curve of the trained models

V. RESULTS

After training the models with the training dataset and testing them with test dataset, these are the results that are observed:

Table 1: Observed results for evaluation

Model	Naive Bayes	Decision Tree	K-Nearest Neighbor	Logistic Regression	Random Forest	Support Vector Machine
Precision	0.6405	0.9945	0.9817	0.8533	0.9950	0.9826
Recall or TPR	0.9827	0.9290	0.8083	0.9096	0.9688	0.9252
Accuracy	0.7175	0.9622	0.8973	0.8774	0.9821	0.9547
F1 Score	0.7755	0.9606	0.8866	0.8805	0.9817	0.9530
FPR	0.5440	0.0051	0.0148	0.1543	0.0048	0.0161
Training Time	6.4880s	238.2752s	0.0151s	33.2732s	47.7207s	15810.6213s

VI. CONCLUSION

In this paper, we tried to find out the application of Gaussian Naïve Bayes, Decision Tree, K-Nearest Neighbor, Logistic Regression and Random Forest in the field of hate speech detection from Twitter. It is observed that out of all these models, Support Vector Machine, Decision Tree and Random Forest achieved state-of-art 95.5%, 96.2% and 98.2% accuracy respectively at finding the hidden meaning inside the large number of comments and therefore determining whether there is any hateful event is going on or not. In future, the aim will be to make these models more efficient so that it can work on other social media.

REFERENCES

[1] Hajibabaee, Parisa, Masoud Malekzadeh, Mohsen Ahmadi, Maryam Heidari, Armin Esmaeilzadeh, Reyhaneh Abdolazimi, and H. James Jr. "Offensive language detection on social media based on text classification." In 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC), pp. 0092-0098. IEEE, 2022.

[2] Xu, Ruilin. "POS weighted TF-IDF algorithm and its application for an MOOC search engine." In 2014 International Conference on Audio, Language and Image Processing, pp. 868-873. IEEE, 2014.

[3] Nobata, Chikashi, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. "Abusive language detection in online user content." In Proceedings of the 25th international conference on World Wide Web, pp. 145-153. 2016.

[4] Turki, Turki, and Sanjiban Sekhar Roy. "Novel Hate Speech Detection Using Word Cloud Visualization and Ensemble Learning Coupled with Count Vectorizer." Applied Sciences 12, no. 13 (2022): 6611.

[5] Waseem, Zeerak, and Dirk Hovy. "Hateful symbols or hateful people? predictive features for hate speech

ISSN (online): 2581-3048
Volume 7, Issue 3, pp 24-28, March-2023
https://doi.org/10.47001/IRJIET/2023.703004

detection on twitter." In Proceedings of the NAACL student research workshop, pp. 88-93. 2016.

[6] Kwok, Irene, and Yuzhou Wang. "Locate the hate: Detecting tweets against blacks." In Proceedings of the AAAI Conference on Artificial Intelligence, vol. 27, no. 1, pp. 1621-1622. 2013.

[7] Ahluwalia, Resham, Himani Soni, Edward Callow, Anderson Nascimento, and Martine De Cock. "Detecting hate speech against women in english tweets." EVALITA Evaluation of NLP and Speech Tools for Italian 12 (2018): 194.

[8] Xu, Zhi, and Sencun Zhu. "Filtering offensive language in online communities using grammatical relations." In Proceedings of the Seventh Annual Collaboration, Electronic Messaging, Anti-Abuse and Spam Conference, pp. 1-10. 2010.

[9] Putri, Shofianina Dwi Ananda, Muhammad OkkyIbrohim, and Indra Budi. "Abusive language and hate speech detection for javanese and sundanese languages in tweets: Dataset and preliminary study." In 2021 11th International Workshop on Computer Science and Engineering, WCSE 2021, pp. 461-465. International Workshop on Computer Science and Engineering (WCSE), 2021.

[10] Dewi, Mila Putri Kartika, and Erwin Budi Setiawan. "Feature Expansion Using Word2vec for Hate Speech Detection on Indonesian Twitter with Classification Using SVM and Random Forest." Jurnal Media Informatika Budidarma 6, no. 2 (2022): 979-988.

[11] Cahyana, Nur Heri, Shoffan Saifullah, YuliFauziah, Agus Sasmito Aribowo, and Rafal Drezewski. "Semi-supervised Text Annotation for Hate Speech Detection using K-Nearest Neighbors and Term Frequency-Inverse Document Frequency." Int. J. Adv. Comput. Sci. Appl 13, no. 10 (2022): 147-151.

[12] Warner, William, and Julia Hirschberg. "Detecting hate speech on the world wide web." In Proceedings of the second workshop on language in social media, pp. 19-26. 2012.

[13] Chakravartula, Nikhil. "HATEMINER at SemEval-2019 task 5: hate speech detection against immigrants and women in Twitter using a multinomial naive Bayes classifier." In Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 404-408. 2019.

[14] Fan, Rong-En, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. "LIBLINEAR: A library for large linear classification." the Journal of machine Learning research 9 (2008): 1871-1874.

[15] Khond, Nupur, Godawari Padwal, Veena Ulgekar, Tejaswini Parsekar, and Sumit Harale. A Preventive Measure on Hate Speech Detection On Online Social Network using Naïve Bayes. No. 2967. EasyChair, 2020.

[16] Asogwa, Doris Chinedu, Chiamaka Ijeoma Chukwuneke, C. C. Ngene, and G. N. Anigbogu. "Hate Speech Classification Using SVM and Naive BAYES." arXiv preprint arXiv:2204.07057 (2022).

[17] Ginting, Purnama Sari Br, BudhiIrawan, and Casi Setianingsih. "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, 2019.

[18] MacAvaney, Sean, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. "Hate speech detection: Challenges and solutions." PloS one 14, no. 8 (2019): e0221152.

Citation of this Article:

Subhajeet Das, Koushikk Bhattacharyya, Sonali Sarkar, "Performance Analysis of Logistic Regression, Naive Bayes, KNN, Decision Tree, Random Forest and SVM on Hate Speech Detection from Twitter" Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 3, pp 24-28, March 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.703004

****

