Suresh Gyan Vihar University Journal of Engineering & Technology
(An International Bi-Annual Journal), Volume 9, Issue 1, 2023, pp 16-23
ISSN: 2395-0196

Hasoc19: Hate Speech Detection on Multimodal Dataset

Archika Jain[1], Dr. Sandhya Sharma[2]
[1]Research Scholar, SGVU, Jaipur, India
[2]Professor, SGVU, Jaipur, India

Abstract: In this paper, we employed a multimodal dataset called hasoc19 in several languages, including English, Hindi, and German. We used a variety of machine learning classifiers on these datasets, including multinomial naive Bayes, k-nearest neighbor, Gaussian naive Bayes, logistic regression, decision tree, and random forest. Using k-nearest neighbor and logistic regression on the hasoc19 English dataset, we get the maximum accuracy of 0.66. The greatest accuracy on the hasoc19 Hindi dataset is 0.74 when using Gaussian naive Bayes, and the highest accuracy on the hasoc19 German dataset is 0.92 when using k-nearest neighbor classifiers.

Keywords– Machine Learning, Accuracy, Hate Speech, Hasoc19

I. INTRODUCTION

Contents are shared by social media is a very common thing now a days. People commonly use social media to communicate their thoughts, opinions, and observations. Despite the fact that social media is extremely popular due to its quick growth. Social networking is a fast, open, free, and easy means to communicate. Its nature is also rather vulnerable. It becomes a platform for wrongdoers to spread various sorts of hatred or prejudice through communication with a different group. Hate speech is forbidden because it is primarily a dialogue that may be extremely damaging to an individual's or group's sentiments and may contribute to violence or insensitivity that exhibits unreasonable and inhuman behavior. While these websites provide an open forum for people to debate and share ideas and perspectives, their nature and the large amount of posts, comments, and communications exchanged makes it nearly difficult to monitor their content. Because of diverse origins, customs, and beliefs, many people use angry and abusive words while conversing with others from different backgrounds [1].

Major contribution of this paper is:

- Work on multimodal datasets like English, Hindi and German. These datasets are from hasoc19 dataset.
- Predicted the accuracy for all the datasets.
- Applied machine learning classifiers to all the hasoc19 datasets.

This paper contains “Related Work” in section II. In section III we will explain the “Proposed Work”. In section IV we will elaborate the “Experimental Setup” and “Conclusion & Future Scope” are covered in section V.

II. RELATED WORK

The word-based technique is the most basic method for detecting offensive/abusive content on social media, however it is insufficient for detecting inappropriate/offensive speech by a person.

Muhammad Sajjad et-al (2019) used deep learning algorithm for utilizing to extract features, which were then combined with related syntactic and n-gram features before being trained and predicted using a basic baseline classifier (SVM, LR, RF) [2]. Muhammad U. S. Khan et-al (2021) used CNN techniques for both multiclass and multilevel classification is sufficiently encouraging and indicates the possibility of these approaches for hate speech categorization on social media [3]. Bhavesh Pariyani et-al (2021) acquired the optimal parameter for the machine learning model, certain preprocessing procedures and grid search were employed. SVM using TF-IDF performs best after preprocessing and utilizing grid search, with 0.7488 F1 Score and 0.9668 Accuracy Score [4]. Angela Marpaung et-al (2021) revealed Bi-GRU approach with Indo-BERT and no stop word removal gets the greatest accuracy of 84.77% [5]. Noor Azeera Abdul Aziz et-al (2021) Using the Twitter dataset, the tests are carried out by taking into account the combination of word n-gram and improved syntactic n-gram. Filter-embedded combining feature selection is used to minimize the feature set. The experimental findings show that combining word n-gram and improved syntactic n-gram with feature selection to categories the data into three classes: hate speech, offensive language, or neither can produce satisfactory results. The accuracy score is 91%, as are the precision, recall, and F1 averages [6].

III. PROPOSED WORK

The main motive of this work is firstly we classify the data into four category that is none, hate, profane and offensive. After that for evaluation categorize this data into two classes that is hated or nan-hated.

A Data Collection

For this work, we have collected Hasoc19 dataset in different languages from Kaggle. The sizes of English, Hindi and German datasets are 5852 tweets, 4665 tweets and 3819 tweets. Firstly we categorize our tweets into four category like none, hate, offensive (OFFN) and profane (PRFN). Table 1 shows the Hasoc19 dataset statistics. In which English dataset are categorize into 3591 none, 1143 hate, 667 profane and 451 offensive. Hindi dataset shows that 2196 none, 556 hate, 1237 profane and 676 offensive. And in German dataset shows that 3412 none, 111 hate, 86 profane and 210 offensive speech. Further we classify our data into two categories like class 0 for non-hated and class 1 for hated. In class 0 we have taken none category and in class 1 we have taken hate, profane and offensive categories, it is showing in table 2.

INDEX-1 HASOC19 DATASET STATICS

Category	**No. of Speech**		
	English dataset	**Hindi dataset**	**German dataset**
None	3591	2196	3412
Hate	1143	556	111
PRFN	667	1237	86
OFFN	451	676	210

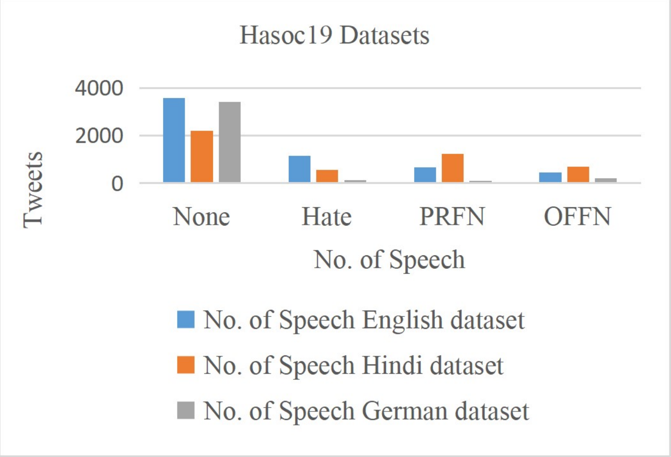


Fig. 1. Graphical representation of Hasoc19 datasets

Figure 1 shows that graphical representation of hasoc19 dataset in different languages like English, Hindi and German. In this we categorize the speech in four category none, hate, profane and offensive.

INDEX-2 NO. OF SPEECH IN HASOC19 DATASET

Clas s	**Meanin g**	**No. of Speech**		
		Englis h Datase t	**Hindi Datase t**	**Germa n Dataset**
0	Non-hated speech	3591	2196	3412
1	Hated speech	2261	2469	407

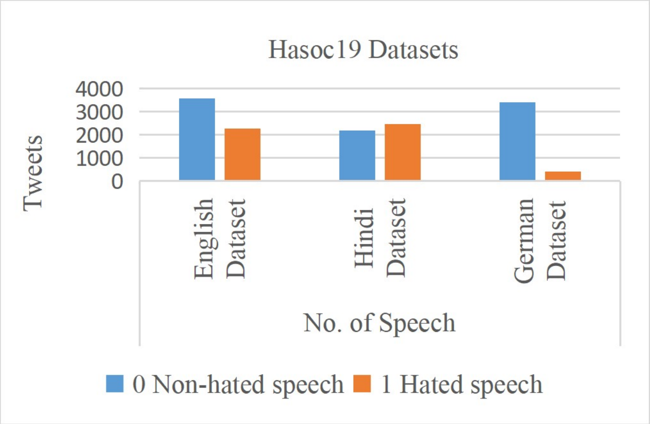


Fig. 2. Graphical representation of Hasoc19 datasets according to Class

Figure 2 shows the graphical representation of hasoc19 dataset in different languages like English, Hindi and German. In this categorize our data in two class 0 and 1. Class 0 is known as non-hated speech and class 1 is known as hated speech.

B. Machine Learning Classifiers

In this paper we used many machine learning models for predicting hated speech.

- ***Multinomial naïve Bayes:*** Multinomial Naive Bayes is a probabilistic learning approach commonly used in Natural Language Processing (NLP). It is extremely beneficial when used to a multinomial distributed dataset.
- ***K-nearest neighbour (KNN):*** The k-nearest neighbours (KNN) technique is a basic, easy-to-implement supervised machine learning algorithm that may be used to handle both classification and regression issues.
- ***Gaussian naive Bayes (GNB):*** Gaussian Naive Bayes (GNB) is a classification technique used in Machine Learning (ML) that is based on the probabilistic approach and Gaussian distribution. Gaussian Naive Bayes presupposes that each parameter (also called features or predictors) has an independent capacity of predicting the output variable.
- ***Logistic regression (LR):*** Logistic Regression may be used to categories the observations using different forms of data and can readily discover the most efficient variables utilized for the classification.
- ***Decision tree (DT):*** The purpose of utilizing a Choice Tree is to develop a training model that can use to predict the class or value of the target variable by learning basic decision rules inferred from past data(training data) (training data).
- ***Random forest (RF):*** Random Forest is a common machine learning algorithm that belongs to the supervised learning approach. It may be utilized for both Classification and Regression tasks in ML. It is based on the notion of ensemble learning, which is a method of integrating numerous classifiers to solve a complicated issue and to enhance the performance of the model.

IV. EXPERIMENTAL SETUP

In this we used hasoc19 datasets in different languages like English, Hindi and German that is available on Keggle. For accuracy prediction we used machine learning classifiers. Table 3 shows the accuracy prediction on the basis of different machine learning classifiers.

INDEX-3 ACCURACY PREDICTION

ML Classifiers	**Accuracy**		
	English Dataset	**Hindi Dataset**	**German Dataset**
Multinomial NB	0.64	0.74	0.87
KNN	0.66	0.59	0.92
GNB	0.53	0.65	0.66
LR	0.66	0.73	0.90
DT	0.61	0.67	0.86
RF	0.65	0.71	0.91

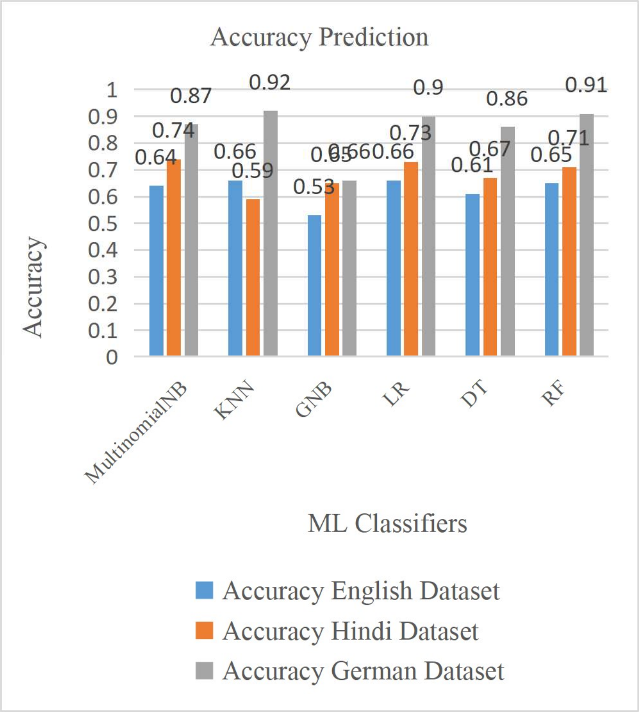


Fig. 3. Graphical representation of accuracy prediction of Hasoc19 datasets in different languages

Figure 3 has shown that accuracy prediction of hasoc19 dataset in different languages by applying different machine learning classifiers.

A. Confusion matrix for Hasoc19 English dataset

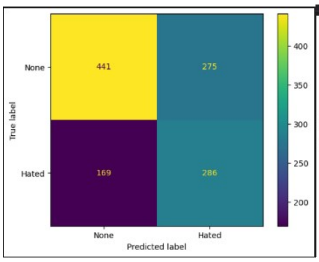


Fig. 4. Multinomial NB confusion matrix

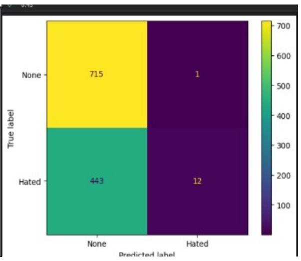


Fig. 5. KNN confusion matrix

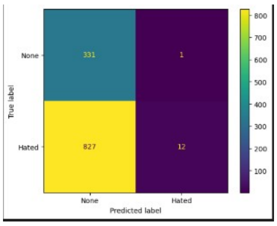


Fig. 6. GNB confusion matrix

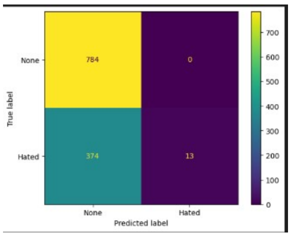


Fig. 7. LR confusion matrix

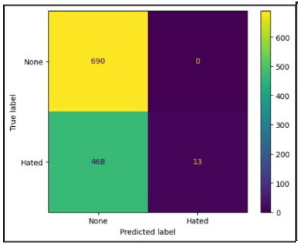


Fig. 8. DT confusion matrix

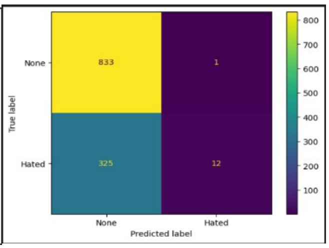


Fig. 9. RF confusion matrix

Figure 4, Figure 5, Figure 6, Figure7, Figure8 and Figure 9 represent the confusion matrix values for hasoc19 English dataset by applying different ml classifiers.

Figure 4 shows 727 true predictions and 444 wrong predictions when using multinomial NB model. Figure 5 shows 727 true predictions and 444 wrong prediction by using KNN model. Figure 6 shows 343 true predictions and 828 wrong predictions when using GNB model. Figure 7 shows 797 true predictions and 374 wrong prediction by using LR model. Figure 8 shows 703 true predictions and 468 wrong predictions when using decision tree model. Figure 9 shows 845 true predictions and 326 wrong prediction by using RF model.

B. Confusion matrix for Hasoc19 Hindi dataset

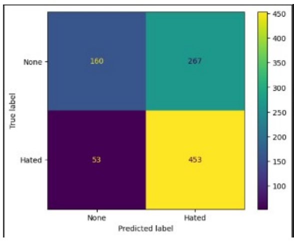


Fig. 10. Multinomial NB confusion matrix

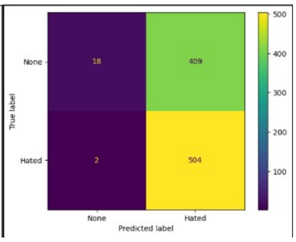


Fig. 11. KNN confusion matrix

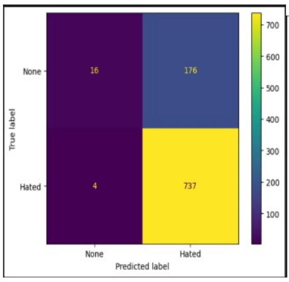


Fig. 12. GNB confusion matrix

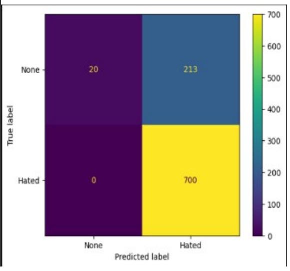
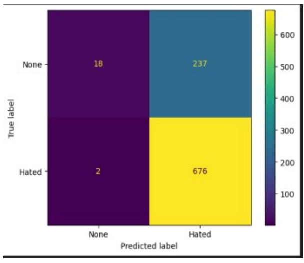


Fig. 13. LR confusion matrix



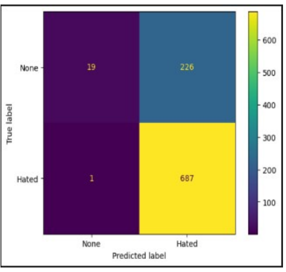


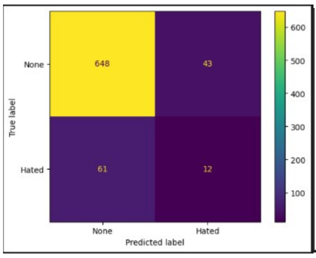
Fig. 14. DT confusion matrix

Fig. 15. RF confusion matrix

Figure 10, Figure 11, Figure 12, Figure 13, Figure 14 and Figure 15 represent the confusion matrix values for hasoc19 Hindi dataset by applying different ml classifiers.

Figure 10 shows 613 true predictions and 320 wrong predictions when using multinomial NB model. Figure 11 shows 522 true predictions and 411 wrong prediction by using KNN model. Figure 12 shows 753 true predictions and 180 wrong predictions when using GNB model. Figure 13 shows 720 true predictions and 213 wrong prediction by using LR model. Figure 14 shows 694 true predictions and 239 wrong predictions when using decision tree model. Figure 15 shows 706 true predictions and 227 wrong prediction by using RF model.

C. Confusion matrix for Hasoc19 German dataset



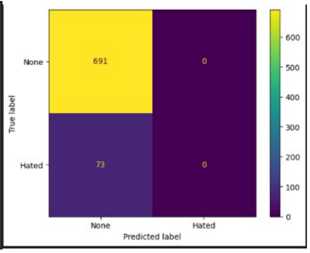
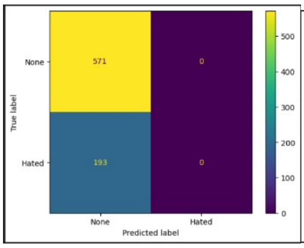


Fig. 16. Multinomial NB confusion matrix

Fig. 17. KNN confusion matrix



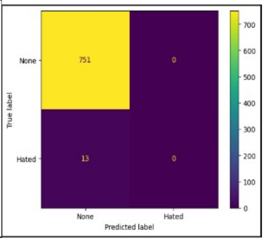
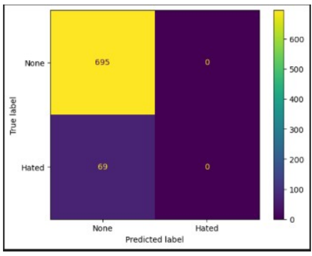


Fig. 18. GNB confusion matrix

Fig. 19. LR confusion matrix



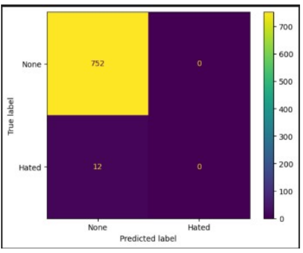


Fig. 20. DT confusion matrix

Fig. 21. RF confusion matrix

Figure 16, Figure 17, Figure 18, Figure 19, Figure 20 and Figure 21 represent the confusion matrix values for hasoc19 German dataset by applying different ml classifiers.

Figure 16 shows 660 true predictions and 104 wrong predictions when using multinomial NB model. Figure 17 shows 691 true predictions and 73 wrong prediction by using KNN model. Figure 18 shows 571 true predictions and 193 wrong predictions when using GNB model. Figure 19 shows 751 true predictions and 13 wrong prediction by using LR model. Figure 20 shows 695 true predictions and 69 wrong predictions when using decision tree model. Figure 21 shows 752 true predictions and 12 wrong prediction by using RF model.

V. CONCLUSION & FUTURE SCOPE

In this paper, we used multimodal dataset named as hasoc19 in different languages like English, Hindi and German. On these datasets we have applied different machine learning classifiers like multinomial naive Bayes, k-nearest neighbor, Gaussian naive Bayes, logistic regression, decision tree and random forest. On hasoc19 English dataset we get highest accuracy 0.66 by using k-nearest neighbor and logistic regression. On hasoc19 Hindi dataset we get highest accuracy 0.74 by using Gaussian naive Bayes and on hasoc19 German dataset we get highest accuracy 0.92 by using k-nearest neighbor classifiers.

In future, we will apply linguistic features on these datasets and find precision, recall and f1-score values.

REFERENCES

[1] H. Watanabe, M. Bouazizi, T. Ohtsuki, "Hate speech on Twitter: a pragmatic

approach to collect hateful and offensive expressions and perform hate speech detection," IEEE Access, vol. 6, 2018.

[2] M. Sajjad, F. Zulifqar, M. U. G. Khan, & M. Azeem, "Hate speech detection using fusion approach" In 2019 International Conference on Applied and Engineering Mathematics (ICAEM), pp. 251-255.

[3] M. Khan, A. Abbas, A. Rehman and R. Nawaz, "Hate classify: a service framework for hate speech identification on social media" in IEEE Internet Computing, vol. 25, pp. 40-49, 2021.

[4] B. Pariyani, K. Shah, M. Shah, T. Vyas and S. Degadwala, "Hate speech detection in Twitter using natural language processing," Third International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV), pp. 1146-1152, 2021.

[5] Marpaung, R. Rismala and H. Nurrahmi, "Hate speech detection in Indonesian Twitter texts using bidirectional gated recurrent unit," 13th International Conference on Knowledge and Smart Technology (KST), pp. 186-190, 2021.

[6] N. A. Abdul Aziz, M. Aizaini Maarof and A. Zainal, "Hate speech and offensive language detection: a new feature set with filter-embedded combining feature selection," 2021 3rd International Cyber Resilience Conference (CRC), pp. 1-6, 2021.

[7] J. Sachdeva, K. K. Chaudhary, H. Madaan and P. Meel, "Text based hate-speech analysis," 2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS), pp. 661-668, 2021.

[8] T. Febriana, & A. Budiarto, "Twitter dataset for hate speech and cyberbullying detection in Indonesian language," in International Conference on Information Management and Technology (ICIMTech) vol. 1, pp. 379-382.

[9] H. Şahi, Y. Kılıç and R. B. Saglam, "Automated detection of hate speech towards woman on Twitter" 2018 3rd International Conference on Computer Science and Engineering (UBMK), pp. 533-536, 2018.

[10] H. Watanabe, M. Bouazizi and T. Ohtsuki, "Hate speech on Twitter: a pragmatic approach to collect hateful and offensive expressions and perform hate speech detection," in IEEE Access, vol. 6, pp. 13825-13835, 2018.

[11] U. A. N. Rohmawati, S. W. Sihwi and D. E. Cahyani, "SEMAR: an interface for Indonesian hate speech detection using machine learning," International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 646-651, 2018.

[12] S. W. A. M. D. Samarasinghe, R. G. N. Meegama and M. Punchimudiyanse, "Machine learning approach for the detection of hate speech in Sinhala Unicode text," 20th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 65-70, 2020.

[13] G. Koushik, K. Rajeswari and S. K. Muthusamy, "Automated hate speech detection on Twitter," 5th International Conference On Computing, Communication, Control And Automation (ICCUBEA), pp. 1-4. 2019.

[14] N. D. T. Ruwandika and A. R. Weerasinghe, "Identification of hate speech in social media," 18th International Conference on Advances in ICT for Emerging Regions (ICTer), pp. 273-278, 2018.