A Truncated SVD Framework for Online Hate Speech Detection on the ETHOS Dataset

Anusha Chhabra
Department of Information Technology
Biometric Research Laboratory
Delhi Technological University
Delhi-110042, India
anusha.chhabra@gmail.com

Dinesh Kumar Vishwakarma
Department of Information Technology
Biometric Research Laboratory
Delhi Technological University
Delhi-110042, India
dinesh@dtu.ac.in

2023 4th International Conference on Innovative Trends in Information Technology (ICITIIT) | 978-1-6654-9414-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICITIIT57246.2023.10068574

***Abstract*—Hate content on social media is currently one of the most significant risks, where the victim is either a single individual or a group of people. In the current scenario, online web platforms are one of the most prominent ways to contribute to an individual's opinions and thoughts. Free sharing of ideas on an event or situation also bulks on the web. Information sharing is sometimes a bane for society if primarily used platforms are utilized with some lousy intention to spread hatred for intentionally creating chaos/ confusion among the public. Users take this as an opportunity to spread hate to get some monetary benefits, the detection of which is of paramount importance. This article utilizes the concept of truncated singular value decomposition (SVD) for detecting hate content on the ETHOS (Binary-Label) dataset. Compared with the baseline results, our framework has performed better in various machine learning algorithms like SVM, Logistic Regression, XGBoost, and Random Forest.**

Keywords*—*Hate Speech, Machine Learning, SVD, Binary-label Classification, TF-IDF

I. Introduction

There has been substantial usage of social media platforms by more people and exponential growth in the data. People share their thoughts and views on almost everything without considering the impact on society. According to statistics, Twitter is the most usable platform having nearly 340 million active users [1] and about 200 million tweets per year. The mentioned statistics and many users are also flooding hate content. Therefore, identifying hate content is a very prominent research area. Hate content can be defined as controversial, attacking group characteristics based on religion, gender, ethnicity, etc. Fig *1* shows that a leader is porting a divisive statement targeting those who raise their voices against CAA, NRC, and NPR [2]. Perhaps, Major social media platforms are curbing hate content at an initial stage. Still, hate content is sowing its roots almost in every form of content characteristics.

To improve the binary classification of social media texts, researchers and practitioners are paying more attention to the upcoming techniques of machine learning and deep learning. Considerable efforts have been spent on creating new and practical features that better classify hate speech on social media [3], [4], [5]. In addition, the challenges related to specific hate content detection lie in the need for more guidelines, benchmarks [6], and the non-availability of multimodal datasets. This paper presents a framework for identifying hate content on the ETHOS dataset.

The Major contributions of this manuscript are:

- Training the models on one dataset and cross-validation is done on another dataset which is approximately 24 times greater.
- To show the vulnerability of a small dataset with another related large dataset.

The rest of the paper is organized as follows: Section II provides an overview of the recent works on hate content detection using unsupervised machine learning approaches.

Section III illustrates the framework to detect hate content, followed by the discussion of the experimental results in Section IV. The conclusion and further scope are discussed in Section V.

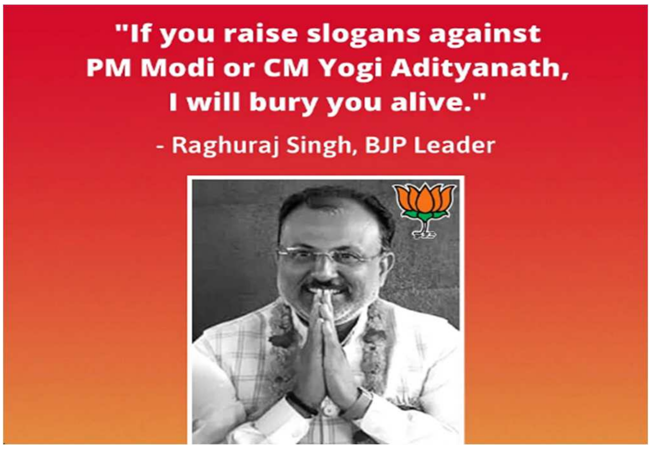
"If you raise slogans against
PM Modi or CM Yogi Adityanath,
I will bury you alive."
- Raghuraj Singh, BJP Leader

Fig 1. Example of Hate Speech

II. Related Work

Identifying hate content is crucial for millions of users to have freedom of expression. Authors and Academicians are focusing on multimodal, multilingual, and multiclass hate speech detection using supervised, unsupervised and semi-supervised machine learning techniques.

Machine learning has played its role very well in the last two decades. Specifically for hate speech and offensive language detection, Naïve Bayes, Support Vector Machine, Logistic Regression, Random Forest Decision Tree, and ensemble techniques are used as machine learning classifiers[7] for hate speech detection. The work in this area is found to be done in various languages. The same probabilistic and predictive analysis techniques are used by [8] for hate speech detection in Indonesian languages. [9] applies a supervised SVM technique for racist text classification. It is also observed that by ignoring the word-order sequence, BoW showed better accuracy in text classification. To overcome the limitation of BoW, Researchers perform N-gram approaches[10]. Manual labeling of large data is a time-consuming task that leads to the requirement of an unsupervised method. It takes advantage of detecting hate speech in a huge stream of data. Authors in [11] used Kohonen maps for the detection of cyberbullying, claiming an accuracy of 72%. PCA is also another class to

978-1-6654-9414-4/23/$31.00 ©2023 IEEE

detect violent or non-violent tweets. The authors examined the classification rate using various machine-learning algorithms after analyzing the sentence and language features[13].

The benefit of manual labeling is that it is effective for domain-specific tasks, but its drawback is execution time. In order to train and test the classifiers, the authors employed unsupervised approach for dimensionality reduction [12], applying k-means clustering and assigning each cluster as one human-annotated data from the Twitter dataset using supervised machine learning text classifiers. Twitter data is categorized into hateful and antagonistic labels using a Bayesian logistic regression model, according to [7]. The effectiveness of several supervised approaches for hate speech identification is evaluated and contrasted by authors [6] with an emphasis on South Asian languages. labeled and unlabeled data can be used by semi-supervised learning systems. Labeling data pertaining to unlabeled data can effectively increase productivity. [14] analyzed that supervised learning can sufficiently catch small-scale events whereas unsupervised learning has limited ability to deal with limited-scale events; yet, the necessity to manually label the data lowers down the models scalability.

III. Dataset Description

ETHOS dataset contains ~1K comments from YouTube and Reddit validated through a Figure-Eight Platform. This dataset repository has two csv files: Binary and Multi- Label. Out of 998 comments in binary file, 565 are manually labeled as non-hate while rest of them are labeled as hate. For Multi-label file, 8 labels are created as violence (if it incites (1) or not (0) violence), directed vs general (if it is directed to a person (1) or a group (0)), and 6 labels about the category of hate speech like gender, race, national origin, disability, religion and sexual orientation.

IV. Proposed Methodology

Hate speech is now a threat to society, affecting the dignity of an individual, unity, and nation. Many hate words are used alternatively. Fig 2 shows the word cloud for hate speech explicitly generated from [3]. Therefore, Eliminating and classifying hate content over social platforms is crucial and requires an hour.

Data preprocessing is a component of data preparation. Several techniques are used to normalize the data before it is fed into any machine learning or AI development pipeline. Fig 3 represents the broader approach used in the classification. Tokenization is the initial step in any NLP pipeline, used to break unstructured data into chunks of discrete values. Then, stemming is used as a normalized technique in which tokenized words are converted into short words to remove redundancies. Finally, the cleaned data is used to create a dictionary for key: value pairs. TF-IDF is used for mapping words to vectors of real numbers then a vector matrix is given as an input to classify as hate or non-hate.

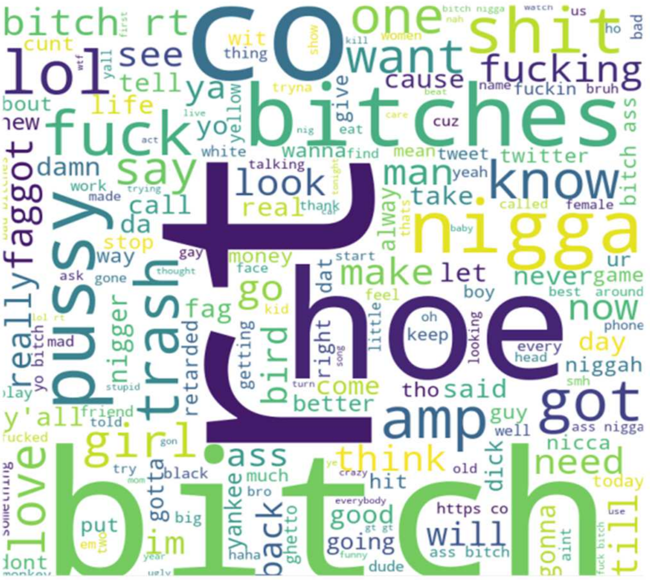

Fig 2. Word Cloud

Error! Reference source not found. represents the flowchart adopted for implementation. In the process flow, tweets are preprocessed in the first stage. Second stage implements TF-IDF, to quantify the words. Truncated SVD is used for dimensionality reduction for simplifying the calculations. Hyper parameter tuning such as L1 regularization is also done for logistic regression, XGBoost and SVM. Finally, the Prediction is done using various machine learning algorithms like Logistic Regression, Random Forest, Support Vector Machines, and XGBoost.

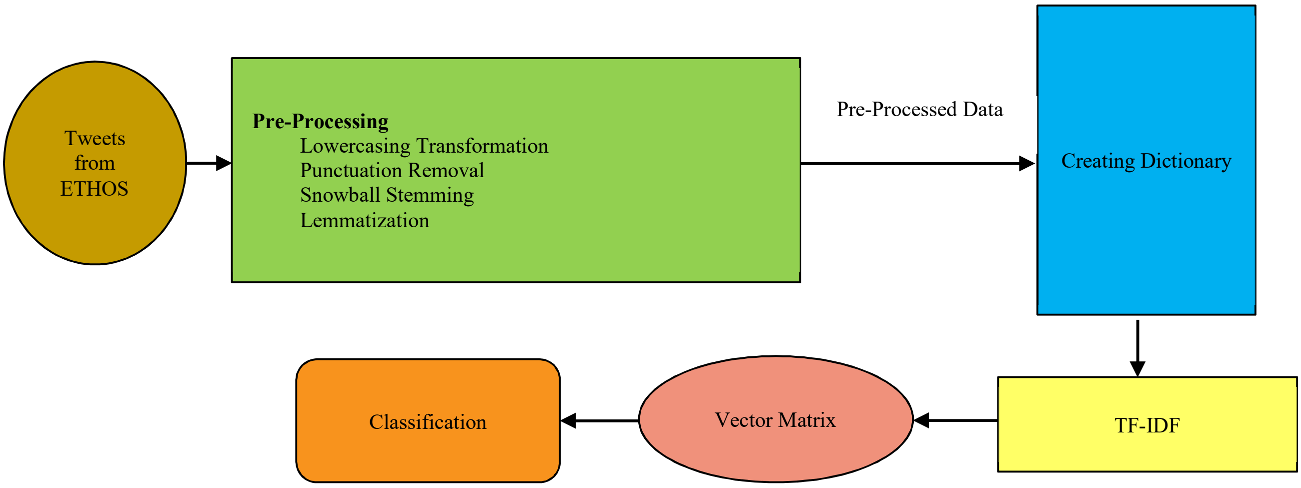

Fig 3. Data Pre Processing Steps

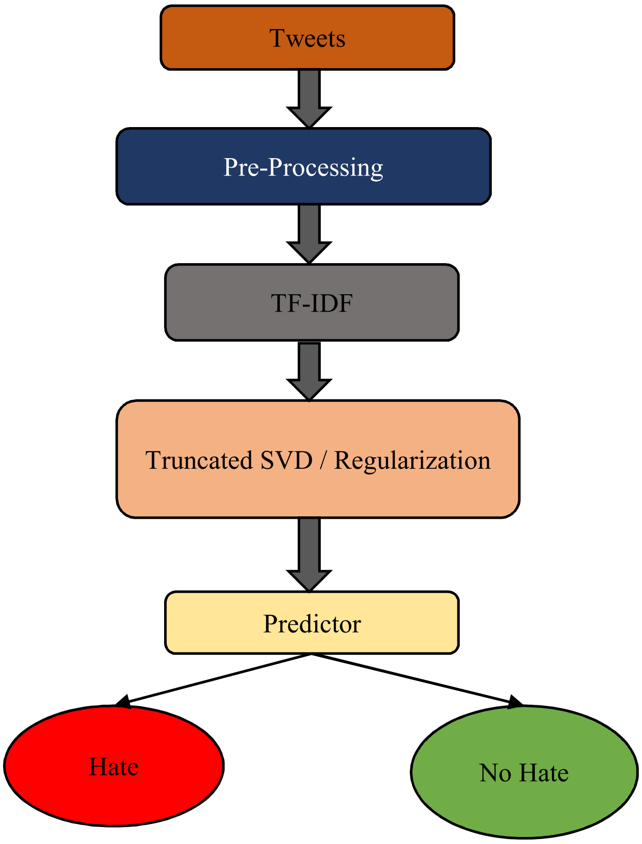

Fig 4. Process Flow

V. Experimental Setup

Although the dataset size is very small. To prove that a dataset of higher quality is more useful than the larger datasets, we have considered a dataset D1[3] which is approximately 24 times greater than ETHOS. In this experiment, we train various machine learning models with default parameters on the ETHOS dataset and compare the results with D1 dataset. The results are compared in terms of F1 score and balanced accuracy.

F1 score (Eq 1) is defined as the combination of precision and recall of a classifier into a single metric by considering their harmonic mean.

$$F1 = \frac{2(Precision*Recall)}{Precision+Recall} \tag{1}$$

Table I F1 Scores of ETHOS and D1 from SVM, LR, RF, XGBoost

Models	ETHOS			D1		
	F1 Score	F1 Score (Hate)	F1 Score (No Hate)	F1 Score	F1 Score (Hate)	F1 Score (No Hate)
SVM	67.71	59.60	73.63	75.47	12.86	79.30
LR	69.13	60.84	75.27	78.76	14.89	82.67
RF	67.01	58.85	73.03	67.21	12.73	70.55
XGBoost	65.30	54.50	73.44	75.39	10.62	79.35

Table I shows the results in the form of overall F1 scores, F1 Score (Hate) and F1 score (No Hate) of four machine learning models implemented on ETHOS and D1 datasets. The results are obtained when the models are trained on ETHOS and cross validation is done on D1 dataset.

Balanced accuracy (Eq 2) is defined as the arithmetic mean of sensitivity and specificity. It is also considered as the further development in standard accuracy metric.

$$Balanced\ Accuracy = \frac{Specificity+Sensitivity}{2} \tag{2}$$

Balanced Accuracies are shown in the Table II representing that our proposed approach using truncated SVD and hyper parameter tuning gives better results than baseline results. The graphical representation of balanced accuracies are shown in Fig 4.

Table II Comparison Table_Balanced Accuracy

Balanced Accuracy		
Models	**ETHOS_Our Approach**	**ETHOS_Baseline**
SVM	***66.70***	66.43
LR	***67.07***	66.94
RF	***68.17***	65.04
XGBoost	***64.42***	64.33

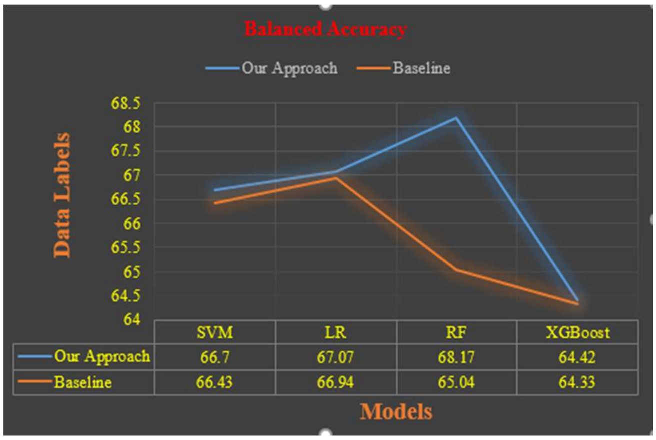

Fig 4. Comparison Graph of Balanced Accuracy: ETHOS_BINARY (Our Approach vs Baseline)

VI. Conclusion & Future Scope

From the empirical evaluation done in the paper, it is seen that reducing features using Truncated SVD along with hyper parameter tuning helped in increasing balanced accuracy and F1 score for algorithms like Logistic Regression, SVM and XGBoost when compared to the baseline results. For Random Forest, only change in hyper parameter is giving good results. The paper covers the basic ML algorithms for detecting hate speech. So, more SOTA algorithms and ensemble techniques can be implemented as a future task. Moreover, ETHOS dataset can be combined with other similar datasets for more evaluations.

References

[1] J clement, "https://www.statista.com/statistics/282087/number-of-monthly-active-twitter-users/," *Number of monthly active Twitter users worldwide*, 2021. .

[2] M. Bose, "https://www.thequint.com/news/politics/senior-bjp-leaders-giving-india-a-free-tutorial-in-hate-speech#read-more," *The Quint*, 2020. .

[3] T. Davidson, D. Warmsley, M. Macy, and I. Weber, "Automated hate speech detection and the problem of offensive language,"

Proc. 11th Int. Conf. Web Soc. Media, ICWSM 2017, pp. 512–515, 2017.

[4] Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” pp. 88–93, 2016, doi: 10.18653/v1/n16-2013.

[5] P. Burnap and M. L. Williams, “Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making,” *Policy and Internet*, vol. 7, no. 2, pp. 223–242, 2015, doi: 10.1002/poi3.85.

[6] M. M. Khan, K. Shahzad, and M. K. Malik, “Hate Speech Detection in Roman Urdu,” *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 20, no. 1, pp. 1–19, 2021, doi: 10.1145/3414524.

[7] P. Burnap and M. Williams, “Hate Speech, Machine Classification and Statistical Modelling of Information Flows on Twitter: Interpretation and Communication for Policy Decision Making,” in *Internet, Policy & Politics*, 2014, pp. 1–18, [Online]. Available: http://orca.cf.ac.uk/id/eprint/65227%0A.

[8] I. Alfina, R. Mulia, M. I. Fanany, and Y. Ekanata, “Hate speech detection in the Indonesian language: A dataset and preliminary study,” in *2017 International Conference on Advanced Computer Science and Information Systems, ICACSIS 2017*, 2018, vol. 2018-Janua, pp. 233–237, doi: 10.1109/ICACSIS.2017.8355039.

[9] E. Greevy and A. F. Smeaton, “Classifying racist texts using a support vector machine,” *Proc. Sheff. SIGIR - Twenty-Seventh Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr.*, no. January 2004, pp. 468–469, 2004, doi: 10.1145/1008992.1009074.

[10] W. B. Cavnar, J. M. Trenkle, and A. A. Mi, “N-Gram-Based Text Categorization,” *Proc. SDAIR-94, 3rd Annu. Symp. Doc. Anal. Inf. Retr.*, pp. 161–175, 1994, [Online]. Available: http://www.let.rug.nl/~vannoord/TextCat/textcat.pdf.

[11] M. Di Capua, E. Di Nardo, and A. Petrosino, “Unsupervised cyber bullying detection in social networks,” *Proc. - Int. Conf. Pattern Recognit.*, vol. 0, pp. 432–437, 2016, doi: 10.1109/ICPR.2016.7899672.

[12] K. E Abdelfatah, G. Terejanu, and A. A Alhelbawy, “Unsupervised Detection of Violent Content in Arabic Social Media,” pp. 01–07, 2017, doi: 10.5121/csit.2017.70401.

[13] Y. Chen, Y. Zhou, S. Zhu, and H. Xu, “Detecting offensive language in social media to protect adolescent online safety,” 2012.

[14] V. Tech, C. Lu, H. J. H. I. I. I. College, and F. Chen, “STED : Semi-Supervised Targeted-Interest Event Detection.”