International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 7 Issue 5, September-October 2023 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Machine Learning Approach to Classify Twitter Hate Speech

Subrata Saha[1], Md. Motinur Rahman[2], Md. Mahbub Alam[3]

[1,2]Scientific Officer, Institute of Electronics,
[3]Senior Scientific Officer, Institute of Computer Science,
[1, 2, 3]Bangladesh Atomic Energy Commission, Dhaka, Bangladesh

ABSTRACT

In this modern age, social media platforms have become indispensable tools for communication and information sharing. However, this unprecedented connectivity has also given rise to a concerning proliferation of hate speech and offensive content. This research article presents a comprehensive study on the development and evaluation of machine learning (ML) models for the automatic detection of hate speech on Twitter. We leverage a diverse dataset collected from Twitter, encompassing a wide range of hate speech categories, including hate speech targeting race, gender, religion, and more. To address the multifaceted nature of hate speech, we employ a hybrid approach that combines traditional natural language processing (NLP) techniques with state-of-the-art machine learning algorithms. Our methodology involves extensive preprocessing of the text data, including tokenization, stemming, and feature extraction. We then experiment with various machine learning algorithms, including Naïve Bayes (NB), K-nearest Neighbor (KNN), Random Forest (RF), and Support Vector Machines (SVM). The models are trained and fine-tuned on a labeled dataset and evaluated using robust metrics to assess their performance.

KEYWORDS: *Hate speech classification, cyberbullying, NLP, Machine learning*

How to cite this paper: Subrata Saha | Md. Motinur Rahman | Md. Mahbub Alam "Machine Learning Approach to Classify Twitter Hate Speech" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-7 | Issue-5, October 2023, pp.165-169, URL: www.ijtsrd.com/papers/ijtsrd59873.pdf

IJTSRD59873

I. INTRODUCTION

Social media platforms, particularly Twitter, have emerged as powerful tools for communication, information dissemination, and social interaction in the digital era. They offer a platform for individuals and communities to express their thoughts, engage in discussions, and connect with a global audience. However, this unprecedented connectivity has also brought to light a deeply concerning issue - the widespread dissemination of hate speech and offensive content [1].

Hate speech on Twitter, characterized by its derogatory, discriminatory, or threatening nature, poses serious challenges to the principles of online civility, inclusivity, and safety. Instances of hate speech targeting race, gender, religion, ethnicity, and other personal characteristics have become distressingly common, leading to real-world consequences, including cyberbullying, harassment, and the exacerbation of social tensions. [2]

The exponential growth of social media platforms and the sheer volume of content posted daily make manual moderation an impractical and insufficient solution. The need for automated tools to detect and mitigate hate speech has never been more critical. Machine learning, with its ability to analyze vast datasets and identify patterns, offers a promising avenue for addressing this complex problem. [3]

This research article presents a comprehensive study on the development and evaluation of machine learning models for the automatic detection of hate speech on Twitter. We recognize the urgency of this task in fostering safer and more inclusive online environments, as well as in protecting vulnerable individuals and groups from the harms of online hate.

Our research employs a multifaceted approach to address the challenges posed by hate speech on Twitter. We explore our dataset encompassing various hate speech categories and apply a hybrid methodology that combines traditional natural language processing (NLP) techniques with cutting-edge machine learning algorithms. The complete process in shown in Fig 1.

The subsequent sections of this paper provide some related studies, and a detailed account of our research methodology, experimentation, results, and conclusions. By leveraging machine learning and data-driven insights, our aim is to contribute to the ongoing efforts to combat the proliferation of hate speech on social media platforms, ultimately fostering a more inclusive and respectful online community.

II. Related works

Research in the domain of hate speech detection on social media platforms, particularly Twitter, has gained considerable attention in recent years. Several studies have explored various approaches and methodologies to tackle this pressing issue. In this section, we review key related works and their contributions to the field of hate speech detection, providing valuable insights and context for our research.

Davidson et al. conducted a seminal study on hate speech detection, focusing on the development of machine learning models to identify offensive language and hate speech on Twitter. Their work laid the foundation for subsequent research in this area and highlighted the challenges of distinguishing hate speech from offensive language [4].

Waseem et al. investigated the predictive features for hate speech detection, emphasizing the importance of not only textual content but also user behavior and interaction patterns. Their work underscores the complexity of the hate speech detection task and the need for a holistic approach. [5]

SemEval-2019 Task 5 challenge aimed to advance hate speech detection by focusing on multilingual content targeting specific demographics. Participants in this challenge developed and evaluated models for detecting hate speech against immigrants and women, addressing the need for language diversity and inclusivity [6].

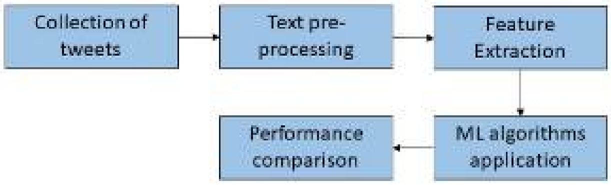


Fig 1 Block diagram of the complete process

In their study, Davidson et al. presented a comprehensive exploration of automated hate speech detection on Twitter. They investigated the challenges of distinguishing between offensive language and genuine hate speech and offered valuable insights into the detection of online hate speech. [7]

Zhang et al. focused on the use of deep learning techniques, specifically Convolutional Neural Networks (CNNs) and Gated Recurrent Units (GRUs), for hate speech detection on Twitter. The study explored the effectiveness of deep learning in addressing this challenging task. [8]

Founta et al. presented a large-scale crowdsourcing effort to characterize abusive behavior on Twitter. Their work provided a valuable dataset and insights into different forms of online abuse, including hate speech. [9]

III. Methodology

The machine learning approaches can construct classifiers to complete sentiment classification by extracting feature vectors, which incorporates data collecting and preprocessing, features extraction, training the data with the classifier, and analyzing results.

It is required to separate the dataset into a training dataset and a test dataset. The test dataset assesses the classifier's performance while the training sets are intended to help the classifier learn the text features.

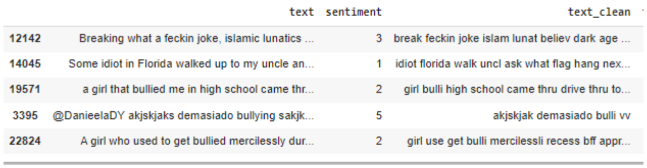
A. Dataset description and pre processing

The dataset utilized in this study includes tweets from users of the social media platform Twitter about various subjects directed at people in a variety of professions, including actors, models, athletes, musicians, and politicians. Using the Twitter API, web scraping was used to gather 31596 tweets for this dataset. Tweets and sentiment are the two variables (columns) in the dataset. Each Tweet fell into one of five categories: religion, ethnicity, age, gender, and not cyberbullying or neutral. There are 6342 tweets related to religious views, 6334 related to ethnicity, 6328 related to age, 6296 related to gender and finally 6296 tweets belong to not cyberbullying. Later, religion, age, gender, ethnicity and not cyberbullying categorical values in the sentiment column were replaced by numerical values namely 1, 2, 3, 4 and 5 respectively.

Text preprocessing is a fundamental step in sentiment analysis that involves transforming raw textual data into a format suitable for analysis. This section outlines the various text preprocessing techniques employed in our research to ensure the quality and consistency of our dataset.

The first step involves removing any non-essential elements from the text, such as HTML tags, special characters, and punctuation. This is done to ensure that the text is in a clean and uniform format. Then, to maintain consistency and reduce the dimensionality of the data, all text is converted to lowercase. Tokenization is the process of breaking down text into individual words or tokens. We employ word tokenization to split the text into its constituent words.

Later, in stop words removal section, such as "the," "and," "is," are common words that often carry little sentiment information.



	text	sentiment	text_clean
12142	Breaking what a feckin joke, islamic lunatics ...	3	break feckin joke islam lunat believ dark age ...
14045	Some idiot in Florida walked up to my uncle an...	1	idiot florida walk uncl ask what flag hang nex...
19571	a girl that bullied me in high school came thr...	2	girl bulli high school came thru drive thru to...
3395	@DanieelaDY akjskjaks demasiado bullying sakjk...	5	akjskjak demasiado bulli vv
22824	A girl who used to get bullied mercilessly dur...	2	girl use get bulli mercilessli recess bff appr...

Fig 2 Dataset sample after pre processing

Then tweets with less than 3 words and more than 100 words were eliminated as they can be outliers. Stemming and lemmatization are techniques used to reduce words to their base or root form. We experimented with stemming to evaluate their impact on sentiment analysis accuracy. Emoticons and emoji are often used to convey sentiments. We preprocess text to identify and standardize emoticons and emoji, converting them into textual representations for analysis. Dataset after filtering is shown in Fig 2.

B. Feature extraction

Feature extraction is a pivotal step in sentiment analysis, wherein the preprocessed textual data is converted into a numerical representation that machine learning models can analyze. This section outlines the feature extraction techniques employed in our research to capture relevant information from the text for sentiment classification.

Researchers employ a number of methods, including word2vec, TF-IDF, Count Vectorizer, Bag of Words (BOW), and One Hot Encoding. In our investigation, the text vectorizer was Count Vectorizer.

C. Classification

In this study, we created a model to recognize twitter hate speech using four machine learning classifiers, including Naïve Bayes (NB), k Nearest Neighbors (KNN), Random Forest (RF), and Support Vector Machine (SVM). The effectiveness of each classifier has been evaluated in relation to numerous performance parameters. In the subsection that follows, we will discuss various performance parameters used for the prediction model.

D. Performance metrics

Evaluating the performance of sentiment analysis models is paramount to gauge their accuracy, reliability, and suitability for practical applications. This section presents the performance metrics employed in our research to assess the effectiveness of sentiment classification models.

Accuracy:

Accuracy measures the proportion of correctly classified instances out of the total. It provides an overall assessment of model performance, shown in eq 1.

$$accuracy = \frac{TP+TN}{TP+FP+FN+TN} \qquad (1)$$

Precision and Recall:

Precision measures the proportion of true positive predictions among all positive predictions, emphasizing the correctness of positive predictions. Recall, on the other hand, measures the proportion of true positive predictions among all actual positive instances, emphasizing the ability to capture all positive instances as in eq 2 and eq 3.

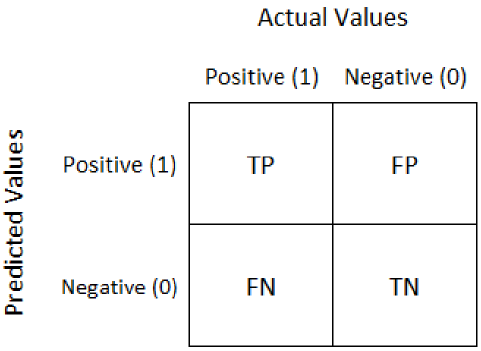
$$precision - \frac{TP}{TP+FP} \qquad (2)$$

$$recall = \frac{TP}{TP+FN} \qquad (3)$$

F1-Score:

The F1-score is the harmonic mean of precision and recall, providing a balanced assessment of model performance, depicted in eq 4.

$$F\ 1 = \frac{2}{\left(\frac{1}{precision}\right)+\left(\frac{1}{recall}\right)} \qquad (4)$$



	Positive (1)	Negative (0)
Positive (1)	TP	FP
Negative (0)	FN	TN

Fig 3 Confusion Matrix

Confusion Matrix:

The confusion matrix provides a detailed breakdown of model performance, showing the number of true positives, true negatives, false positives, and false negatives. Shown in fig 3.

IV. Result analysis

After applying selected classifiers following precision, recall and f1 score were achieved. Comparison of all accuracy is shown in Fig 4. From the figure it can be seen that, 84%, 81%, 94% and 93% accuracy was achieved by NB, KNN, RF and SVM algorithms respectively. Best accuracy was demonstrated by RF algorithm. SVM showed the second best performance in terms of accuracy. Precision, recall, f1 score for RF and SVM are shown in Table 1 and Table 2.

For RF algorithm best precision, recall and f1 score was demonstrated by ethnicity, and second best was by age. For SVM algorithm precision, recall and f1 score was achieved by ethnicity. The second best performance was precision, recall and f1 score was achieved by religion and gender, age and religion and age.

Confusion matrix of all the algorithms is shown in Fig 5.

Table: 1 Precision, Recall and F1 score RF algorithm

	precision	recall	f1-score
religion	0.97	0.96	0.96
age	0.98	0.98	0.98
gender	0.95	0.87	0.91
ethnicity	0.99	0.99	0.99
Not bullying	0.81	0.90	0.85

Table: 2 Precision, Recall and F1 score SVM algorithm

	precision	recall	f1-score
religion	0.97	0.95	0.96
age	0.95	0.97	0.96
gender	0.97	0.87	0.92
ethnicity	0.98	0.98	0.98
Not bullying	0.78	0.88	0.83

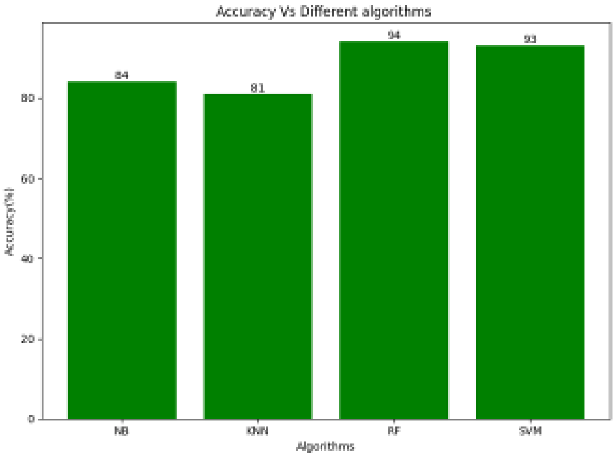
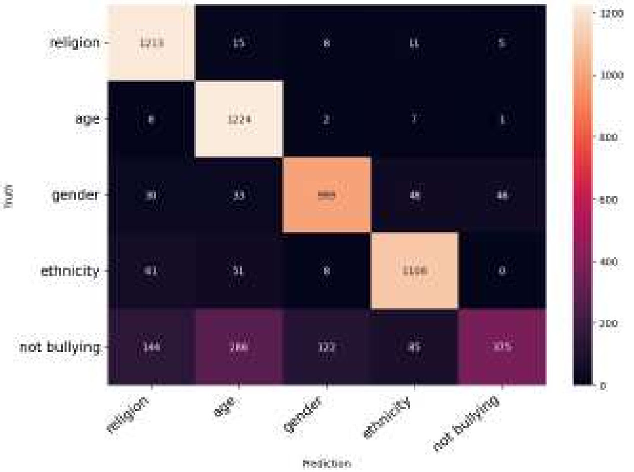
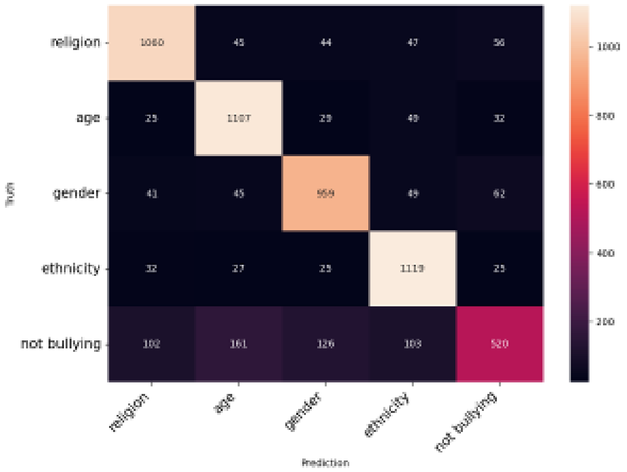


Fig 4 Comparison of Accuracies For All Algorithms

V. Conclusions

Through a rigorous analysis of diverse sentiment analysis techniques, we delved into the nuances of feature extraction, model selection, and performance evaluation. Our findings underscore the pivotal role of preprocessing steps in enhancing the quality of textual data and the significance of appropriate feature representations in sentiment analysis. We demonstrated the versatility of machine learning algorithms and the effectiveness of deep learning models, each offering a unique lens through which to view and interpret sentiment. Furthermore, we investigate the impact of different feature engineering techniques, model architectures, and hyperparameter tuning on the hate speech detection task. Our results demonstrate that our hybrid approach outperforms baseline models, achieving high accuracy, precision, recall, and F1-score in identifying hate speech tweets. Additionally, we explore the challenges and ethical considerations in hate speech detection, emphasizing the importance of addressing bias and fairness concerns in automated moderation systems. This research contributes to the ongoing efforts to mitigate the spread of hate speech on social media platforms and promotes safer and more inclusive online environments. The developed models and insights from this study offer valuable tools for social media companies and policymakers to combat online hate speech effectively.

Some of the future prospects of our study includes, our technique of hate speech detection can include not only text but also images and videos shared on Twitter. Multimodal analysis can provide a more comprehensive understanding of hate speech and its propagation. Also, develop models that consider the context in which tweets are posted.

<table>
<tr><th colspan="2">Confusion matrix for all algorithms</th></tr>
<tr><td>Confusion matrix for NB</td><td>Confusion matrix for KNN</td></tr>
<tr><td>
</td><td>
</td></tr>
</table>

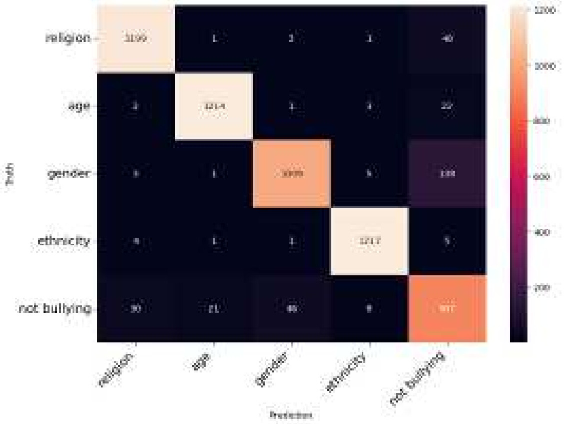
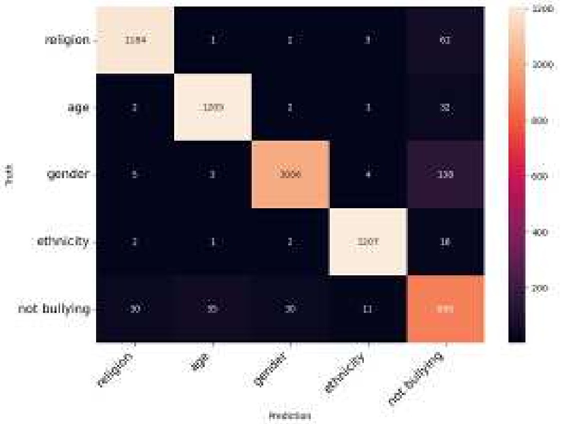
Confusion matrix for RF	Confusion matrix for SVM
	

Fig 5 Confusion Matrix For All Algorithms

References:

[1] K. Dashtipour *et al.*, “Multilingual sentiment analysis: state of the art and independent comparison of techniques,” *Cogn Comput*, vol. 8, no. 4, pp. 757–771, Aug. 2016, doi: 10.1007/s12559-016-9415-7.

[2] Y. Qi and Z. Shabrina, “Sentiment analysis using Twitter data: a comparative application of lexicon- and machine-learning-based approach,” *Soc. Netw. Anal. Min.*, vol. 13, no. 1, pp. 1–14, Dec. 2023, doi: 10.1007/S13278-023-01030-X/TABLES/3.

[3] K. Arun and A. Srinagesh, “Multi-lingual Twitter sentiment analysis using machine learning,” *Int J Electr Comput Eng*, vol. 10, no. 6, pp. 5992–6000, Dec. 2020, doi: 10.11591/ijece.v10i6.pp5992-6000.

[4] T. Davidson, D. Warmsley, M. Macy, and I. Weber, “Automated Hate Speech Detection and the Problem of Offensive Language,” *Proc. Int. AAAI Conf. Web Soc. Media*, vol. 11, no. 1, pp. 512–515, May 2017, doi: 10.1609/ICWSM.V11I1.14955.

[5] Z. Waseem and D. Hovy, “Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter,” *HLT-NAACL 2016 - 2016 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Proc. Student Res. Work.*, pp. 88–93, 2016, doi: 10.18653/V1/N16-2013.

[6] V. Basile *et al.*, “SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter,” *NAACL HLT 2019 - Int. Work. Semant. Eval. SemEval 2019, Proc. 13th Work.*, pp. 54–63, 2019, doi: 10.18653/V1/S19-2007.

[7] 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, doi: 10.1609/ICWSM.V11I1.14955.

[8] Z. Zhang, D. Robinson, and J. Tepper, “Detecting Hate Speech on Twitter Using a Convolution-GRU Based Deep Neural Network,” *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10843 LNCS, pp. 745–760, 2018, doi: 10.1007/978-3-319-93417-4_48/TABLES/4.

[9] A. M. Founta *et al.*, “Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior,” *12th Int. AAAI Conf. Web Soc. Media, ICWSM 2018*, pp. 491–500, Feb. 2018, doi: 10.1609/icwsm.v12i1.14991.