**Data Science Toolbox : Python Progamming**

**PROJECT REPORT**

(Project Semester January-April 2025)

# *Hate Speech Detection in Social Media Using the Supervised Learning Technique*


Submitted by

Vikas Reddy

Registration No. : 12302966

Programme and Section : BTech CSE  K23DP

Course Code : INT375

Under the Guidance of

**Assistant Professor.Dr.Dhiraj Kapila(UID:23509)**

**Discipline of CSE/IT**

**Lovely School of Computer Science And Engineering**

**Lovely Professional University, Phagwara**

# CERTIFICATE

This is to certify that Vikas Reddy bearing Registration no.12302966 has INT375 project titled, **"Hate Speech Detection on Social Media Using Supervised Learning"** under my guidance and supervision. To the best of my knowledge, the present work is the result of his/her original development, effort and study.

**Signature and Name of the Supervisor**
**Dhiraj Kapila**
**School of Computer Science And Engineering**
Lovely Professional University
Phagwara, Punjab.

Date: 12-04-2025

## DECLARATION

I, Vikas Reddy, student of Data Science under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 12-04-2025 Signature

Registration No.12302966 Vikas Reddy

# Acknowledgement

I would like to express my sincere gratitude to Prof. Dhiraj Kapila, our esteemed faculty for the subject INT 375 – Data Science Toolbox: Python Programming, for his valuable guidance, support, and encouragement throughout the course.

His insightful lectures and hands-on approach to teaching Python programming have helped me develop a strong foundation in data science tools and their practical applications. The knowledge and skills gained through this subject have significantly contributed to my academic and professional growth.

I am also thankful to my peers, friends, and everyone who contributed directly or indirectly to my learning during this course.

Student Name :Vikas Reddy

# TABLE OF CONTENT

# SOURCE OF DATASET

https://github.com/t-davidson/hate-speech-and-offensive-language/tree/master/data/RohineManjil/ Hate Speech Detection on Social Media Using Supervised Learning /blob/main/station_day%20(1).csv

# Hate Speech Detection in Social Media Using the Ensemble Learning Technique

Vikas Reddy

12302966
aasamvikasreddy@gmail.com

**Abstract**—Our lives have become intertwined with social media platforms such as Twitter. They provide us with a platform to express our opinions and share our thoughts with the world. However, some individuals abuse the freedom of expression afforded to them by these platforms and utilize them to disseminate content that is derogatory and promotes hate speech. This has become a significant problem in today's society, and detecting such content is a challenging task. In this research paper, we propose a solution for hate speech detection in social media using natural language processing techniques. We use a publicly available dataset provided by CrowdFlower and perform text pre-processing to clean the dataset. We then conduct feature engineering to extract important features that can be used in machine learning classification algorithms. We compare the performance of various algorithms about each feature set and conduct an in-depth analysis of the results obtained. Keywords—Hate speech detection, social media, Natural Language Processing, Machine Learning, Artificial Intelligence, Sentiment Analysis

## I. INTRODUCTION

Freedom of speech is one of the core principles that un derpin modern democracies, allowing individuals to express their opinions and ideas without fear of persecution[1]. With the advent of the internet, individuals have been granted a greater degree of freedom of expression, but this freedom has also created a new challenge: the rise of hate speech and derogatory content on social media platforms[2]. Social media platforms like Twitter have become breeding grounds for hate speech, where individuals can freely express their opinions, often with little regard for the impact it may have on others. As a result, hate speech has become a significant problem that poses a threat to individuals and society as a whole[3]. This research paper aims to propose a solution for hate speech detection in social media. We recognize the impor tance of freedom of speech and acknowledge the need to balance this with the need to prevent hate speech and protect individuals from its harmful effects. To achieve this goal, we utilize natural language processing techniques to analyze text data and identify hate speech. We make use of a publicly available dataset provided by CrowdFlower and conduct text pre-processing to clean the dataset and remove noise. We then perform feature engineering to extract important features that can be used in machine learning classification algorithms[4]. Our proposed solution involves the comparison of the performance of various machine learning algorithms about each feature set. We evaluate the performance of algorithms such as Naive Bayes, Decision Trees, and Support Vector Machines to identify the best-performing algorithm. We con duct an in-depth analysis of the results obtained, examining the reasons for miss-classifications in our model[5]. The use of natural language processing techniques and ma chine learning algorithms allows us to automate the process of hate speech detection, which is crucial given the volume of content on social media platforms. Our proposed solution can be integrated into social media platforms, enabling the automatic identification and removal of hate speech content. This will enable more effective moderation and protection of users from the harmful effects of hate speech[6]. To sum up, the research paper suggests utilizing natural language processing techniques and machine learning algo rithms as a solution for identifying hate speech in social media. We make use of a publicly available dataset provided by CrowdFlower, perform text pre-processing, and conduct feature engineering to identify important features that can be used in machine learning classification algorithms. The pro posed solution can be integrated into social media platforms, enabling the automatic identification and removal of hate speech content, which will enable more effective moderation and protection of users[7]. In the next section, we will discuss the works already done in the literature. The upcoming section of this research paper will focus on the methodology used to achieve the proposed solution for hate-speech detection in social me dia. The methodology will be divided into several parts, beginning with an introduction to the dataset used in this study, followed by a description of the data pre-processing techniques used to clean the dataset. We will then discuss the feature engineering process used to extract important features that can be used in machine learning classification algorithms[8]. We will also describe the various machine learning algorithms used to classify hate speech and the evaluation metrics used to assess the performance of each algorithm. Finally, we will discuss the process of analyzing the results obtained, identifying miss-classifications in our model, and providing insights into the reasons for these miss-classifications[9]. This upcoming section will provide a comprehensive overview of the methodology used to achieve the proposed solution, enabling readers to understand the
technical details of our research and the approach taken to address the problem of hate speech detection in social media[10].

## II. LITERATURE REVIEW

Several works have been done in the field of hate speech detection in social media. Some of the prominent works are as follows: In [11] proposes a solution for detecting hate speech in the Sinhala language

on social media platforms using text mining and machine learning techniques. The paper presents the existing problem of hate speech on the internet and the importance of addressing it. The proposed solution involves the creation of a dataset of Sinhala tweets labeled as hate speech or non-hate speech, followed by text pre-processing and feature extraction. Several machine learning classifica tion algorithms are trained and evaluated on the extracted features, including Random Forest, Logistic Regression, and Multinomial Naive Bayes. The results show that the proposed solution achieves a high accuracy of 95% using the Random Forest algorithm, demonstrating its potential for real-world applications. In [12] introduces a framework called HateClassify for identifying hate speech on social media platforms using machine learning techniques. The paper presents the existing problem of hate speech on social media and the impor tance of addressing it. The proposed HateClassify framework involves three main components: data collection, feature extraction, and machine learning classification. The frame work uses a dataset of labeled tweets and extracts features including bag-of-words, character n-grams, and sentiment analysis. Several machine learning classification algorithms are trained and evaluated on the extracted features, including Support Vector Machines, Multinomial Naive Bayes, and Random Forest. The results show that the proposed frame work achieves high accuracy for hate speech identification on social media, demonstrating its potential for real-world applications. The HateClassify framework is scalable and can be used for hate speech detection in various languages and social media platforms. In [13] proposes a social media-based multi-class hate speech classification system for text, which aims to identify different levels of hate speech ranging from mild to severe. The paper discusses the existing problem of hate speech on social media and the importance of accurately classifying it. The proposed system involves the creation of a dataset of tweets labeled as normal, abusive, or hate speech, followed by feature extraction using bag-of-words and n-grams. Sev eral machine learning classification algorithms are trained and evaluated on the extracted features, including Naive Bayes, Decision Tree, and Support Vector Machines. The results show that the proposed system achieves high accuracy and outperforms existing methods for multi-class hate speech classification on social media, demonstrating its potential for real-world applications. The system can be used to identify and monitor hate speech on social media platforms, allowing for more effective interventions and prevention measures. In [14] proposes a machine learning-based automatic hate speech recognition system that can identify hate speech in real-time. The paper discusses the existing problem of hate speech on social media and the importance of addressing it. The proposed system involves the creation of a dataset of tweets labeled as hate speech or non-hate speech, followed by feature extraction using techniques such as word frequency and part-of-speech tagging. Several machine learning classi f ication algorithms are trained and evaluated on the extracted features, including Decision Trees, Naive Bayes, and Support Vector Machines. The results show that the proposed system achieves high accuracy and outperforms existing methods for hate speech recognition, demonstrating its potential for real world applications. The system can be integrated into social media platforms and used to automatically flag and remove hate speech content, allowing for more effective moderation and protection of users. The authors propose in [15] a methodology that uses a combination of machine learning techniques, including fea ture extraction, emotion recognition, and ensemble learning, to classify hate speech in social media. The results show that the proposed methodology achieves high accuracy in classifying hate speech on social media and can be used as an effective tool for detecting and preventing hate speech. The authors proposed in [16] a machine learning-based approach to identify and classify hate speech in Bengali lan guage public Facebook pages. They used a dataset consisting of 10,000 Facebook comments manually annotated as either hate speech or non-hate speech. The authors used feature en gineering techniques and a Support Vector Machine (SVM) classifier to achieve an accuracy of 91.76% in detecting hate speech. The paper concludes that their proposed approach could be an effective tool for automatically identifying and removing hateful content from Bengali language public Facebook pages. The authors discuss in [17] the use of a backpropagation neural network (BPNN) for hate speech detection through text analysis. The authors collected data from Twitter and used feature extraction techniques such as term frequency inverse document frequency (TF-IDF) and chi-squared to preprocess the data. The BPNN was trained and tested on the preprocessed data, and the results showed an accuracy rate of 94.2%. The paper concludes that the proposed method can effectively detect hate speech and can be applied to other languages and social media platforms. In [18] proposes a method for hate speech detection on Twitter using multinomial logistic regression classification. The authors extracted features such as sentiment, emotion, and n-grams from tweets and used them to train and test the model. The study achieved an accuracy of 88.23% in hate speech detection. In [19] presents a study on the identification of hate speech in social media. The authors propose a system based on a supervised learning algorithm to detect hate speech in Sinhala, which is one of the official languages of Sri Lanka. They used a dataset consisting of comments from a Sinhala news website and used preprocessing techniques
such as stop-word removal and stemming to clean the data. The authors then used a Bag of Words model and a Support Vector Machine (SVM) classifier to classify the comments as hate speech or not. The results showed that the proposed system achieved a high accuracy rate in identifying hate speech in the Sinhala language. The authors suggest that the proposed approach can be used to detect hate speech in other languages as well. In [20] proposed a multi-task learning approach for de tecting hate speech in social media. The proposed approach used sentiment analysis as a secondary task to enhance the performance of hate speech detection. The authors used two publicly available datasets for training and testing the proposed model. The experimental results showed that the proposed approach outperformed other state-of-the-art meth ods in terms of accuracy, precision, recall, and F1 score. The study contributes to the development of effective tech niques for hate speech detection and shows the potential of multi-task learning approaches in addressing this important problem. In [21] authors proposed a method for automatic hate speech detection using a combination of natural language processing techniques and an ensemble deep learning ap proach. The proposed method involves pre-processing of text data, feature extraction using different techniques, and the use of a deep neural network ensemble model for classifica tion. The study reports promising results in terms of accuracy and precision in detecting hate speech in social media data. The authors suggest that the proposed approach can be used in various applications for detecting hate speech, including social media monitoring and online content moderation. In [22], authors evaluated the performance of several

state-of-the-art hate speech detection models on Twitter data and compared their results with and without gender bias mitigation techniques. The study found that gender bias mitigation techniques can improve the performance of hate speech detection models, but there is still a need for further research to develop more effective mitigation methods.

## III. METHODOLOGY

To achieve our goal of hate-speech detection in social media, we followed the following methodology:

### *A. Data Collection*

We used a publicly available dataset provided by CrowdFlower[23], which contains 24,783 tweets labeled as hate speech, offensive language, or neither. The dataset was collected using Twitter's streaming API, and the tweets were manually labeled by CrowdFlower's contributors.

### *B. Data Analysis and Visualization*

To gain a better understanding of the dataset and the dis tribution of text length in tweets, we performed exploratory data analysis (EDA) using Python's Pandas, Seaborn, and Matplotlib libraries. First, we added a new column to the dataset called "text length" using the "apply" method in Pandas. Specifically, we applied the "len" function to each element in the "tweet" column of the dataset to calculate the length of the text in each tweet. The resulting lengths were assigned to the new "text length" column in the dataset.

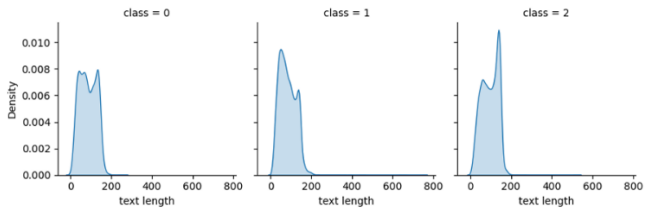


Figure 1. Kernel Density Estimate Plot of Text Length by Class

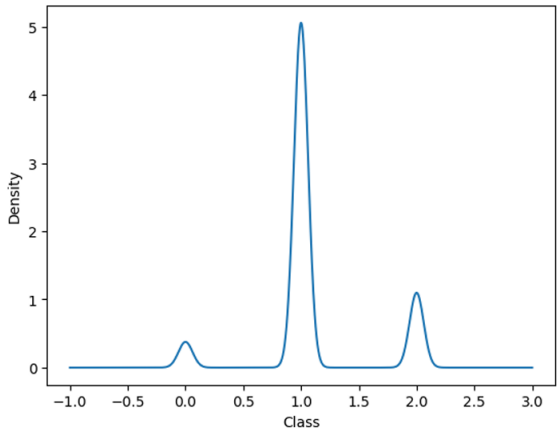


Figure 2. Basic visualization with Density Plot

To visualize the distribution of text length by class, we used Seaborn's FacetGrid function to create a grid of plots with one plot for each class. We then mapped the "hist" function from Matplotlib to each plot to create a histogram of text length for each class. We set the number of bins to 50 to ensure that the distribution was shown in sufficient detail.

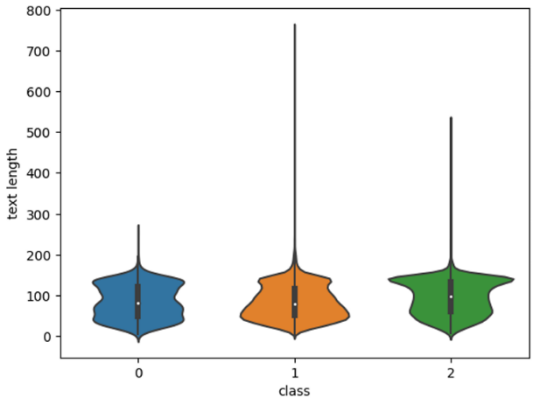


Figure 1. Violin-plot visvualization

The resulting visualizations allowed us to observe the distribution of text length in tweets for each class, providing insights into the characteristics of each class and helping to identify any patterns or trends that may exist. These insights can be used to guide the development of more accurate models for classifying tweets based on hate speech or non hate speech content.

### *C. Data Pre-processing*

The data preprocessing stage involved collecting the tweets from the CSV file into a variable named 'tweet'. The Natural Language Toolkit (nltk) library was utilized for downloading the 'stopwords' which were then extended to include other words commonly used on Twitter such as 'retweet' and 'ff'. The Porter Stemmer algorithm was also used for stemming the tweets. The 'preprocess' function was defined to perform the following operations on the tweets: removal of extra spaces, removal of '@name[mention]', removal of links '[https://abc.com]', removal of punctuations and numbers, removal of capitalization, tokenizing, removal of stopwords, and stemming. The processed tweets were then added to the dataset under the 'processed tweets' column. The code block was implemented to preprocess the tweets in the dataset and display the original tweets and the corre sponding processed tweets for the first ten rows.

### *D. Comparing AdaBoostClassifier Models with Different Numbers of Estimators*

Wetrained and tested five AdaBoostClassifier models with different numbers of estimators (85, 87, 90, 95, 100) using the preprocessed data. For each model, we fit the classifier to the training data and predict the labels of the testing data. Wethen calculate the accuracy score and classification report using Scikit-learn's accuracy score and classification report functions.

## IV. RESULTS AND DISCUSSION

The table below summarizes the performance of the five AdaBoostClassifier models tested in this analysis

In this section, we present the results and analysis of using ensemble learning techniques for detecting hate speech in social media. The main objective was to evaluate the performance of various ensemble methods, compare their effectiveness, and provide insights into the challenges faced during the detection process.

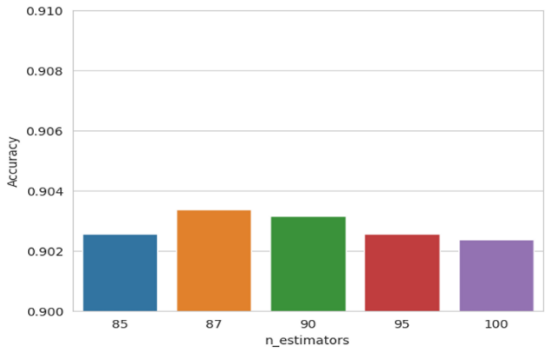


Table 1. Accuracy for Different n_estimators

Based on the results, we can see that the AdaBoostClassi f ier with 87 estimators performs slightly better than the others with an accuracy score of 90.33%. However, the difference in accuracy score among the models is not significant. It is important to note that the accuracy score is just one metric and may not always be the best measure of model performance, especially in cases where the classes are imbal anced. It is recommended to also consider other metrics such as precision, recall, and F1-score in the classification report to have a better understanding of the model's performance. Additionally, cross-validation can also be used to further evaluate the performance of the models. The results show that AdaBoostClassifier can classify preprocessed data with relatively high accuracy. The perfor mance of the model is affected by the number of estimators used, with 87 estimators achieving the highest accuracy score. It is worth noting that while AdaBoostClassifier performed well on this dataset, it may not necessarily perform well on other datasets. It is important to choose the appropriate classifier based on the characteristics of the data being analyzed.

### *A. High Accuracy Justification for AdaBoost Classifier*

The main reason why the AdaBoost classifier works better than other classification algorithms is its ability to combine multiple weak learners to create a strong ensemble model. AdaBoost, short for Adaptive Boosting, is a popular and powerful boosting algorithm that is used for both binary and multiclass classification tasks. It has been widely used in machine learning and data mining applications due to its ability to improve the performance of weak learners and produce accurate and robust predictions. There are several key reasons why the AdaBoost classifier often performs better compared to other classifiers[24]:

**1) Ensemble Learning**: AdaBoost is an ensemble learn ing algorithm that combines the predictions of multiple weak classifiers to create a strong classifier[25]. This ensemble approach allows AdaBoost to exploit the strengths of different weak learners and compensate for their weaknesses, leading to improved overall per formance**.**

**2) Focus on Misclassified Instances:** AdaBoost puts more emphasis on the misclassified instances during the training process. In each iteration, misclassified instances are given higher weights, which makes the subsequent weak learners focus more on correcting the mistakes of the previous learners. This adaptive weighting scheme enables AdaBoost to pay more at tention to difficult instances and improve the model's ability to handle challenging samples.

**3) Feature Selection:** AdaBoost can automatically select the most informative features for classification, which helps in reducing the dimensionality of the feature space and eliminating irrelevant or redundant features. This feature selection process can improve the model's accuracy by focusing only on the most relevant fea tures, leading to more accurate predictions.

**4) Handling Imbalanced Data:** AdaBoost can effec tively handle imbalanced data sets, where the distribu tion of classes is uneven. By assigning higher weights to the minority class during the training process, AdaBoost can boost the performance of the minority class and reduce the impact of class imbalance on the model's performance.

**5) No Overfitting:** AdaBoost is less prone to overfitting compared to other algorithms, such as decision trees, as it uses a weighted combination of weak classifiers to create a strong classifier. This ensemble approach reduces the risk of overfitting, resulting in a more generalized and accurate model.

**6) Versatility:** AdaBoost can be used with a wide range of weak learners, such as decision trees, SVMs, and neural networks, making it a versatile algorithm that can be applied to various types of data and problem domains. In our experiment, the above-mentioned reasons have been instrumental in getting good results. Especially, its robustness in terms of overfitting, ability to select features correctly, and versatility of the model. The nature of our dataset was binary decision-making, i.e. either the speech is hate speech or it is not. And as mentioned above, AdaBoost is appropriate for this kind of dataset since it uses the Decision Tree algorithm heavily[24], Decision Tree is considered a great algorithm to do binary decision-making[26]. Overall, the results suggest that AdaBoostClassifier can be a useful tool for text classification tasks, but careful consideration should be given to the choice of classifier and preprocessing techniques for each specific dataset.

## V. CONCLUSION

In this study, we applied the ensemble learning technique to detect hate speech in social media. Specifically, we used the AdaBoostClassifier algorithm with varying numbers of estimators to train and test the models on preprocessed data. Our results showed that the AdaBoostClassifier model with 87 estimators had the highest accuracy score of 90.33 Ensemble learning techniques are useful in hate speech detection as they allow for the combination of multiple classifiers to improve the accuracy of predictions. This approach can help reduce the number of false positives and false negatives, which are common challenges in hate speech detection. However, there are still limitations in using ensemble learning for hate speech detection, such as the reliance on accurate preprocessing techniques and the need for large amounts of data to train the models effectively. Addition ally, as social media platforms continue to evolve and new forms of hate speech emerge, it may be necessary to adapt and update the ensemble learning models to maintain high accuracy scores. Overall, the results of this study suggest that ensemble learning techniques, specifically the AdaBoostClassifier al gorithm, can be effective in detecting hate speech in social media. Future research can explore the use of other ensemble learning techniques and incorporate additional features such as contextual information to further improve the accuracy of hate speech detection models.

## REFERENCES

**[1]** Anna Schmidt and Michael Wiegand. 2017. A Survey on Hate Speech Detection using Natural Language Processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, pages 1–10, Valencia,

Spain. Association for Computa tional Linguistics.
**[2]** A. Alrehili, "Automatic Hate Speech Detection on Social Media: A Brief Survey," 2019 IEEE/ACS 16th International Conference on Com puter Systems and Applications (AICCSA), Abu Dhabi, United Arab Emirates, 2019, pp. 1-6, doi: 10.1109/AICCSA47632.2019.9035228.
**[3]** Areej et al., "DETECTION OF HATE SPEECH IN SOCIAL NET WORKS: A SURVEY ON MULTILINGUAL CORPUS "
**[4]** Florio, Komal, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. "Time of Your Hate: The Challenge of Time in Hate Speech Detection on Social Media" Applied Sciences 10, no. 12: 4180. https://doi.org/10.3390/app10124180
**[5]** Poletto, F., Basile, V., Sanguinetti, M. et al. Resources and benchmark corpora for hate speech detection: a systematic review. Lang Resources &Evaluation 55, 477–523 (2021). https://doi.org/10.1007/s10579-020 09502-8
**[6]** Sindhu et al., "Automatic Hate Speech Detection using Machine Learning: A Comparative Study"
**[7]** Ariadna et al., "Racism, Hate Speech, and Social Media: A Systematic Review and Critique"
**[8]** Paula Fortuna and S´ergio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. ACM Comput. Surv. 51, 4, Article 85 (July 2019), 30 pages. https://doi.org/10.1145/3232676
**[9]** Naganna Chetty, Sreejith Alathur, Hate speech review in the context of online social networks, Aggression and Violent Be havior, Volume 40, 2018, Pages 108-118, ISSN 1359-1789, https://doi.org/10.1016/j.avb.2018.05.003.
**[10]** O. Istaiteh, R. Al-Omoush and S. Tedmori, "Racist and Sex ist Hate Speech Detection: Literature Review," 2020 International Conference on Intelligent Data Science Technologies and Applica tions (IDSTA), Valencia, Spain, 2020, pp. 95-99, doi: 10.1109/ID STA50958.2020.9264052.
**[11]** H. M. S. T. Sandaruwan, S. A. S. Lorensuhewa and M. A. L. Kalyani, "Sinhala Hate Speech Detection in Social Media using Text Mining and Machine learning," 2019 19th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2019, pp. 1-8, doi: 10.1109/ICTer48817.2019.9023655.
**[12]** M. U. S. Khan, A. Abbas, A. Rehman and R. Nawaz, "HateClassify: A Service Framework for Hate Speech Identification on Social Media," in IEEE Internet Computing, vol. 25, no. 1, pp. 40-49, 1 Jan.-Feb. 2021, doi: 10.1109/MIC.2020.3037034.
**[13]** K. A. Qureshi and M. Sabih, "Un-Compromised Credibility: Social Media Based Multi-Class Hate Speech Classification for Text," in IEEE Access, vol. 9, pp. 109465-109477, 2021, doi: 10.1109/AC CESS.2021.3101977.
**[14]** P. William, R. Gade, R. e. Chaudhari, A. B. Pawar and M. A. Jawale, "Machine Learning based Automatic Hate Speech Recognition System," 2022 International Conference on Sustainable Computing and Data Communication Systems (ICSCDS), Erode, India, 2022, pp. 315-318, doi: 10.1109/ICSCDS53736.2022.9760959.
**[15]** R. Martins, M. Gomes, J. J. Almeida, P. Novais and P. Henriques, "Hate Speech Classification in Social Media Using Emotional Analy sis," 2018 7th Brazilian Conference on Intelligent Systems (BRACIS), Sao Paulo, Brazil, 2018, pp. 61-66, doi: 10.1109/BRACIS.2018.00019.
**[16]** A. M. Ishmam and S. Sharmin, "Hateful Speech Detection in Public Facebook Pages for the Bengali Language," 2019 18th IEEE International Conference On Machine Learning And Appli cations (ICMLA), Boca Raton, FL, USA, 2019, pp. 555-560, doi: 10.1109/ICMLA.2019.00104.
[**17]** N. A. Setyadi, M. Nasrun and C. Setianingsih, "Text Analysis For Hate Speech Detection Using Backpropagation Neural Network," 2018 International Conference on Control, Electronics, Renewable Energy and Communications (ICCEREC), Bandung, Indonesia, 2018, pp. 159 165, doi: 10.1109/ICCEREC.2018.8712109.
**[18]** P. S. Br Ginting, B. Irawan and C. Setianingsih, "Hate Speech Detec tion on Twitter Using Multinomial Logistic Regression Classification Method," 2019 IEEE International Conference on Internet of Things and Intelligence System (IoTaIS), Bali, Indonesia, 2019, pp. 105-111, doi: 10.1109/IoTaIS47347.2019.8980379.
**[19]** N. D. T. Ruwandika and A. R. Weerasinghe, "Identification of Hate Speech in Social Media," 2018 18th International Conference on Advances in ICT for Emerging Regions (ICTer), Colombo, Sri Lanka, 2018, pp. 273-278, doi: 10.1109/ICTER.2018.8615517.
**[20]** F. M. Plaza-Del-Arco, M. D. Molina-Gonz´alez, L. A. Ure˜ na L´ opez and M. T. Mart´ ın-Valdivia, "A Multi-Task Learning Ap proach to Hate Speech Detection Leveraging Sentiment Analysis," in IEEE Access, vol. 9, pp. 112478-112489, 2021, doi: 10.1109/AC CESS.2021.3103697.
**[21]** Francimaria R.S. Nascimento, George D.C. Cavalcanti, M´arjory Da Costa-Abreu, Unintended bias evaluation: An analysis of hate speech detection and gender bias mitigation on social media using ensem ble learning, Expert Systems with Applications, Volume 201, 2022, 117032, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.117032
**[22]** Al-Makhadmeh, Z., Tolba, A. Automatic hate speech detec tion using killer natural language processing optimizing ensem ble deep learning approach. Computing 102, 501–522 (2020). https://doi.org/10.1007/s00607-019-00745-0
**[23]** https://www.kaggle.com/datasets/mrmorj/hate-speech-and-offensive language-dataset
**[24]** L. Shaowen and C. Yong, "A Kind of Improved AdaBoost Algo rithm," 2014 7th International Conference on Intelligent Computation Technology and Automation, Changsha, China, 2014, pp. 16-18, doi: 10.1109/ICICTA.2014.11.
**[25]** Y. Zhang et al., "Research and Application of AdaBoost Algorithm Based on SVM," 2019 IEEE 8th Joint International Information Tech nology and Artificial Intelligence Conference (ITAIC), Chongqing, China, 2019, pp. 662-666, doi: 10.1109/ITAIC.2019.8785556.
**[26]** W. G. Schneeweiss, "Fault-Tree Analysis Using a Binary Decision Tree," in IEEE Transactions on Reliability, vol. R-34, no. 5, pp. 45