KEC AI DSNLP@LT-EDI-2024:Caste and Migration Hate Speech Detection using Machine Learning Techniques

Kogilavani Shanmugavadivel[1], Malliga Subramanian[1], Aiswarya M[1], Aruna T[1], Jeevaananth S[1]

[1]Department of AI, Kongu Engineering College, Perundurai, Erode.

{kogilavani.sv, mallinishanth72}@gmail.com
{aiswaryam.22aid, arunat.22aid}@kongu.edu
{jeevaananths.22aid}@kongu.edu

Abstract

Commonly used language defines "hate speech" as objectionable statements that may jeopardize societal harmony by singling out a group or a person based on fundamental traits (including gender, caste, or religion). Using machine learning techniques, our research focuses on identifying hate speech in social media comments. Using a variety of machine learning methods, we created machine learning models to detect hate speech. An approximate Macro F1 of 0.60 was attained by the created models.

1 Introduction

Caste and Migration hate speech refer to the use of discriminatory language or expressions that target individuals or groups based on their caste or migration status (Chakravarthi, 2022). These forms of hate speech can manifest in various ways, including verbal abuse, written messages, online content, or even physical actions (Mehta and Passi, 2022). The effects of caste and migration hate speech can be profound and detrimental on both individual and societal levels. Hate speech can cause significant psychological distress and emotional harm to individuals who are targeted (Gaydhani et al., 2018). It can lead to feelings of fear, anxiety, depression, and a sense of isolation. Hate speech contributes to the division and polarization of communities. It can create or exacerbate existing tensions between different caste or migrant groups, leading to social fragmentation (Al-Hassan and Al-Dossari, 2019). Caste and migration hate speech can have a significant impact on social media due to the widespread reach and instantaneous nature of online platforms (Alkomah and Ma, 2022). Detecting and addressing hate speech in social media comments is important for maintaining a safe and inclusive online environment (Razdan et al., 2021). For example, Pakistan is a country where caste still persists. Caste-based hate speech is severe and pervasive in Pakistan's countryside (Fortuna and Nunes, 2018). Social media users accused Dalits of spreading the COVID-19 pandemic because they are dirty and consume dead animals during the lockdown days. Activities related to caste discrimination also occur in India.

Detecting hate speech in online comments can be challenging due to several factors, and these difficulties pose significant obstacles for both automated systems and human moderators (Asogwa et al., 2022). Hate speech often relies on context, cultural nuances, and sarcasm. Automated systems may struggle to understand these subtleties, leading to false positives or negatives (Kumar and Kumar, 2023). The same words or phrases may have different meanings in different contexts. Hate speech detection becomes more complex in multilingual and multicultural environments (Karim et al., 2022). Different languages, dialects, and cultural norms contribute to variations in expression, making it challenging for automated systems to cover a diverse range of content accurately (Ayo et al., 2020). In this paper we mainly work on Tamil-English social media comments.

The shared task on "Caste and Migration Hate Speech Detection"[1] focuses on identifying character offsets of hate speech while handling code-mixed Tamil-English comments (Rajiakodi et al., 2024). Hate speech can be extracted using a variety of methods. In this work, we approach token labeling from the perspective of hate speech detection. To detect hate speech, we assessed KNN, the Decision Tree algorithm, and the Naïve Bayes-based token labeling system.

This is how the remainder of the paper is structured. First, the literature on studies relevant to caste hate speech identification is briefly discussed

[1] https://codalab.lisn.upsaclay.fr/competitions/16089

Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion, pages 206–210
March 21, 2024 ©2024 Association for Computational Linguistics

in section 2. After a thorough description of our system in Section 3, the tests and findings are reported in Section 4. We wrap off by discussing potential implications for further research.

2 Literature Review

Hate speech has grown to be a serious issue in today's world, with the ability to hurt both individuals and communities (William et al., 2022). Using machine learning techniques to automatically identify and flag hate speech in text-based data is one possible answer to this issue (Anjum and Katarya, 2023). In order to prevent hate speech and objectionable content from proliferating on social media platforms, it is important to monitor the speech and information that users are disseminating. To remove these harmful elements, a strong automated filter system is needed (Pereira-Kohatsu et al., 2019). A wide range of topics, including politics, religion, gender, caste, race, and color, are covered by hate speech and offensive content, which has the ability to polarize society (Biere et al., 2018).

"Hate Speech Detection Using ML" - In this paper, a decision tree algorithm-based hate speech detection system is proposed. Large datasets can be handled via the straightforward and efficient machine learning algorithm known as decision trees, which has been used to a variety of classification tasks with success. The decision tree model is trained using a dataset of labeled hate speech and non-hate speech material (El-Sayed et al., 2023). The decision tree algorithm is then used to classify the input text as hate speech or non-hate speech after they preprocess the text by eliminating stop words and stemming the words, extracting pertinent characteristics using the TF-IDF approach, and so on.

"Social Shout – Hate Speech Detection Using Machine Learning Algorithm" by Ohol et al.. They investigated a number of methodologies and strategies for machine learning-based hate speech identification in this research, including feature engineering, deep learning, supervised and unsupervised learning approaches, and natural language processing. They also talked on the difficulties and constraints associated with using machine learning to detect hate speech, including the scarcity of annotated datasets, the complexity of defining and classifying hate speech, and the possibility of bias in machine learning algorithms. This paper's overall goal is to give a summary of the state of machine learning-based hate speech identification today and to draw attention to the opportunities and difficulties that await further study in this significant and quickly developing area.

"SVM for Hate Speech and Offensive Content Detection" by Ratan et al. (2021). Support Vector Machine (SVM) was the traditional machine learning method used in this paper's experiments on Hindi and English datasets. In this article, the system and its outcomes were examined, with a focus on identifying hate speech and objectionable content. In Hindi, the model performs worse (0.7195 Macro F1), but it even managed an Macro F1 Score of 0.7563 in English.

3 Problem and System Description

Figure 1 illustrates a hate speech identification example. The aim is to determine the hate speech content given the input sentence. Thuluvavellalar and Agamudayar are the names of two different castes in the example above. Thuluvavellalar caste members propagate hate speech directed at the Agamudayar caste. Once the hate speech in that specific comment has been identified, it is labeled as 1 (contains hate speech) or 0 otherwise. This dataset's description is provided in Section 3.1.

3.1 Dataset Description

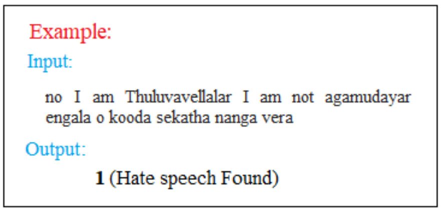
There are three columns in the publicly available shared task dataset, which is written in Tamil. These columns are labeled "text", which contains comments from social media platforms that contain hate speech as well as comments that do not, "id", which is the ID of those comments, and "label", which is set to 1 for hate speech and 0 otherwise. 5,355 samples make up the training set based on classes, and 1,575 samples without labels make up the testing set based on classes.

Dataset	No. of Comments
Train	5,355
Test	1,575

Table 1: Dataset Description

3.2 Development Pipeline

Figure 2 shows the overall development process that was employed for this project. Two modules might be separated out of our pipeline: (a) Feature Extraction and (b) Machine Learning Model. which are all exactly as said.



Example:

Input:

no I am Thuluvavellalar I am not agamudayar engala o kooda sekatha nanga vera

Output:

1 (Hate speech Found)

Figure 1: Example of Caste Hate Speech

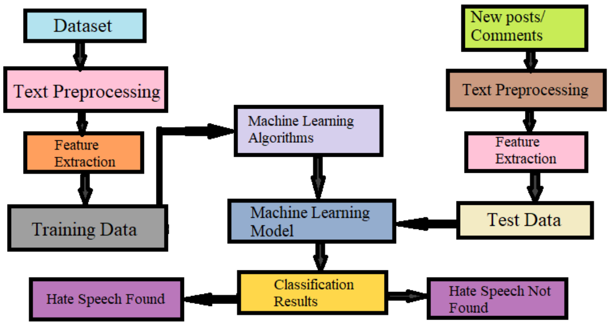


Figure 2: Proposed System Workflow

3.2.1 Feature Extraction

Since the dataset was in text format, we used a Python program to extract features so that machine learning techniques could support it. The process of converting unstructured text input into a format that machine learning algorithms may use for additional processing is known as feature extraction. To extract features for our model, we utilized TFIDFVectorizer, a Python module. Representing the significance of a word or phrase to a document is one of the most crucial information retrieval approaches. Consider the following scenario: we need to extract information from a string, or Bag of Words, and we may utilize this method to do it. TFIDF does not immediately transform unusable data into features. First, it creates vectors from raw strings or datasets, with a vector for every word. The characteristic will then be retrieved using a specific method, such as Cosine Similarity, which is applicable to vectors, etc. We are aware that the string cannot be passed straight to our model. Thus, TFIDF gives us the numerical values for each and every instances of the dataset.

3.2.2 Machine Learning Models

In recent years, machine learning has advanced significantly, altering people's perceptions of crucial applications like data mining, image recognition, and natural language processing (NLP). The machine learning models used in the current study for text classification are described in this section. We employed a variety of machine learning methods, including KNN, GaussianNB, and decision trees.wherein we chose a model based on its performance in comparison to the other two models.

KNN algorithm-The k-nearest neighbors algorithm, sometimes referred to as KNN or k-NN, is a non-parametric supervised learning classifier that groups individual data points based on closeness in order to classify or predict data. KNN can use the output of TFIDF as the input matrix and then predicts class label. Figure 3 illustrates the performance of our KNN model.

Decision Tree algorithm-A decision tree is a type of supervised learning algorithm that is commonly used in machine learning to model and predict outcomes based on input data. The decision trees implemented uses only numerical features and these features extracted by TFIDF are always as numeric variables. Figure 4 illustrates the performance of our Decision tree model.

GaussianNB-Machine learning techniques for classification based on a probabilistic method and Gaussian distribution are known as Gaussian Naive Bayes (GNB) techniques. Based on the premise that every parameter, also known as features or predictors, has the ability to independently predict the output variable, Gaussian Naive Bayes makes this

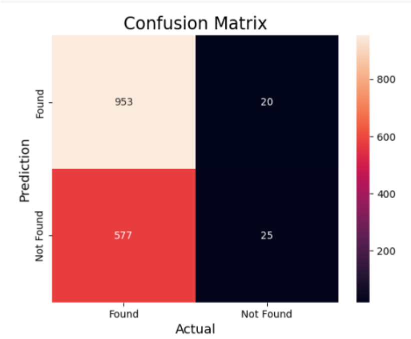


Figure 3: Performance of KNN model

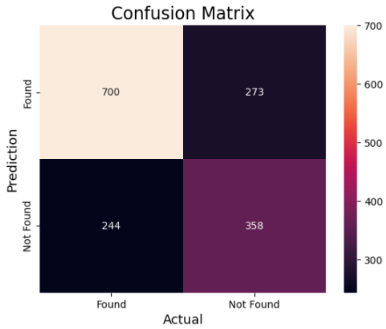


Figure 4: Performance of Decision tree model

assumption. Figure 5 illustrates the performance of our Naive Bayes model.

4 Experiments and Results

We have presented the most effective iteration of our model after doing a number of experiments to examine the model's effectiveness. The outcomes are displayed in Table 2. Although the KNN model did not yield improved accuracy, the other two models did yield better accuracy than the KNN model. Our model will get textual comments as input and then it classifies the texts whether it contains hate speech or not. As such, we intend to return to this challenge using more advanced architectures and language models.

5 Conclusion

Due to social media's accessibility and anonymity, as well as the shifting political situation in many regions of the world, hate speech has become more prevalent in recent years. The comment is identified as hate speech if the detection reveals that two or more of the individual outputs are positive for hatred; otherwise, it is identified as non-hate speech. Through text analysis, hate and offensive content were found in this study. The technique we employed in this work was designed to predict hate speech from code-mixed Tamil comments.

References

Areej Al-Hassan and Hmood Al-Dossari. 2019. Detection of hate speech in social networks: a survey on multilingual corpus. In *6th international conference on computer science and information technology*, volume 10, pages 10–5121.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Information*, 13(6):273.

Anjum and Rahul Katarya. 2023. Hate speech, toxicity detection in online social media: a recent survey of state of the art and opportunities. *International Journal of Information Security*, pages 1–32.

Doris Chinedu Asogwa, Chiamaka Ijeoma Chukwuneke, CC Ngene, and GN Anigbogu. 2022. Hate speech classification using SVM and NAIVE BAYES. *arXiv preprint arXiv:2204.07057*.

Femi Emmanuel Ayo, Olusegun Folorunso, Friday Thomas Ibharalu, and Idowu Ademola Osinuga. 2020. Hate speech detection in Twitter using hybrid embeddings and improved cuckoo search-based neural networks. *International Journal of Intelligent Computing and Cybernetics*, 13(4):485–525.

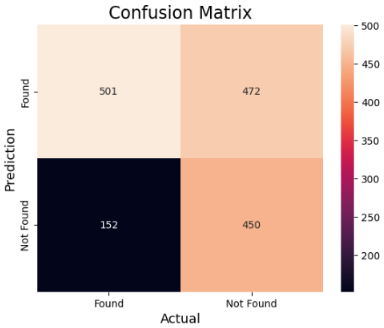


Figure 5: Performance of Naive Bayes model

Model	Macro F1-score
K-Nearest Neighbour	0.0772
Decision Tree	0.5862
Naive Bayes	0.5905

Table 2: Evaluation Metrics

Shanita Biere, Sandjai Bhulai, and Master Business Analytics. 2018. Hate speech detection using natural language processing techniques. *Master Business AnalyticsDepartment of Mathematics Faculty of Science*.

Bharathi Raja Chakravarthi. 2022. Hope speech detection in YouTube comments. *Social Network Analysis and Mining*, 12(1):75.

Tharwat El-Sayed, Abdallah Mustafa, Ayman El-Sayed, and Mohamed Elrashidy. 2023. Hate speech detection by classic machine learning. In *2023 3rd International Conference on Electronic Engineering (ICEEM)*, pages 1–4. IEEE.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Aditya Gaydhani, Vikrant Doma, Shrikant Kendre, and Laxmi Bhagwat. 2018. Detecting hate speech and offensive language on Twitter using machine learning: An n-gram and TFIDF based approach. *arXiv preprint arXiv:1809.08651*.

Md Rezaul Karim, Sumon Kanti Dey, Tanhim Islam, Md Shajalal, and Bharathi Raja Chakravarthi. 2022. Multimodal hate speech detection from Bengali memes and texts. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 293–308. Springer.

Ashwini Kumar and Santosh Kumar. 2023. Hate speech detection in multi-social media using deep learning. In *International Conference on Advanced Communication and Intelligent Systems*, pages 59–70. Springer.

Harshkumar Mehta and Kalpdrum Passi. 2022. Social media hate speech detection using explainable artificial intelligence (xai). *Algorithms*, 15(8):291.

VB Ohol, Siddhi Patil, Ishwari Gamne, Sayali Patil, and Shweta Bandawane. Social shout–hate speech detection using machine learning algorithm.

Juan Carlos Pereira-Kohatsu, Lara Quijano-Sánchez, Federico Liberatore, and Miguel Camacho-Collados. 2019. Detecting and monitoring hate speech in Twitter. *Sensors*, 19(21):4654.

Saranya Rajiakodi, Bharathi Raja Chakravarthi, Rahul Ponnusamy, Prasanna Kumar Kumaresan, Sathiyaraj Thangasamy, Bhuvaneswari Sivagnanam, and Charmathi Rajkumar. 2024. Overview of Shared Task on Caste and Migration Hate Speech Detection. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity and Inclusion*, Malta. European Chapter of the Association for Computational Linguistics.

Shyam Ratan, Sonal Sinha, and Siddharth Singh. 2021. SVM for Hate Speech and Offensive Content Detection.

Aditya Razdan et al. 2021. Hate speech detection using ml algorithms. In *2021 International Conference on Artificial Intelligence and Machine Vision (AIMV)*, pages 1–6. IEEE.

P William, Ritik Gade, Rup esh Chaudhari, AB Pawar, and MA Jawale. 2022. Machine learning based automatic hate speech recognition system. In *2022 International conference on sustainable computing and data communication systems (ICSCDS)*, pages 315–318. IEEE.