Selam@DravidianLangTech 2024:Identifying Hate Speech and Offensive Language

Selam Abitte Kanta, Grigori Sidorov and **Alexander Gelbukh**
Instituto Politécnico Nacional (IPN),
Centro de Investigación en Computación (CIC), Mexico City, Mexico
Corresponding: selaminadady300@gmail.com

Abstract

Social media has transformed into a powerful tool for sharing information while upholding the principle of free expression. However, this open platform has given rise to significant issues like hate speech, cyberbullying, aggression, and offensive language, negatively impacting societal well-being. These problems can even lead to severe consequences such as suicidal thoughts, affecting the mental health of the victims. Our primary goal is to develop an automated system for the rapid detection of offensive content on social media, facilitating timely interventions and moderation. This research employs various machine learning classifiers, utilizing character N-gram TF-IDF features. Additionally, we introduce SVM, RL, and Convolutional Neural Network (CNN) models specifically designed for hate speech detection. SVM utilizes character N-gram TF-IDF features, while CNN employs word embedding features. Through extensive experiments, we achieved optimal results, with a weighted F1-score of 0.77 in identifying hate speech and offensive language.

1 Introduction

In the digital age, social media plays an important role in online communication, allowing users to create and share material while also giving accessible means to express their views and thoughts on anything at any time (Edosomwan et al., 2011). However, with the advent of social media, platforms such as YouTube, Facebook, and Twitter not only aided in information sharing and networking, but they also became a place where people were targeted, defamed, and marginalized based solely on their physical appearance, religion, or sexual orientation(Keipi et al., 2016) Social media platforms have become more integrated with this digital era and have impacted various people's perceptions of networking and socializing(Tonja et al., 2022b).

Not only human beings, the hate content can corrupt the chatbots as well. Microscoft's chatbot 'Tay' which was developed to engage people through casual and playful conversation started using filthy language, which it learned from the conversation with people. The chatbot was unable to understand and avoid the hate content. So the detection of hate speech in tweets and social media sites has important applications in Chatbot building, content recommendation, sentiment analysis, etc.

India being a diverse country in terms of its culture and language has a huge population using code-mixed language in social media. Around 44% of the Indian population speak Hindi. So the usage of Hindi-English code-mixed language is very high on Twitter and Facebook. It is mainly seen among bilingual and multilingual communities. Code-mixing is the usage of certain words, phrases, or morphemes of one language in other languages.

This influence allowed different users to communicate via various social media platforms using a mix of texts. NLP technology has advanced rapidly in many applications, including machine translation(Tonja et al., 2022a),(Tash et al., 2022) Although considerable progress has been achieved in identifying offensive English language and hate speech, most research has mostly concentrated on identifying the abusive and offensive language in monolingual settings. This subject still appears to be at a very early stage of research for under resourced languages such as Tamil, Malayalam, and Kannada, which lack tools and datasets (Chakravarthi et al., 2020), (Yigezu et al., 2023d).

The task at hand involves identifying hate or offensive content in Telugu code-mixed text, with annotations made at the comment or post level. Given the complexity of language mixing in the Telugu context, where expressions may span multiple sentences, a tailored approach is crucial. In response, we deploy a Convolutional Neural Network (CNN)

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 91–95
March 22, 2024 ©2024 Association for Computational Linguistics

(Yigezu et al., 2023c,a)to enhance the effectiveness of detection methods. This research not only contributes to hate speech detection but also addresses the specific challenges posed by code-mixed expressions in Telugu. The subsequent sections delve into the deployment of CNN in our methodology, present experimental results, and discuss the implications of our approach.

2 Related Work

Currently, solving NLP problems in code-mixed data is getting attention from many researchers.

Social platforms are inundated with hate speech and offensive content, necessitating swift filtration (Yigezu et al., 2023e,b). However, manual filtering is nearly impossible due to the immense volume of incoming posts. This issue has garnered significant attention from the research community. Numerous studies have focused on predicting offensive and hate speech posts on social platforms, encompassing various languages. However, a majority of these studies have predominantly utilized English languages. languages

Detecting offensive content in social media comments is not a novel concept for the English language(Mandl et al., 2021). Several systems have also been developed for languages other than English, such as Hindi, Germany (Rajalakshmi et al., 2022),(Rajalakshmi and Reddy, 2019). However, there is limited research focused on identifying offensive content in low-resource Dravidian languages such as Tamil, Malayalam, and Kannada (Garain et al., 2021). The study proposes a method for identifying offensive language in code-mixed Kannada-English, Malayalam-English, and Tamil-English language pairs sourced from social media.(Ojo et al., 2023) The authors advocate for a multi-label classification approach, leveraging an ensemble of IndicBERT and generic BERT models, to recognize and mitigate offensive content on social media platforms.(Yigezu et al., 2022),(Yigezu et al., 2021) Addressing a multi-label classification challenge with various sub-categories, the system employs tokenization of the data before model training.

The calculation of class-wise confidence scores, subsequently combined into an output vector, facilitates effective classification. Across the Malayalam–English, Tamil-English, and Kannada-English datasets, the achieved F1 scores are 0.54, 0.72, and 0.66, respectively. This research task is outlined in the study conducted by(Kedia and Nandy, 2021). Introduced is a potent multiclass abusive detection model, showcasing an impressive accuracy and F1-score of 0.99 on a well-balanced dataset. The model detected abusive comments within Tamil and Telugu English code-mixed text, and employed the TF-IDF technique in conjunction with SVM, as demonstrated by (Balakrishnan et al., 2023).

3 Datasets

The dataset employed in this research is sourced from the Hate and Offensive Language Detection in Telugu Code-mixed Text (HOLD-Telugu) dataset provided by DravidianLangTech@EACL(Balakrishnan et al., 2023) The data look below in 1 The data set contains Telugu-English code-mixed language comments, along with the labels and text IDs. The labels indicate whether the comment is offensive language or hate speech Since there were only 2 category comments in the training set and no such labeled comments found in the test set, we have discarded those hate or offensive labeled samples from our corpus. Hence, we have viewed this as a binary classification In this dataset, we have a total of 6930 Telugu tweet comments, of which tried 5355. The remaining 1575 samples were part of the testing set we have tried to find a suitable representation for Telugu language text data and also studied the effect of stemming and stop word removal by applying various methods. The details of the proposed methodology are presented in the following sections.

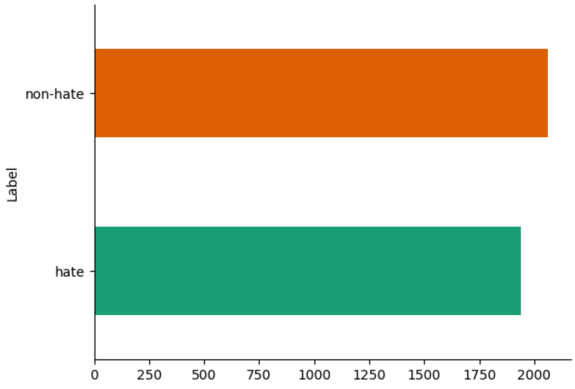

Figure 1: Categorical distributions

4 Methodology

The described methodology is extensively detailed in this section. Figure 2 illustrates the comprehensive flowchart outlining the process of identifying hate speech and offensive tweets. The section commences with an overview of the dataset,1 followed by an in-depth exploration of the proposed models. We used the DravidianLangTech dataset to carry out our experimentations. Each comment in the dataset is labeled as offensive (Hate) or not offensive (NOT hate). We experimented with different conventional classifiers such as (i) Support Vector Machine (SVM), (ii) Naïve Bayes (NB), (iii) Random Forest (RF), and (iv) Convolutional Neural Network (CNN.

We employed character N-gram TF-IDF features in conventional machine learning classifiers and the DNN model. In the case of CNN, we used a 100-dimensional word embedding vector to represent each word of the tweet. In our case, we fixed the maximum word length for comments to 10000 words. As in our dataset, we found most of the comments were less than 100 words, due to this we chose the maximum length of 1000 words for the experiments. It means we curtailed the words of the comments that have more than 42 words and padded for the comments that have less than 100 words. It means a comments matrix with a (42 × 100) dimension is passed through the CNN network to extract features from the comments to classify them into hate and NOT hate classes. A convolutional neural network is a multi-layered neural network with a unique design that is used to recognize complex data features.

We implemented one layered convolutional neural network with 128 filters of 3-gram to extract features from the text and then this feature is passed through the Max Pooling operation and the activation function to get the feature map and this feature map is then used by the dense layer to classify tweets into hate and not-hate classes.

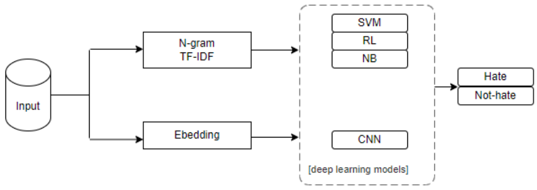

Figure 2: Proposed model architecture

4.1 Conclusion

Our Task focused on evaluating the effectiveness of diverse machine learning models in the identification of hate speech and offensive/non-offensive content in comments. We conducted a comparative analysis, assessing Support Vector Machine (SVM), Naive Bayes, Random Forest, and Convolutional Neural Network (CNN) models. The performance evaluation relied on the weighted F1 score, a pivotal metric that considers both precision and recall. show in figure 3 The results underscored the exceptional success of the CNN model, achieving a weighted F1 score of 0.7711.

Model	**Hate**	**Not-hate**
CNN	0.76	0.78
SVM	0.74	0.75
NB	0.68	0.76
RF	0.7	0.73

Table 1: F1 Score in each model

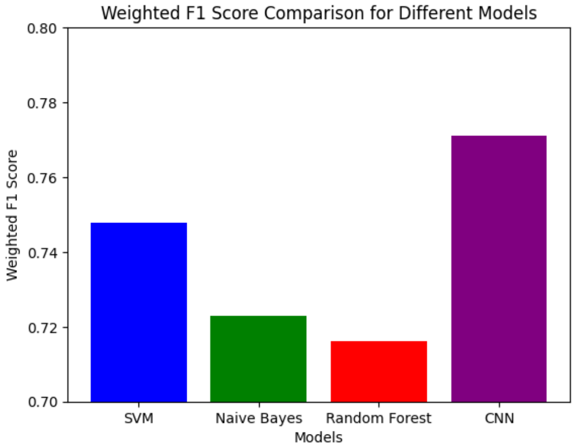

Figure 3: F1- Score Weighted

This emphasizes CNN's ability to effectively balance accurate identification of offensive or non-offensive content while minimizing false positives and negatives. Although SVM, Naive Bayes, and Random Forest demonstrated commendable performances, the superior performance of the CNN model highlights the significance of deploying deep learning techniques for intricate tasks such as identifying hate and offensive speech. The visual representation through the bar graph further emphasized the comparative performance, with the CNN model standing out due to its higher weighted F1 score. Looking ahead, the integration of machine learning and deep learning models, alongside ongoing parameter tuning for the CNN-based model, presents promising avenues for enhancing the efficiency of hate speech identification in comments.

Acknowledgments

The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Vimala Balakrishnan, Vithyatheri Govindan, and Kumanan N Govaichelvan. 2023. Tamil offensive language detection: Supervised versus unsupervised learning approaches. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(4):1–14.

Bharathi Raja Chakravarthi, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2020. A sentiment analysis dataset for code-mixed malayalam-english. *arXiv preprint arXiv:2006.00210*.

Simeon Edosomwan, Sitalaskshmi Kalangot Prakasan, Doriane Kouame, Jonelle Watson, and Tom Seymour. 2011. The history of social media and its impact on business. *Journal of Applied Management and entrepreneurship*, 16(3):79.

Avishek Garain, Atanu Mandal, and Sudip Kumar Naskar. 2021. Junlp@ dravidianlangtech-eacl2021: Offensive language identification in dravidian langauges. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 319–322.

Kushal Kedia and Abhilash Nandy. 2021. indicnlp@ kgp at dravidianlangtech-eacl2021: Offensive language identification in dravidian languages. *arXiv preprint arXiv:2102.07150*.

Teo Keipi, Matti Näsi, Atte Oksanen, and Pekka Räsänen. 2016. *Online hate and harmful content: Cross-national perspectives*. Taylor & Francis.

Thomas Mandl, Sandip Modha, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Prasenjit Majumder, Johannes Schäfer, Tharindu Ranasinghe, Marcos Zampieri, Durgesh Nandini, et al. 2021. Overview of the hasoc subtrack at fire 2021: Hate speech and offensive content identification in english and indo-aryan languages. *arXiv preprint arXiv:2112.09301*.

Olumide Ebenezer Ojo, Olaronke Oluwayemisi Adebanji, Hiram Calvo, Alexander Gelbukh, Anna Feldman, and Grigori Sidorov. 2023. Hate and offensive content identification in indo-aryan languages using transformer-based models.

R Rajalakshmi and B Yashwant Reddy. 2019. Dlrg@ hasoc 2019: An enhanced ensemble classifier for hate and offensive content identification. In *FIRE (Working Notes)*, pages 370–379.

Ratnavel Rajalakshmi, Ankita Duraphe, and Antonette Shibani. 2022. DLRG@DravidianLangTech-ACL2022: Abusive comment detection in Tamil using multilingual transformer models. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 207–213, Dublin, Ireland. Association for Computational Linguistics.

M Shahiki Tash, Z Ahani, Al Tonja, M Gemeda, N Hussain, and O Kolesnikova. 2022. Word level language identification in code-mixed kannada-english texts using traditional machine learning algorithms. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 25–28.

Atnafu Lambebo Tonja, Olga Kolesnikova, Muhammad Arif, Alexander Gelbukh, and Grigori Sidorov. 2022a. Improving neural machine translation for low resource languages using mixed training: The case of ethiopian languages. In *Mexican International Conference on Artificial Intelligence*, pages 30–40. Springer.

Atnafu Lambebo Tonja, Olumide Ebenezer Ojo, Mohammed Arif Khan, Abdul Gafar Manuel Meque, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2022b. Cic nlp at smm4h 2022: a bert-based approach for classification of social media forum posts. In *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, pages 58–61.

Mesay Gemeda Yigezu, Girma Yohannis Bade, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Multilingual hope speech detection using machine learning.

Mesay Gemeda Yigezu, Selam Kanta, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Habesha@ dravidianlangtech: Abusive comment detection using deep learning approach. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 244–249.

Mesay Gemeda Yigezu, Tadesse Kebede, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023c. Habesha@ dravidianlangtech: Utilizing deep and transfer learning approaches for sentiment analysis. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 239–243.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023d. Transformer-based hate speech detection for multi-class and multi-label classification.

Mesay Gemeda Yigezu, Moges Ahmed Mehamed, Olga Kolesnikova, Tadesse Kebede Guge, Alexander Gelbukh, and Grigori Sidorov. 2023e. Evaluating the effectiveness of hybrid features in fake news detection on social media. In *2023 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 171–175. IEEE.

Mesay Gemeda Yigezu, Atnafu Lambebo Tonja, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbukh. 2022. Word level language identification in code-mixed kannada-english texts using deep learning approach. In *Proceedings of the 19th International Conference on Natural Language Processing (ICON): Shared Task on Word Level Language Identification in Code-mixed Kannada-English Texts*, pages 29–33.

Mesay Gemeda Yigezu, Michael Melese Woldeyohannis, and Atnafu Lambebo Tonja. 2021. Multilingual neural machine translation for low resourced languages: Ometo-english. In *2021 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 89–94. IEEE.