Fida @DravidianLangTech 2024: A Novel Approach to Hate Speech Detection Using Distilbert-base-multilingual-cased

Fida Ullah, Muhammad Tayyab Zamir, Muhammad Arif, M.Ahmad, E Felipe-Riveron, A. Gelbukh
Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC)
Corresponding: fullah-2022@cic.ipn.mx

Abstract

In the contemporary digital landscape, social media has emerged as a prominent means of communication and information dissemination, offering a rapid outreach to a broad audience compared to traditional communication methods. Unfortunately, the escalating prevalence of abusive language and hate speech on these platforms has become a pressing issue. Detecting and addressing such content on the Internet has garnered considerable attention due to the significant impact it has on individuals. The advent of deep learning has facilitated the use of pre-trained deep neural network models for text classification tasks. While these models demonstrate high performance, some exhibit a substantial number of parameters. In the DravidianLangTech@EACL 2024 task, we opted for the Distilbert-base-multilingual-cased model, an enhancement of the BERT model that effectively reduces the number of parameters without compromising performance. This model was selected based on its exceptional results in the task. Our system achieved a commendable macro F1 score of 0.6369, securing the 18th position among the 27 participating teams.

1 Introduction

In recent times, the exponential growth of Internet technology has led to a surge in the user base, accompanied by the emergence of various social platforms. These platforms provide a space for netizens to freely express their opinions, often leveraging anonymous features. Consequently, this freedom has led to an increase in hate speech (Shahiki-Tash et al., 2023a; Yigezu et al., 2023a) and offensive content on the Internet. Addressing this issue is crucial, as it not only causes distress but also poses severe risks, including mental health concerns and potential instances of suicide. Given the enormous volume of comments generated daily on the Internet, manual moderation is impractical. Therefore, the integration of artificial intelligence methods becomes imperative. However, identifying hate speech and offensive content poses several challenges. Firstly, social media posts encompass multiple languages and diverse writing styles. The presence of irregular writing and the evolution of new Internet expressions further complicate the detection task. Additionally, some comments may not overtly contain derogatory language but instead employ implicit or ironic attacks, adding another layer of complexity.

Furthermore, the absence of a clear standard for the definition of hate speech contributes to the intricacy of the task. Model performance is highly contingent on the training dataset, influenced by the annotator's perspectives to a considerable extent. In response to these challenges, the NLP community has introduced various tasks focused on hate speech identification, including the Dravidianlangtech-eacl2021 shared task (Saha et al., 2021; Priyadharshini et al., 2023b; Yigezu et al., 2023b). This particular task is dedicated to identifying hate speech and offensive content in both Telugu and English languages.In the realm of NLP, numerous tasks such as identifying hope speech (Shahiki-Tash et al., 2023b), sentiment analysis (Tash et al., 2023), and detecting hate speech utilize various models like deep learning (Yigezu et al., 2022), transformers (Tonja et al., 2022), and traditional machine learning (Tash et al., 2022).

In this paper, our focus is on hate speech using the pre-trained model Distilbert-base-multilingual-cased (Ghosh and Senapati, 2022; Yigezu et al., 2023c), which builds upon the BERT model (Renjit and Idicula, 2020; Yigezu et al., 2023d), by significantly reducing the number of parameters, consequently enhancing training speed. The structure of this article encompasses an exploration of related research on hate speech and offensive speech recognition (Section 2), an explanation of the model used in the task (Section 3), an elucidation of the exper-

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 85–90

imental procedure (Section 4), a presentation of the experimental results and their analysis (Section 5), and a comprehensive summary of this work (Section 6).

2 Related work

Numerous research endeavors have sought to identify and address abusive comments across various languages; however, there is a noticeable research gap in the domain of low-resource languages. (Priyadharshini et al., 2023b, 2022) and addressed this gap by conducting a shared task at ACL 2022, focusing on detecting categories of abusive comments on social media. Their study encompassed comments in Tamil and a code-mixed language featuring both Tamil and English scripts (Akhter et al., 2021) contributed a comprehensive investigation into abusive language detection, specifically in Urdu and Roman Urdu comments. Employing a diverse set of machine learning and deep learning models, the author evaluated the performance of five ML models (Naive Bayes, Support Vector Machine, Instance-Based Learning, Logistic Regression, and JRip) and four DL models (CNN, LSTM, BLSTM, and Convolutional LSTM) across two datasets— one comprising tens of thousands of Roman Urdu comments and another with over two thousand comments in Urdu. The CNN exhibited notable superiority, achieving accuracy rates of 96.2% for Urdu and 91.4% for Roman Urdu, establishing itself as the most adept model in identifying abusive language in these linguistic contexts. Some researchers have explored multiple methods independently to determine the most effective model rajalakshmi2022dlrg employed three methodologies—Machine Learning, Deep Learning, and Transformer-based modeling. For Machine Learning, eight algorithms were implemented, with Random Forest yielding the best results for the Tamil-English dataset. In Deep Learning, Bi-Directional LSTM outperformed other models, especially with pre-trained word embeddings. In Transformer-based modeling, IndicBERT and mBERT with fine-tuning were employed, with mBERT delivering the most favorable results. (Eshan and Hasan, 2017) delved into machine learning algorithms for Bengali abusive text detection, revealing that SVM with a Linear kernel performed optimally using trigram TfidfVectorizer features. (Djuric et al., 2015)proposed a binary classification model for hate speech detection, utilizing advanced deep learning techniques such as continuous bag-of-words (CBOW) and paragraph2vec to represent text in a low-dimensional vector space, achieving an AUC value of 0.80. (Kedia and Nandy, 2021) developed an offensive content classification model for Dravidian code-mixed languages (Tamil, Malayalam, and Kannada) using transformer-based models—BERT and RoBERTa. Renjit and Idicula (2020) employed word embedding for the Manglish dataset, achieving a weighted F1-score of 0.53 and 0.48 using Keras Embedding and Doc2Vec approaches for sentence representation, respectively. (Burnap and Williams, 2015) focused on hate speech detection by extracting unito-five-gram features from a dataset of 450,000 tweets. Using ensemble learning, logistic regression, and SVM classification techniques, they predicted hate speech with an accuracy of 89% using the ensemble learning model.

3 Methodology

The methodology employed in this study leverages the Distilbert-base-multilingual-cased model (Sanh et al., 2019), a pre-trained transformer-based language model, to address the research objectives. The model, developed by Hugging Face, has been fine-tuned for multilingual understanding and exhibits capabilities across various languages. To train and evaluate the Distilbert-base-multilingual-cased model a diverse and representative dataset encompassing multiple languages is gathered. This dataset spans domains relevant to the research focus. The collected data undergoes preprocessing to ensure uniformity and compatibility with the model's requirements. This step involves tokenization, stemming, and the removal of irrelevant information to enhance the model's efficiency. The Distilbert-base-multilingual-cased model is configured with specific parameters to align with the research objectives. This includes setting appropriate learning rates, batch sizes, and training epochs for optimal performance. The pre-trained Distilbert-base-multilingual-cased model is fine-tuned on the task-specific dataset as seen in figure 1.

This involves updating the model's weights based on the unique characteristics of the dataset and the objectives of the research. Fine-tuning enhances the model's ability to grasp nuances within the target domain. To assess the generalizability of the model, a cross-validation strategy is employed. The dataset is partitioned into training, validation,

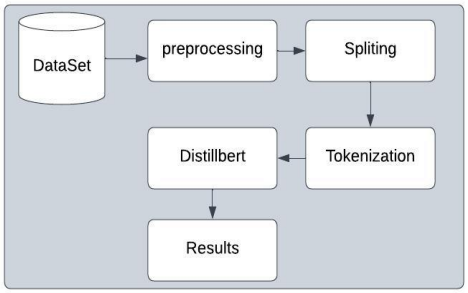

Figure 1: workflow our proposed method

and test sets, ensuring that the model is rigorously evaluated on diverse samples. This step aids in mitigating over-fitting and enhances the robustness of the model. The model's performance is evaluated using appropriate metrics tailored to the research task. Common metrics such as precision, recall, and F1 score are computed to quantify the model's effectiveness in capturing patterns and making accurate predictions. The outcomes of the experiments are analyzed comprehensively to draw meaningful conclusions. This involves an in-depth examination of the model's predictions, identification of potential challenges, and exploration of areas for improvement. Throughout the implementation of the Distilbert-base-multilingual-cased model, ethical considerations are prioritized. Data privacy and confidentiality are ensured, and the research adheres to established guidelines for responsible AI usage.

3.1 Data Description

The Shared Task on Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu) hosted by DravidianLangTech@EACL 2024, as proposed by (B et al., 2024; Priyadharshini et al., 2023a) addresses the challenge of mitigating offensive content within social media through a post-classification approach. This task seeks to advance methodologies and language models specifically tailored for code-mixed data in languages with limited linguistic resources. It recognizes the inadequacy of models trained on monolingual data in capturing the intricate semantics inherent in code-mixed datasets. The task at hand involves hate speech classification within a dataset consisting of 4000 sentences expressed in both native Telugu script and Romanized Telugu. The training dataset encompasses 2061 sentences labeled as non-hate and 1939 as hate. Furthermore, a test dataset comprising 500 sentences is provided, devoid of labeled categories. The primary objective is to employ machine learning models to predict whether these 500 test sentences can be classified into hate speech or non-hate categories. This classification is to be based on the discernment of patterns learned from the labeled training data. Figure 2 illustrating the distribution of Hate and Non-Hate labels.

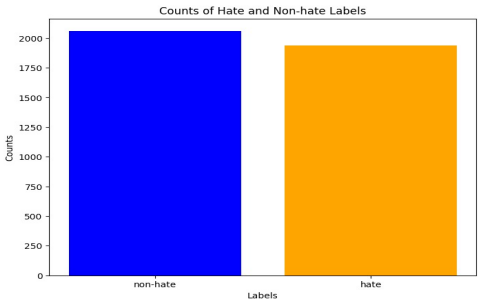

Figure 2: Value counts for hate and non hate comments in training dataset

3.2 Data Preprocessing

Upon reception from the organizer (B et al., 2024), the received data was partitioned into two distinct segments: the training set and the testing set. However, prior to deploying this data for training purposes, it is imperative to undergo a pre-processing phase. The current state of the data renders it unsuitable for effective model training, necessitating a transformation into a structured and intelligible format compatible with the requisites of the training process. An essential facet of pre-processing involves the removal of extraneous elements inherent in the data. Such elements encompass hyperlinks, HTML tags, numeric values, and symbols, which possess the potential to impede the training process or introduce noise into the dataset. The elimination of these undesirable components serves to enhance the cleanliness and focus of the data, thereby empowering the model to discern patterns and relationships within the textual content more effectively. Upon the successful removal of these extraneous elements, the data becomes more amenable to model training. Pre-processing, in this context, facilitates a concentrated focus on pertinent linguistic features and patterns within the text, thereby augmenting the model's capacity for generalization and accurate predictions or classifications. These pre-processing steps are instrumental in optimizing the dataset for subsequent model training, ultimately contributing to enhanced accuracy and meaningful outcomes in subsequent analytical pur-

suits or practical applications.

4 Result

The performance parameter used to assess the detection model's overall efficacy is the macro average F1-score. It is computed by taking the average of all classes after determining the F1-score for each class separately. The macro average F1-score assigns equal weight to each class by computing the average after considering each class's performance individually, regardless of class size or imbalance. In Telugu, we obtained a macro F1-score of 0.6369. securing the 18th position among the 27 participating teams. For a comprehensive summary of the outcomes attained by all competing teams, as outlined in Table 1, offering a detailed overview of the performance metrics and points garnered by each participant in the competition.

Table 1 Result of all participants in Telugu hate speech

Team	Run	F1-score	Rank
Sandalphon	1	0.7711	1
Selam	2	0.7711	1
Kubapok	1	0.7431	3
DLRG1	1	0.7101	4
DLRG	1	0.7041	5
CUET_Binary	2	0.7013	6
CUET_OpenNLP	1	0.6878	7
Zavira	1	0.6819	8
IIITDWD-zk_lstm	2	0.6739	9
lemlem	1	0.6708	10
Mizan	1	0.6616	11
byteSizedLLM	1	0.6609	12
pinealai	1	0.6575	13
IIITDWD_SVC	2	0.6565	14
MUCS	3	0.6501	15
Lemlem-eyob	2	0.6498	16
Tewodros	2	0.6498	16
Fida	2	0.6369	18
Lidoma	1	0.6151	19
MasonTigers	1	0.5621	29
Habesha	1	0.5284	21
MasonTigers	1	0.4959	22
CUET_DASH	3	0.4956	23
Fango	1	0.4921	24
Tayyab	1	0.4653	25

5 Error analysis

The distilbert model excels in detecting hate speech, boasting high accuracy with a notable true positive count. However, challenges arise with false positives, even in balanced datasets. This necessitates careful analysis, urging strategic adjustments like fine-tuning parameters to enhance overall efficiency. Ongoing evaluations on validation and test sets are vital for adapting the model and ensuring reliable performance.

6 Limitations

The utilization of pre-trained transformers, specifically multilingual distilbert, presents a significant consideration in our endeavors to detect hate speech. While leveraging these pre-trained models can enhance our comprehension of textual data, their effectiveness may be limited by the specificity of the pre-training corpus. This potential limitation could result in a mismatch with the unique characteristics of hate speech, underscoring the necessity for meticulous fine-tuning to ensure optimal performance. Moreover, the inherent linguistic complexities of Telugu Codemixed Text may pose challenges impacting the model's ability to discern subtle patterns. Consequently, further investigation and refinement are warranted to address these challenges and enhance the model's accuracy in fake news detection for Telugu Codemixed Text.

7 Conclusion

Hate and offensive posts on social media can put the victim in hazardous circumstances and increase their risk of mental health issues like depression, insomnia, and in extreme cases, suicide. As a result, identifying such hateful and harmful social media information is crucial for jobs involving natural language processing. Our work presents an improved Distilbert-base-multilingual-cased model for identifying hateful and abusive tweets from Telugu text. For Telugu language tweets, the suggested fine-tuned Distilbert-base-multilingual-cased model achieved 0.6369% accuracy. The role of embedding with an improved BERT model for higher classification performance may be investigated in subsequent work.

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

Muhammad Pervez Akhter, Zheng Jiangbin, Irfan Raza Naqvi, Mohammed AbdelMajeed, and Tehseen Zia. 2021. Abusive language detection from social media comments using conventional machine learning and deep learning approaches. *Multimedia Systems*, pages 1–16.

Premjth B, Bharathi Raja, Prasanna Kumar Kumaresan, Saranya Rajiakodi, Sai Prashanth Karnati, Sai Rishith Reddy Mangamuru, and Janakiram Chandu. 2024. Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu). In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics.

Pete Burnap and Matthew L Williams. 2015. Cyber hate speech on twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & internet*, 7(2):223–242.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Shahnoor C Eshan and Mohammad S Hasan. 2017. An application of machine learning to detect abusive bengali text. In *2017 20th International Conference of Computer and Information Technology (ICCIT)*, pages 1–6. IEEE.

Koyel Ghosh and Apurbalal Senapati. 2022. Hate speech detection: a comparison of mono and multilingual transformer model with cross-language evaluation. In *Proceedings of the 36th Pacific Asia Conference on Language, Information and Computation*, pages 853–865.

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

Bharathi Raja andS Malliga andCN SUBALALITHA Priyadharshini, Ruba andChakravarthi, Premjith and-Murugappan Abirami S V, Kogilavani andB, and Prasanna Kumar Kumaresan. 2023a. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, Varna, Bulgaria. Recent Advances in Natural Language Processing.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Cn, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumaresan. 2022. Overview of abusive comment detection in tamil-acl 2022. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 292–298.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, S Malliga, Subalalitha Cn, SV Kogilavani, B Premjith, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023b. Overview of shared-task on abusive comment detection in tamil and telugu. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 80–87.

Sara Renjit and Sumam Mary Idicula. 2020. Cusatnlp@ hasoc-dravidian-codemix-fire2020: identifying offensive language from manglishtweets. *arXiv preprint arXiv:2010.08756*.

Debjoy Saha, Naman Paharia, Debajit Chakraborty, Punyajoy Saha, and Animesh Mukherjee. 2021. Hate-alert@ dravidianlangtech-eacl2021: Ensembling strategies for transformer-based offensive language detection. *arXiv preprint arXiv:2102.10084*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Moein Shahiki-Tash, Jesús Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023a. Lidoma at homomex2023@ iberlef: Hate speech detection towards the mexican spanish-speaking lgbt+ population. the importance of preprocessing before using bert-based models. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023)*.

Moein Shahiki-Tash, Jesús Armenta-Segura, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. Lidoma at hope2023iberlef: Hope speech detection using lexical features and convolutional neural networks. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS. org*.

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.

Moein Tash, Jesus Armenta-Segura, Zahra Ahani, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Lidoma@ dravidianlangtech: Convolutional neural networks for studying correlation between lexical features and sentiment polarity in tamil and tulu languages. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 180–185.

Atnafu Lambebo Tonja, Mesay Gemeda Yigezu, Olga Kolesnikova, Moein Shahiki Tash, Grigori Sidorov, and Alexander Gelbuk. 2022. Transformer-based model for word level language identification in code-mixed kannada-english texts. *arXiv preprint arXiv:2211.14459*.

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

Mesay Gemeda Yigezu, Tadesse Kebede, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023b. 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. 2023c. 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. 2023d. 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.