Tewodros@DravidianLangTech 2024: Hate Speech Recognition in Telugu Codemixed Text

Tewodros Achamaleh[2],Lemlem Eyob Kawo [1], Ildar Batyrshin[1], and Grigori Sidorov[1]
[1]Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City, Mexico
[2]Rift Valley University (RVU), Department of Technology and Engineering, Addis Ababa, Ethiopia

Abstract

This study goes into our team's active participation in the Hate and Offensive LanguageDetection in Telugu Codemixed Text (HOLDTelugu) shared task, which is an essential component of the DravidianLangTech@EACL 2024 workshop. The ultimate goal of this collaborative work is to push the bounds of hatespeech recognition, especially tackling the issues given by codemixed text in Telugu, where English blends smoothly. Our inquiry offers a complete evaluation of the task's aims, the technique used, and the precise achievements obtained by our team, providing a full insight into our contributions to this crucial linguistic and technical undertaking.

1 Introduction

The prevalence of hate speech and provocative language in online interactions has raised serious concerns about their negative impact on people and society. With its unequal reach and simplicity of use, social media has become a breeding ground for the spread of such toxic information. While considerable strides have been made in identifying hate speech in English, the issue remains largely unexplored for Dravidian languages such as Telugu.

Telugu, which is spoken by almost 80 million people in India, is a sophisticated language that incorporates words and phrases from other languages, mostly English. This linguistic feature creates additional challenges for effectively detecting hate speech in Telugu literature. Recognizing the need to bridge this knowledge gap and support innovation in Telugu hate speech detection, the DravidianLangTech@EACL 2024 workshop included the Shared task on Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu).The collaborative effort aims to bring together diverse experts in order to create effective methods for detecting hate and abuse terms in Telugu codemixed text.

2 Related Work

While hate speech identification in English has gained significant study interest, the topic for Dravidian languages like Telugu remains relatively new and unexplored. However, in the past year, tasks such as sentiment analysis have been done on the Tulu-English code-mixing language dataset (Tash et al., 2023), and the following earlier research gives vital facts and insights linked to the HOLD-Telugu joint task:

In the work of (Priyadharshini et al., 2022), the authors attempt to present an overview of detecting abusive comments and hope speech involving homophobia, misandry, counter-speech, misogyny, xenophobia, and transphobia using data in Tamil and Tamil-English code-mixed languages. along with dataset details and participant findings.

In (Pavlopoulos et al., 2019), the authors provide the evaluation of two powerful baselines for offensive language identification (Perspective) and categorization (BERT). Their experiment shows that Perspective outperformed BERT in detecting toxicity, whereas BERT outperformed Perspective in categorizing the offensive type. In the SEMEVAL-2019 OFFENSEVAL competition, Perspective ranked 12th in detecting an offensive post, while BERT ranked 11th in categorizing it.

In (Chakravarthi et al., 2022), the authors constructed a multilingual, manually annotated dataset and experimented with machine learning and deep learning algorithms. The dataset contains around 60,000 YouTube comments, including approximately 44,000 comments in Tamil-English, 7000 comments in Kannada-English, and 20,000 comments in Malayalam-English. They make it publicly available on GitHub and Zenodo.

The authors in (Ayele et al., 2022b) build a dataset of 5,267 tweets, and machine learning methods LR and SVM achieve an F1 score of 0.49, whereas NB achieves an F1 value of 0.46. The

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 96–100

deep learning algorithms (LSTM, BiLSTM, and CNN) achieve an equal F1 score of 0.44, the lowest of all models. Am-FLAIR and Am-RoBERTa, two contextual embedding models, attain F1 scores of 0.48 and 0.50, respectively.

The authors in (Shahiki-Tash et al., 2023) conducted experiments using transformer architectures and BERT-based models, present hate speech detection toward the Mexican Spanish-speaking LGBT+ population, and achieve results with a Macro F1 score of 0.73.

The authors in (Shahiki-Tash et al., 2023) presented a word-based tokenization approach to train a convolutional Neural network (CNN).

In (Yigezu et al., 2023), To examine language patterns, the authors use deep learning models such as Long Short-Term Memory (LSTM) and Recurrent Neural Networks (RNNs). The LSTM model, which is a form of RNN, is used to comprehend the context by capturing long-term relationships and detailed patterns in input sequences.

The authors in (Ayele et al., 2022a) examine the primary concerns associated with crowdsourcing annotation for the collection of Amharic hate speech data using Yandex Toloka.An estimated 1,000 tweets are annotated annually, or 5,400 in total. Classification models based on deep learning were developed utilizing LSTM and BiLSTM. Both models achieved an F1-score of 0.44.

3 Methodology

Various machine learning algorithms, including classical methods and deep learning techniques, can be used. Neural networks and deep learning are among the most helpful AI approaches that have been developed (Ahani et al., 2024). RNN is often used as a building block in modern neural networks to detect hate speech (Yigezu et al., 2022). Our team deployed a task of automatic hate and offensive language detection in Telugu codemixed text using a deep learning-based simple neural network (Simple RNN) model.

This design was chosen for its efficacy in capturing temporal dependencies within text data vital for understanding the sequential nature of language, and their ability to model short-term dependencies is useful for applications such as hate speech and offensive language detection. They are computationally simple compared to advanced architectures like Long Short Term Memory (LSTM) or Gated Recurrent Unit (GRU), which allow faster training, especially on small datasets. The interpretability of simple RNNs is advantageous in applications requiring a transparent decision-making process, such as hate speech detection.

The model is trained on a training and validation dataset and evaluated on the test set. We perform pre-processing procedures, including tokenization and padding, removing punctuation marks, and filtering out stop words, to turn the text data into a format appropriate for neural network computation. The general methodology pipeline is shown in Figure 2

3.1 Dataset Analysis

The main point of our research resides in the HOLD-Telugu dataset which is collected from social media (Priyadharshini et al., 2022), carefully curated with 4000 items of Telugu codemixed text. Each entry is extensively annotated for hate or non-hate content, capturing varied linguistic phrases and cultural nuances. Code-Mix is used in nearly all social media networks where individuals speak many languages. The use of code-mixed data in natural language processing (NLP) research is receiving a lot of interest right now (Tash et al., 2022).

Due to the expansion and significance of social media in communication, hate speech detection of social media code-mixed text has been an attractive subject of study in recent years (Tonja et al., 2022). Our investigation delves into the nature of the dataset, analyzing the distribution of hate vs. non-hate samples, the intricacies of codemixing patterns, and the possible issues these aspects represent for hate speech recognition programs. We show the Distribution of Data in Figure 1

3.2 Shared Task Description

The HOLD-Telugu shared task supplied users with a rich dataset of Telugu code-mixed text(Priyadharshini et al., 2023; Premjth et al.,2024"), painstakingly annotated for hate and offensive content. This dataset includes different online sources, including social networking platforms, discussion forums, and news websites. Participants were tasked with constructing models capable of reliably detecting whether a particular comment contained hate or derogatory language. The performance of the suggested models was evaluated using the macro-F1 score, a balanced metric that combines both precision and recall across both classes,ensuring a full and trustworthy assessment.

3.3 Model Architecture

We painstakingly study the model architecture, uncovering the rationale behind each design decision. The embedding layer transforms the discrete word tokens into dense vectors, capturing semantic links between words. The simple RNN layer then analyzes these vectors sequentially, allowing the model to learn from the context and sequence of the words. Dropout regularization is employed to prevent overfitting and increase model generalization. Finally, a thick layer with softmax activation classifies each input as hate or non-hate content. We show the parameters we use in Table 1:

Table 1: parameter Setting

Parameters	Values
embed_units	64
hidden_units	128
dropout	0.5
optimizer	adam
batch_size	64
loss	categorical_crossentropy
epoch	5
activation	SoftMax
restore best weights	TRUE
SimpleRNN layer	early stop
callbacks	32 units

3.4 Experimental setup

Our model attained a test accuracy of 64.9 %, assessed using the macro-F1 score as stated by the shared task. Each entry is extensively annotated for hate or non-hate content and further divided into 70% training, 15% testing, and 15% validation sets. We deconstruct the results, assessing the performance on several types of hate speech and exploring the effects of codemixing on model effectiveness. We identify areas for improvement and discuss critical lessons learned during the trial process.

3.5 Predictions on Unseen Data

To highlight the real-world applicability of our model, we apply it to a separate test dataset, proving its capacity to generalize to previously unseen data. The anticipated categories are saved in a conveniently accessible format, enabling additional research and review.

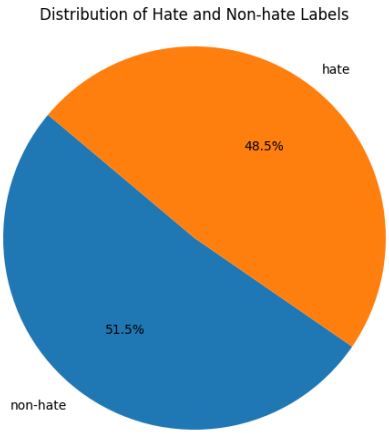

Figure 1: Distribution of Data

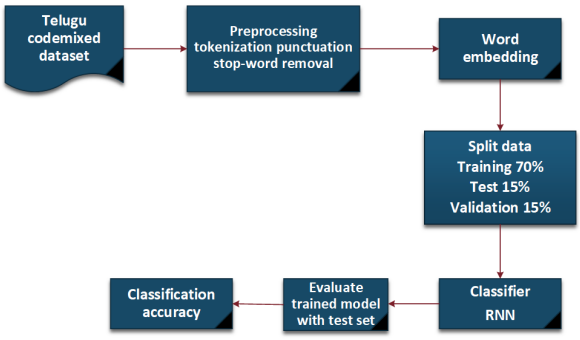

Figure 2: Steps for task evaluations

4 Discussion

While our model achieved promising results, we acknowledge the limits inherent in the simple RNN design and the issues provided by codemixing. In future work, we will have a plan to address potential routes for development, including studying more complicated neural networks,adding additional variables like sentiment analysis, and leveraging larger and more diverse datasets.

5 Conclusion

The HOLD-Telugu collaborative initiative effectively addressed the critical problem of identifying hate speech in Telugu code-mixed text. The shared employment enabled significant discoveries in this under-researched subject by bringing together divergent researchers and promoting teamwork. The large and diverse dataset, high-quality submissions, and intelligent analysis have paved the way for the continuing development of powerful hate speech detection algorithms for Telugu code-mixed text.

The successes of collaborative work go beyond technical advancements. The HOLD-Telugu work

helps to a more inclusive and healthier online environment for Telugu communities by reducing the prevalence of toxic language. The shared task resources and results enable researchers and developers to continue this endeavor, resulting in better online interactions and protecting people from the negative consequences of hate speech.

Looking forward, the HOLD-Telugu collaboration lays the groundwork for future research on hate speech detection in Dravidian languages and code-mixed text. More research into sophisticated NLP methods, such as multilingual language models and fine-tuning procedures, has the potential to significantly improve the accuracy and general applicability of hate speech detection systems. Furthermore, the data from the shared job may be used to create tools and resources that enable people and organizations to reject hate speech and promote online safety.

The achievement of the HOLD-Telugu joint endeavor demonstrates the enormous potential of collaborative research in addressing difficult social issues. Research communities may positively contribute to the establishment of safer and more inclusive online environments for everybody by facilitating open data exchange, fostering diverse perspectives, and concentrating on practical applications.

6 Limitations

The problems faced while interacting with a Telugu language dataset originate from the lack of resources that hinder preprocessing techniques. The dynamics of spoken language create challenges in adaptability for models that were trained on historical data. Proper identification of relevant features for successful training is also a challenging task, especially when dealing with a language that has certain unique linguistic.

Acknowledgements

We extend our heartfelt gratitude to the organizers of the HOLD-Telugu shared task and the DravidianLangTech@EACL 2024 workshop for providing a useful forum for this joint study. We also recognize the contributions of the dataset producers and annotators, whose hard efforts made this research feasible.

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

Zahra Ahani, Moein Shahiki Tash, Yoel Ledo Mezquita, and Jason Angel. 2024. Utilizing deep learning models for the identification of enhancers and super-enhancers based on genomic and epigenomic features. *arXiv preprint arXiv:2401.07470*.

Abinew Ali Ayele, Tadesse Destaw Belay, Seid Muhie Yimam, Skadi Dinter, Tesfa Tegegne Asfaw, and Chris Biemann. 2022a. Challenges of amharic hate speech data annotation using yandex toloka crowdsourcing platform. In *Proceedings of the The Sixth Widening NLP Workshop (WiNLP), Abu Dhabi, United Arab Emirates. Association for Computational Linguistics*.

Abinew Ali Ayele, Skadi Dinter, Tadesse Destaw Belay, Tesfa Tegegne Asfaw, Seid Muhie Yimam, and Chris Biemann. 2022b. The 5js in ethiopia: Amharic hate speech data annotation using toloka crowdsourcing platform. In *2022 International Conference on Information and Communication Technology for Development for Africa (ICT4DA)*, pages 114–120. IEEE.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Vigneshwaran Muralidaran, Navya Jose, Shardul Suryawanshi, Elizabeth Sherly, and John P McCrae. 2022. Dravidiancodemix: Sentiment analysis and offensive language identification dataset for dravidian languages in code-mixed text. *Language Resources and Evaluation*, 56(3):765–806.

John Pavlopoulos, Nithum Thain, Lucas Dixon, and Ion Androutsopoulos. 2019. Convai at semeval-2019 task 6: Offensive language identification and categorization with perspective and bert. In *Proceedings of the 13th international Workshop on Semantic Evaluation*, pages 571–576.

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.

Moein Shahiki-Tash, Jesús Armenta-Segura, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. 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, Selam Kanta, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. 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, 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.