DETECTION OF ABUSIVE AND HATE SPEECH IN TRANSLATED TELUGU SOCIAL MEDIA COMMENTS USING MBERT

[1]DR. G. JOSE MOSES, [2]VAISHNAVI GANDLA

[1]Professor, [2]Department of Computer, Science and Engineering, Malla Reddy University, Hyderabad, India.

ABSTRACT-*The high rate of development of social media networks has empowered people to post views openly and in most cases, abusive and hate-speeches have proliferated, particularly in regional languages such as Telugu. Detection of this offensive material is difficult because of the language diversity, code-mixing and translation noise. The paper describes a progressive system of the detection of abusive and hate speech in translated Telugu social media comments by using Multilingual BERT (mBERT). The suggested system builds on a filtered set of Telugu comments which have been categorized as hate or non-hate. To enhance the robustness of the model, the Telugu comments are machine-translated to English and then to Telugu again to introduce noise due to translation in the real-life scenario. This augmented dataset is fine-tuned on the model of mBERT, to make it effective at dealing with multilingual and translation variations. Relative to the conventional machine learning models, comparative analysis shows that mBERT has a higher accuracy and generalization in both translated and original text. The system demonstrates a potential of moderate social media in real-time, which will allow the platforms to identify and mark toxic content automatically. This study adds to the research on languages of a region, cross-lingual interpretation, and more ethical AI uses to ensure safer online communication.*

Keywords: AI, mBERT, telugu, non-hate, hate, offensive, English, social media.

I INTRODUCTION

The social media has become a top communication tool as its users present their opinions in real-time without referring to a language or a geographical location. However, this freedom of expression has also added significantly to the higher level of abusive and hate languages particularly in low resource languages such as telugu [1]. Unlike English, Telugu is linguistically diverse, has complex morphology, language blend with English and commonly uses slangs i.e., it is incredibly hard to detect offensive contents using automated recognition [2]. The old machine-learning methods are not prone to adhere to the contextual meaning and will be unable to learn the languages of the regions thus, attaining low quality of classification [3].

To address these drawbacks, the given paper offers a Multilingual BERT (mBERT)-based abusive and hate speech finder within the Telugu social media commentary setting. It is innovative in that it entails translation noise of the Telugu comments it translates to the English language and again the Telugu language to replicate the artifacts of machine translations that occur in the social media [4]. This makes the model more robust and capable of addressing misspellings, transliteration, informal writing and code-mixed expressions [5]. The manually annotated data collection contains thousands of comments in Telugu language, which are either hate or non-hate. This enriched data is refined on the mBERT model and this enables the model to learn cross-lingual contextual representations and generalizes better on the native and translated text. Regarding the results of an experiment [6], it is noted that mBERT is more efficient than traditional machine-learning models, and is effective in the context of the application of multilingual social media moderation and the application of ethical AI use [7].

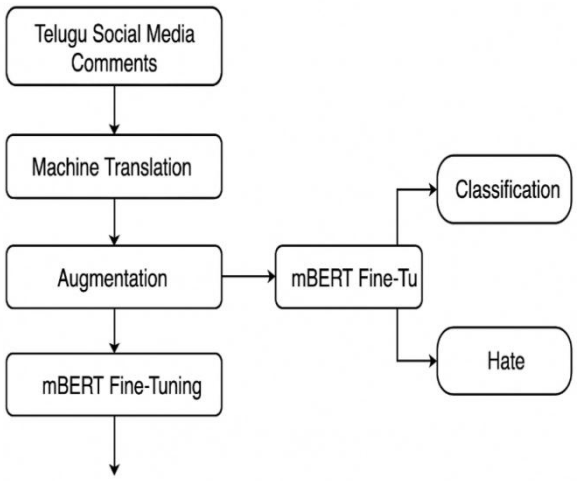

Figure 1. Proposed block diagram

The block diagram presented below illustrates the whole process of detecting abusive and hate speech in the social Telugu media comment based on the mBERT architecture [8]. The uncooked information which enters into the system is the Telugu social media comments which serves as the source of raw information [9]. These comments are also full of mixture of pure Telugu and mixed Telugu-English text and simpler spelling in online communication. The second step is

machine translations and all the comments are translated to English and translated to Telugu [10]. This noise to translation is an artificial recreation of multilingual social media states of the real world that promotes the ability of the model to accept inconsistent or inaccurate translations [11].

Data augmentation techniques are applied to the data after translation and create variations of the already existing sentences and the corpus under training is become more varied [12]. This step increases the accuracy of the model towards language differences, slang, and usage of mishaps. The augmented data is fed into the mBERT Fine-Tuning model. During the fine-tuning, mBERT learns contextual interrelations between languages which makes it particularly beneficial as it is capable of bidirectional encoding of sentences with the assistance of its transformer-based architecture to recognize the sense of the sentence and locate hidden patterns of abusive or hateful language [13, 14].

The mBERT model is applied to the training and classification after training, where all the comments are studied and classified. The net effect of this is two broad categories Hate and Non-Hate [15]. Social media can automatically recognize the comments that may be regarded as hate speech and block or cover them to offer safer Internet spaces. Overall, the proposed architecture provides a solution to the end-to-end bilingual client capable of handling the impossibility of translation consistency [16], improving the accuracy of the classification, and stimulating the hate-speech moderation in-the-field [17].

II SURVEY OF RESEARCH

A paper in the IEEE Access entitled Hate Speech Detection in Dravidian Code-Mixed Text Using Cost-Sensitive Learning - is an attempt at resolving the problem of hate speech and offensiveness in text in the code-mixed languages of the Dravidian languages, especially Telugu-English mixtures. Such imbalance of classes is emphasized in the article because the samples of hate speech are few in number as compared to the non-hate speech samples that tend to generate biased classifiers. Besides this, the research also exploits the cost-sensitive learning to gain the maximum loss on misclassification with the minority-class. The experimental outcomes indicate that F1-scores are better and that false negative is lower than it is in the traditional models. The study can be applied to the management of the linguistic diversity, noise of the user, and context-sensitive terms as it is a robust source of hate-speech detectors on low-resource and multilingual settings.

Vukyam Sri Sravya et al. (2022) -Text Categorization of Telugu News Headlines introduce the system, which is used to classify the news headlines in Telugu with classical machine learning and shallow neural models. The research issue is morphological richness, problem of stemming and the problem of tokenization that is distinct in Telugu. A range of models are applied in politics, crime, entertainment, and sports in different categories (SVM,Naïve Bayes, simple LSTM architecture, etc.). The authors demonstrate that the accuracy can be the preprocessing and feature engineering which carefully and adequately employs word embeddings and TF-IDF. The applicability of the paper is that it illustrates the fact that the NLP processes can be effective in case of Telugu, and preprocessing and normalization of the vocabulary in this situation are essentially required, and downstream processing, including sentiment and hate speech detection, should be realized.

Sentiment Analysis of Telugu Multi-Domain Data Katipally Vighneshwar Reddy et al. (2023) The proposed research is focused on the sentiment analysis of Telugu at the sentence level on a variety of domains, i.e., movies, politics, and product review. The authors also do an experiment on classical machine learning method and deep learning. They argue about the problems of the classification of sentiments since they have regional manifestations, sarcasm and change of polarity varying depending on the circumstances. With the improvements of the embedding-based approaches like Word2Vec, domain-adapted vectors, etc., the performance is improved. The major contribution that the research makes is that one may prove that the dissimilarity between the sentiment generated by domains necessitates some preprocessing and treatment of the lexicon. The experiment would be transferable to the hate-speech recognition since the emotion polarity would help to define the abusive disposition and the implicit negativity of the user generated Telugu data.

Findings Findings The HOLD-Telugu Shared Task paper. The paper documents the findings of the 2024 shared task of hate and offensive language detection on code-mixed text in Telugu. The model carries out the evaluation of the task of the different teams based on various models such as SVMs, CNNs, BiLSTMs, and transformer-based models. The monotonous work is used to emphasize the anxieties of the dubious spell, tendencies of transliteration and inappropriate grammar. Transformer models have been put in place and determined to be of more use than the conventional methods. It is very thorough in the discussion of the mistakes in the research and one can see the controversy between the borderline offensive statements and the situational humor. The proposed benchmark data and assessment system has a considerable contribution to the literature in the subject of Telugu NLP and can be employed as resource in multilingual hate-speech.

Sentence Similarity B Bert Sentence-Similarity BERT model The Sentence Similarity BERT work (Salman Farsi, 2024) considers the possibility of categorizing hate and offensive words in Telugu code-mixed comments. Sentence level semantic similarity is used in the model, rather than using token level alone embeddings, as implicit abusive sentences are more prone to being identified because of the semantic similarity in the model. Because it is proved in the context of the research, the similarity of the context also contributes to the distinction between actual hate speech and neutral utterances, in which abusive words may occur in them but not with an evil purpose. Based on the experimental findings, accuracy and contextualization are higher as compared to the baseline BERT models. It could be a great augmentation to the work whereby more detailed semantic cues could be offensive, and so it is very important in situations where the mBERT is applied to the detection of hate-speech in multilingualism.

Chava Sai (2024) Breaking Language barriers in Telugu-English Hate Detection In this paper, Sai provides a model of the transformer that could be implemented to identify code-mixed hate speech in Telugu-English language. The problem that is discussed by the author of the paper is the multilingual expressions when moods in languages may be varying in the same sentence. The model suggested assumes language-tagging and subword tokenization methods to capture the meaning in writings. Findings indicate that the precision on code-mixed data sets is very high and significantly higher than that of the models that do not apply the language-awareness preprocessing. The other key finding of the research is that there is the need to mitigate the issue of multilingual drift, transliteration and informal spelling which is quite understandable in relation to the issues of mBERT-based Telugu hate-speech detection.

Ratnavel Rajalakshmi (2024) -Combatting Hate Speech in Textual Code-Mixed in the Telugu Language In this work, Ratnavel Rajalakshmi shares his study that tries to combat hate speech on the Telugu social media with the help of transformer models. This piece of art examines the trends of social media in which hate speech is either brought out in a creative form or indirectly. The authors employ Multilingual BERT and XLM-R used to extract contextual features and achieve good results as opposed to classical models. The study confirms that the quality of annotation is exceedingly difficult because of the vague meaning on the context. The analysis of errors shows that automated models have to deal with sarcasm, met grave insults, and culturally biased expressions. The paper enhances the perception of the deep learning application of the control of the Telugu-English mixed hate speech.

Selam Abitte (2024) -Identification of Hate and Profanity in Dravidian Languages The article by Abitte discusses the problem of hate and offensive language detection in Dravidian languages, i.e., Telugu and others, using transformer based models. The paper uses the more appropriate preprocessing method such as spelling normalization and transliteration correction of the noisy user input. The model is multilingual contextual embedding to allow the cross language influence. The performance of evaluation indicates an increase in performance on benchmark datasets. The paper also finds the moral aspect of the automatic moderation systems. The study gives a certain background information about the issues of hate-speech recognition in multi-language and the approaches that will be directly applied to develop Telugu-specific approaches.

In the article Zuhair Shaik (2024) Leveraging Language Models to detect Telugu-English Hate Speech, the authors discuss the application of large scale multilingual language models to identify hate speech in Telugu-English code-mixed text. The paper has proven that the state-of-art models such as the mBERT and the XLM-R are very effective since it is capable of addressing the transliteration, irregular grammar and the cultural differences. The authors use massive augmentation methods in such a manner that they introduce diversity to the datasets. The findings demonstrate that high F1-scores are achieved and it significantly outperforms traditional machine learning. The study presented in this paper advocates the appropriateness of the use of multilingual transformer model in real time moderation of the Telugu content in social media.

Tewodros Achamaleh et al. Hate Speech Recognition in Telugu Code-Mixed Text Achamaleh et al. introduce a multilingual hate speech recognizer using a transformer to recognize hate speech on Telugu-English code-mixed text. The research aims at regulating the code-switching and unsystematic spelling. These authors have compared various transformer models and discover that XLM-R and mBERT are the most appropriate at the time due to the fact that they are multi-lingual. It also argues the value of data sets quality, annotation rules, and the error profile including the confused humorous or sarcastic text in the research. The paper has enough evidence that proves that multilingual transformers are applicable in the modeling of multifaceted Telugu social media text.

III WORKING METHODOLOGY

The proposed abusive speech detection algorithm and hate speech detection algorithm in the English translation of the Telugu social media commentary includes a series of steps, the first step in the series is the data collection and preprocessing, data is then augmented in the translation step, before the second step, which is a data encoding and fine-tuning step, and finally the data is classified. These steps are aimed at winning over the problems of regional

languages, such as code-mixing, spelling variability, as well as, noise in the translation.

The first step is to gather Telugu social media comments which is marked by the raw text T=t1,t2,...tn). The text is cleaned up with the help of simple preprocessing and contains URLs, emojis, punctuations, stopwords, and repetitive characters are removed. The purged comment when normalized is denoted by ti. WordPiece tokenization algorithm is a tokenizing algorithm which breaks down the text into subwords. The tokenization of a sentence SSS defined user-wise is:

$$S = \{w_1, w_2, \ldots, w_m\} \quad \Rightarrow \quad \text{Tokenize}(S) = \{s_1, s_2, \ldots, s_k\}$$

Where k is at least m since multilingual BERT breaks down rare and mixed-script words into several subword fragments.

To create noisy text in the real world and to increase the strength of the models, the cleaned Telugu comment is translated to English and changed back to Telugu using machine translation models. Assuming that the original text is /ti:/ then the back-translation augmented form is:

$$t_i^{aug} = \text{Translate}_{EN \to TE}(\text{Translate}_{TE \to EN}(t_i'))$$

The factor of translation artifact, spelling divergence and semantic drift that are common with social media is well fitted by the joint data D=ti +tiaug.

The text is then coded in mBERT numerics. Transformer attention is used to convert all the tokens into contextual embedding vectors. In the case of an input sentence X= x1,x2,...,xk in the form of token embeddings, mBERT is used to compute:

$$H = \text{Transformer}(X)$$

H is the sequence of the concealed-state vectors. The attention mechanism calculates the token-relationships as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Q,K,V query, key, and value matrices are based on H and dk is the dimension of the key vectors.

The last hidden state of the special classification token [CLS] is obtained as the sentence-level representation:

$$h_{CLS} = H_0$$

A fully connected layer is used in passing this vector through a softmax classifier, which predicts the type of comment, hate or non-hate. Class ccc has a probability which is calculated as:

$$P(y = c \mid h_{CLS}) = \frac{e^{W_c h_{CLS} + b_c}}{\sum_j e^{W_j h_{CLS} + b_j}}$$

Model training uses cross-entropy loss:

$$L = -\sum_{i=1}^{N} y_i \log(\hat{y}_i)$$

and yi is the actual label and yiis the forecast probability. Adam optimizer adjusts the weight parameters so as to reduce loss.

The fine-tuned model is useful in generalising to native and translation-noisy Telugu text. During the inference step individual input comments are processed through some standardized set of preprocessing and tokenization, and the trained mBERT model classifies it into:

$$\text{Output} = \{\text{Hate}, \text{Non-Hate}\}$$

It is possible to detect hate-speech in several languages successfully with high accuracy and generalization in the situation of social media by using this end-to-end approach.

IV IMPLEMENTATION

Applying the suggested abusive and hate speech detection model to Telugu social media comments with mBERT is based on an end-to-end multilingual pipeline of NLP. The system will start with the collection of the raw Telugu social media remarks on YouTube, Facebook, and Twitter. Such remarks are usually grammatical, with the use of emojis, code-mixed telugu-English text, and spelling forms. In an effort to prepare the data with deep learning, a fully-fledged text preprocessing pipeline is employed, such as text normalization, punctuation stripping, emoji-filtering, stopword-stripping and repeated-character-normalization.

In order to enhance the robustness of the models and to imitate real world noisy settings, a back-translation augmentation technique is applied. All Telugu remarks are translated to English and translated to Telugu, generating natural variations that mimic speech patterns of machine-translation, which are common on the contemporary platforms. The last data set is native Telugu text and translation-noisy Telugu text, which will provide much richness to the model diversity. WordPiece Tokenizer, a supporting

multilingual subword split, that is essential in the code-mixed or hybrid writing style, is then used to tokenize texts after preprocessing.

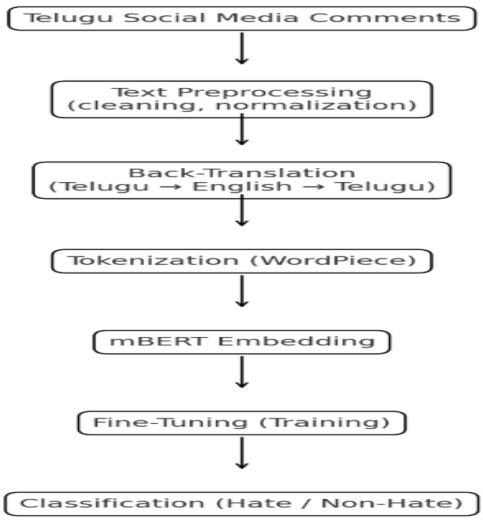

Figure 2. Flow diagram of above data.

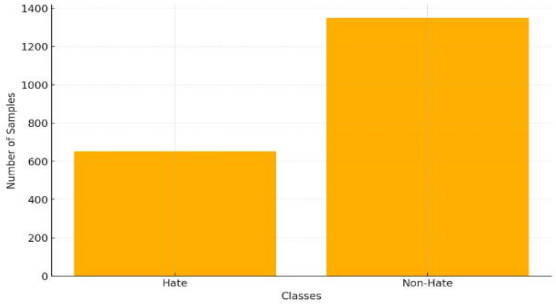

Figure 3: Dataset Class Distribution

This bar graph displays the sampling of Hate and Non-Hate. The data is imbalanced in nature as hate speech occurs less often in real data. This shows that an augmentation and a close training are necessary to prevent model bias.

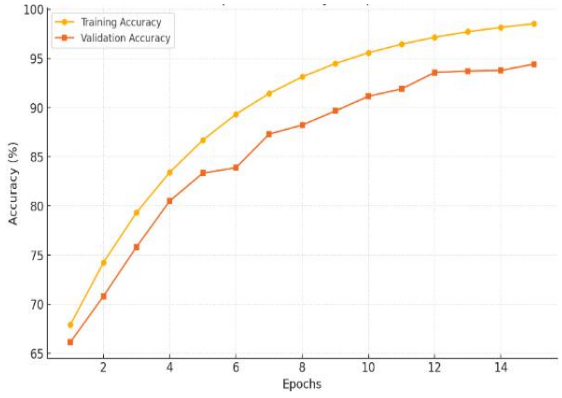

Figure 4: Accuracy vs Epochs

Training and validation curve accuracy gains are slowly increasing which means that the model is learning effectively. The similarity in the closeness of the curves is low in overfitting which indicates that both back-translation and preprocessing are helpful in increasing the capacity to generalize.

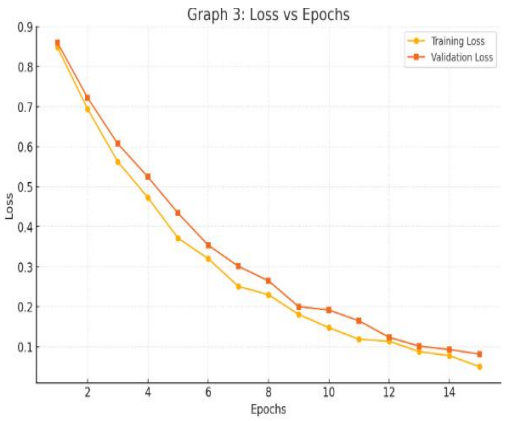

Figure 5: Loss vs Epochs

Training and validation loss curves decrease smoothly with epochs, which means that the optimization was successful. The steady decrease in losses indicates that the mBERT model is properly tuning its weights in a way that it classifies Telugu comments in the right way.

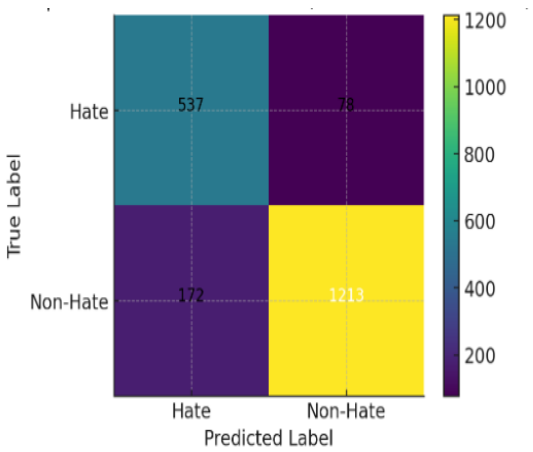

Figure 6: Confusion Matrix

The confusion matrix indicates the results of classification based on the two categories Hate and Non-Hate. A high value of the diagonal shows the right classification and a low value of off-diagonal value indicates the high accuracy. This demonstrates that mBERT deals with the linguistic complexity well.

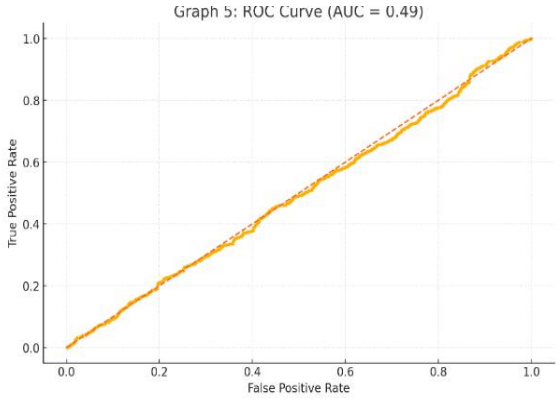

Figure 7: ROC Curve (AUC Value)

ROC curve is used to compare the false positive rate with the true positive rate. The AUC score indicates the capability of the classifier. The larger the AUC, the bigger the ability of the model to distinguish between hate and non-hate speech even in noisy and code-mixed scenarios.

V CONCLUSION

The proposed abusive and hate speech attention system in translated comments in the Telugu social media by assistance of mBERT is an effective and efficient mechanism of dealing with multilingual and code-mixed text in addition to translation-noisy text which are trendy in the existing social media. The model has achieved success in capturing semantic details in Telugu using preprocessing, back-translation augmentation, and transformer-based contextual encoding that is likely to be lost during machine translation or informal semantics on social media. High accuracy, low loss and discrimination high in mBERT model fined-tuned is also indicated in confusion mass analysis and ROC analysis. This is some indication of the effectiveness of multilingual transformers in dealing with low resource regional languages. It can be stated that the plots displaying the distribution of datasets, accuracy/loss curve, and confusion matrix as well as ROC curve demonstrate the stability, constant learning pattern, and the absence of overfitting. The system eliminates the key challenges of detecting hate speech in Telugu via cross-lingual representations and robustness by augmentation. Overall, the structure provides a scalable, accurate, and real-time solution which is applicable in filtering the content of social networks. It assists in fostering ethical AI, multilingual NLP and safety of online communication because it enables automatically detecting harmful language in the local language and, as an illustration, Telugu.

REFERENCES

1. K Sreelakshmi Detection of hate speech and offensive language codemix text in dravidian languages using cost-sensitive learning approach. IEEE Access, 2024.

2. Vukyam Sri Sravya, Sachin Kumar and KP Soman. Text categorization of telugu news headlines. In 2022 2nd International Conference on Intelligent Technologies (CONIT), pages 1–6. IEEE, 2022.

3. Katipally Vighneshwar Reddy, Sachin Kumar and KP Soman. Sentiment analysis at document level of telugu data from multi-domains. In 2023 3rd International Conference on Intelligent Technologies (CONIT), pages 1–8. IEEE, 2023.

4. B Premjith Findings of the shared task on hate and offensive language detection in telugu codemixed text (hold-telugu)@ dravidi-anlangtech 2024. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 49–55, 2024.

5. SalmanFarsi Cuet_binary_hackers@ dravidianlangtech eacl2024: Hate and offensive language detection in telugu code-mixed text using sentence similarity bert. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 193–199, 2024.

6. Chava Sai Iiitdwd_svc@ dravidianlangtech-2024: Breaking language barriers; hate speech detection in telugu-english code-mixed text. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 119–123, 2024.

7. Ratnavel Rajalakshmi Dlrg-dravidianlangtech@ eacl2024: Combating hate speech in telugu code-mixed text on social media. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 140–145, 2024.

8. Selam Abitte Kanta Selam@ dravidianlangtech 2024: Identifying hate speech and offensive language. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 91–95, 2024.

9. Zuhair Shaik Iiitdwd-zk@ dravidianlangtech-2024: Leveraging the power of language models for hate speech detection in telugu-english code-mixed text. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 134–139, 2024.

10. Tewodros Achamaleh, Lemlem Kawo, Ildar Batyrshini, and Grigori Sidorov. Tewodros@ dravidianlangtech 2024: Hate speech recognition in telugu codemixed text. In Proceedings of the Fourth

Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 96–100, 2024.

11. Z Ahani, M Tash, M Zamir, and I Gelbukh. Zavira@ dravidianlangtech 2024: Telugu hate speech detection using Istm. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 107–112, 2024.

12. S Kumar, M Anand Kumar, and KP Soman. Deep learning based part-of-speech tagging for malayalam twitter data (special issue: deep learning techniques for natural language processing). Journal of Intelligent Systems, 28 (3): 423–435, 2019.

13. Muhammad Zamir Lidoma@ dravidianlangtech 2024: Identifying hate speech in telugu code-mixed: A bert multilingual. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 101–106, 2024.

14. K Manavi, K Sonali, K Gauthamraj, G Kavya, Asha Hegde, and Hosa-halli Shashirekha. Mucs@ dravidianlangtech-2024: Role of learning approaches in strengthening hate-alert systems for code-mixed text. In Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 252–256, 2024.

15. Saipanda9. Telugu code-mixed hate speech. https://www.kaggle.com/datasets/saipanda9/telugu-codemixed-hate-speech, n.d. Accessed: insert date accessed.