Findings of the Shared Task on Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu)@DravidianLangTech 2024

Premjith B[1], **Bharathi Raja Chakravarthi** [2], **Prasanna Kumar Kumaresan**[3],
Saranya Rajiakodi[4], **Sai Prashanth Karnati**[1], **Sai Rishith Reddy Mangamuru**[1],
Chandu Janakiram[1]
[1]Amrita School of Artificial Intelligence, Coimbatore, Amrita Vishwa Vidyapeetham, India
[2]School of Computer Science, University of Galway, Ireland
[3]Data Science Institute, University of Galway, Ireland
[4]Central University of Tamil Nadu, India

Abstract

This paper examines the submissions of various participating teams to the task on Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu) organized as part of DravidianLangTech 2024. In order to identify the contents containing harmful information in Telugu codemixed social media text, the shared task pushes researchers and academicians to build models. The dataset for the task was created by gathering YouTube comments and annotated manually. A total of 23 teams participated and submitted their results to the shared task. The rank list was created by assessing the submitted results using the macro F1-score.

1 Introduction

In the present technological era, detecting hate comments on social media has become a crucial and challenging task (Chakravarthi et al., 2023b; Priyadharshini et al., 2022; Prasanth et al., 2022). The growth of internet platforms made it easier for people to disseminate information, including offensive and violent postings and comments. Consequently, it is now crucial to address and mitigate hazardous content in order to automatically maintain online platforms clean (Chakravarthi et al., 2022a,b, 2023b; Chakravarthi, 2023). This is a challenging endeavour because of the complexity of the languages and codemix nature of the contents. However, recently, sophisticated machine learning algorithms and methods were presented to automatically detect and flag offensive remarks, ranging from threats and harassment to hate speech and cyberbullying (Premjith et al., 2023). These technologies analyze the post content and context for hate language. In a codemixed Dravidian language, it is much harder to find hateful words because the text is codemixed and has linguistic properties like morphological richness and agglutinative characteristics (Premjith et al., 2018). Furthermore, sizable datasets of tagged offensive content are needed to train and optimize the AI-based models and make them capable of identifying trends and differentiating between benign and harmful texts (Premjith et al., 2023).

The shared task on hate and offensive language detection in Telugu codemixed text intends to detect hate and offensive content in social media posts and comments written using codemixed Telugu data. The shared task was conceived as a binary class problem, where the dataset has two labels for each data - hate and non-hate. This paper discusses the task in detail and the models submitted to task by the participants.

2 Related Works

(Chakravarthi et al., 2023b) presents a compilation of four datasets extracted from YouTube, which comprise abusive remarks in Tamil and codemixed Tamil-English. Polarity has been ascribed to each dataset's annotations at the comment level. In order to establish baselines for these datasets, the authors conducted experiments utilizing various machine learning classifiers. They subsequently presented their results in F1-score, precision, and recall. In (Chakravarthi et al., 2020), the authors discussed the shared task on offensive language detection in codemixed Dravidian languages conducted as part of the HASOC shared task. (Kumaresan et al., 2021) discuss the overview of the shared task conducted for detecting hate and offensive language detection in Dravidian languages as part of HASCO-FIRE.

(Chakravarthi et al., 2023a) proposed a fusion of multilingual MPNet and CNN for classifying offensive content in social posts written in codemixed Dravidian languages such as Kannada, Malayalam and Tamil. (Subramanian et al., 2022) employed transformer-based and conventional machine learning models to categorize the codemixed text into

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 49–55
March 22, 2024

offensive and non-offensive categories. Moreover, the authors utilized an adapter-based approach to fine-tune the pre-trained transformer models. (Vadakkekara Suresh et al., 2021) discusses a meta-learning approach for detecting offensive content in Dravidian language codemixed text.

(Chakravarthi et al., 2022c) introduced a codemixed dataset for sentiment analysis and offensive language identification in Dravidian languages. The dataset was prepared in codemixed Kannada, Malayalam and Tamil.

3 Task Description

We used the CodaLab platform to conduct the task [1]. The task aims to develop models to identify hate and offensive language content in Telugu-English codemixed social media comments. The hateful remarks on YouTube were gathered to create the dataset. Finding the videos where the hate comments could be found was the first challenge. When generating the dataset, consideration was given to comments containing both Telugu and English words written in their respective scripts and comments that wrote Telugu characters using Latin scripts. According to YouTube's rules [2], we annotated the comments into hate and non-hate categories. Slang was taken into account when annotating the Telugu remarks with additional care. Furthermore, an additional obstacle was presented by the existence of spam content, which was extraneous to the dataset due to its lack of contextual information. Those remarks were disregarded with respect to the intended dataset. The effective analysis and categorization of YouTube comments may present a challenge due to the prevalence of incorrect syntax, typographical errors, and non-standard language usage in social media posts and comments. Before annotating the text, these remarks were reviewed to ensure that the annotators understood the context properly. The annotators were native Telugu speakers with strong academic credentials and fluency in English. In conclusion, the dataset comprised 4,500 annotated comments, of which 4,000 were training data and 500 were test data. Some statistics about the dataset is given in Table 1.

The test data contained 250 hate and non-hate data, while the training dataset contained 1,939 hate and 2,061 non-hate comments. There is not a significant issue with class imbalance based on the distribution of data points in each class. Table 2 provides the train-test split of the dataset as well as the quantity of data points in each class.

Table 1: Statistics of the dataset

Statistics	Value
Total no.of words	43,432
No.of tokens	18,600
Maximum sentence length	71
Average sentence length	9.65

Table 2: Distribution of training and test datasets used for the shared task on abusive language detection in Telugu-English

Category	Train	Test
Hate	1,939	250
Non-hate	2,061	250
Total	4,000	250

Sixty-nine teams signed up for the competition. Only twenty-five teams, though, turned in their predictions for the test set. Each team was allowed to submit up to three runs, and the run with the best performance score was considered for creating the rank list, which is displayed in Table 3. The rank list was created, and the performance of the supplied findings was assessed using the macro F1-score.

4 System Description

This section discusses the models submitted to the shared task.

4.1 Sandalphon

This team used a fine-tuned Telugu-BERT model (Joshi, 2022) for implementation. The authors used a transliteration-based augmentation technique. A transliteration model was utilized for transliterating all the texts to the Telegu script, and another model to transliterate all the texts to Romanized script. This team scored the highest F1 score of 0.7711 in the shared task and shared first place.

4.2 Selam

This team shared first place with team Sandalphon. The submitted models were based on Convolutional Neural Network (CNN) and logistic regression for the classification.

[1] https://codalab.lisn.upsaclay.fr/competitions/16095
[2] http://tinyurl.com/ys56hrr5

Table 3: Rank list for the Telugu-English subtask

Team Name	macro F1	Rank
Sandalphon (Tabassum et al., 2024)	0.7711	1
Selam (Kanta et al., 2024)	0.7711	1
Kubapok	0.7431	3
DLRG1	0.7101	4
DLRG (Rajalakshmi et al., 2024)	0.7041	5
CUET_Binary_Hackers (Farsi et al., 2024)	0.7013	6
CUET_OpenNLP_HOLD	0.6878	7
Zavira (Ahani et al., 2024)	0.6819	8
IIITDWD-zk (Shaik et al., 2024)	0.6739	9
lemlem - Moein Tash	0.6708	10
Mizan	0.6616	11
byteSizedLLM	0.6609	12
pinealai	0.6575	13
IIITDWD_SVC (Sai et al., 2024)	0.6565	14
MUCS (KK et al., 2024)	0.6501	15
Lemlem-eyob	0.6498	16
Tewodros (Achamaleh et al., 2024)	0.6498	16
Fida (Ullah et al., 2024)	0.6369	18
Lidoma (Zamir et al., 2024)	0.6151	19
MasonTigers - Dhiman Goswami	0.5621	20
Habesha	0.5284	21
MasonTigers - AL Nahian Bin Emran	0.4959	22
CUET_DASH	0.4956	23
Fango	0.4921	24
Tayyab	0.4653	25

4.3 Kubapok

This team trained the transfer model for text classification. The model was trained with the following hyperparameters: warmup_ratio=0.1 and num_epochs=30. The team selected the best epoch checkpoint based on the F1 score computed over the development set to fix the model. The final score was fixed by taking the average of the five models' probabilities. The threshold was set at 0.5; class 0 was selected when the score fell below 0.5, and class 1 for data with a score greater than 0.5.

4.4 DLRG1

The team employed a Bi-LSTM (Bidirectional Long Short-Term Memory) to process sequential data by considering past and future contexts. They used stacked ensembles to combine predictions from multiple models to improve accuracy, leveraging the strength of diverse model architectures. A custom stacking model was employed by combining diverse classifiers, swiftly pinpointing hate speech with heightened accuracy, ensuring a safer and more inclusive online environment in Telugu-speaking communities.

4.5 DLRG

The team initially performed transliteration using the ai4bharat library's XlitEngine (Madhani et al., 2022) for Hate and Offensive Language Detection in Telugu Codemixed Text. The text was transliterated to enhance uniformity and facilitate subsequent processing. Following transliteration, two detection methods were implemented. Firstly, logistic regression with TF-IDF features was employed. Secondly, a single-cell Bi-GRU model was built. The model architecture included an embedding layer, two bidirectional GRU layers, and dense layers with ReLU activations. Training of the models included the hyperparameters such as binary cross-entropy loss and the Adam optimizer (Kingma and Ba, 2014). The combined approach leverages transliteration for text normalization and employs diverse models to capture linguistic nuances and sequential patterns in the Telugu Codemixed Text.

4.6 CUET_Binary_Hackers

The team used the pre-trained BERT large models such as MuRIL (Khanuja et al., 2021), indicBERT (Kakwani et al., 2020) and mBERT (Devlin et al., 2018) by fine-tuning the learning rates and batch size. Out of all the BERT models tried, the team's submission focuses on a fine-tuned indicBERT model, which gives better accuracy with a good F1 score.

4.7 CUET_OpenNLP_HOLD

This team used a transformer-based approach. The team fine-tuned XLM-R-base with the given training data.

4.8 Zavira

This team used a BI-LSTM network for classification.

4.9 IIITDWD-zk

The team utilized large language models such as Zephyr-7b-beta (Tunstall et al., 2023) and OpenChat-3.5 (Wang et al., 2023). In the second work, the team used an LSTM model to understand the context of the comments.

4.10 lemlem - Moein Tash

This team used Support Vector Machine (SVM), Simple Recurrent Neural Network (RNN) and Logistic Regression for the classification.

4.11 Mizan

This team used Simple Recurrent Neural Network (RNN) and Logistic Regression for the classification.

4.12 byteSizedLLM

They utilized embeddings generated from a subset of AI4Bharat's data (Kakwani et al., 2020), encompassing 100,000 randomized lines. These embeddings were created using our custom-built subword tokenizers for Telugu (with a size of 7.6 MB) and Tamil (with 1.3 MB) languages. The team employed a Bi-LSTM classifier to perform classification tasks.

4.13 pinealai

The team opted for fasttext (Bojanowski et al., 2017) and SVM for building the model. They applied GridSearch for the SVM model to know the best parameters for the model without overfitting the dataset. They also shuffled the dataset before any preprocessing to ensure that each observation was random.

4.14 IIITDWD_SVC

This team used the transliteration method to bring the text into the Telugu format ultimately and then used the translation model to translate the Telugu sentences into the English format and then trained with different models such as BERT model (cased), hate BERT and used that translated text and saved the model in h5 file and then used that model to predict the labels for the test dataset.

4.15 MUCS

The team used three models - Transfer learning with BERT model (Mathew et al., 2020), an ensemble of classifiers trained with Rchar and word-level TF-IDF features and a logistic regression classifier trained with word, subword and rchar concatenated features.

4.16 Tewodros

This team used Naive Bayes, Simple Recurrent Neural Network (RNN) and Logistic Regression for the classification.

4.17 Fida

The team used multimodels like BERT (Devlin et al., 2018), roBERTa (Liu et al., 2019) and DistilBERT (Sanh et al., 2019) for classification.

4.18 Lidoma

The team used BERT models (Devlin et al., 2018) for classification.

4.19 MasonTigers

They used XLM-R model (Conneau et al., 2019) for classification.

4.20 Habesha

The team used character-based RNN and distilBERT models.

4.21 CUET_DASH

The submission employs a multi-faceted approach using three distinct models for hate and offensive language detection in Telugu codemixed text. Logistic Regression was applied with feature extraction, incorporating n-grams and syntactic features for simplicity and interpretability. Telugu BERT enhances contextual understanding through fine-tuning on task-specific data, leveraging deep contextual embeddings..

4.22 Fango

They used Logistic regression with encoder-decoder method and SVM with encoder-decoder models were used.

4.23 Tayyab

They used BERT models for classification.

The majority of the teams used BERT-based models for feature extraction. Vairants of the BERT such as Telugu-BERT (Joshi, 2022), indicBERT (Kakwani et al., 2020), hate-BERT

(Mathew et al., 2020), and other multilingual BERT models achieved significant performance in classifiying a Telugu codemixed comment into hate and non-hate. Teams also used BERT classifier in addition to BERT embeddings for classification. There were submissions based on LSTMs and Bi-LSTMs. However, the performance of those models were not at par with the performance of the transformer models.

5 Conclusion

This paper discussed the findings of the shared task on Hate and Offensive Language Detection in Telugu Codemixed Text (HOLD-Telugu) conducted as part of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages (DravidianLangTech-2024) at EACL 2024. The datasets used for the competition were collected from YouTube comments and annotated with experts' help in compliance with YouTube's regulations. There were twenty-five submissions for this task. Most teams used multilingual BERT-based pre-trained models to transform the input text into the feature vector. The other submissions consisted of models using TF-IDF features and machine learning classifiers. We used macro F1-score to compute the classification performance and prepared the rank list accordingly.

Acknowledgements

This work was conducted with the financial support of the Science Foundation Ireland Centre for Research Training in Artificial Intelligence under Grant No. 18/CRT/6223, supported in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289_P2(Insight_2).

References

Tewodros Achamaleh, Lemlem Eyob Kawo, Ildar Batyrshin, and Grigori Sidorov. 2024. 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*, Malta. European Chapter of the Association for Computational Linguistics,.

Z Ahani, M. Shahiki Tash, M.T Zamir, I Gelbukh, and A Gelbukh. 2024. Zavira@DravidianLangTech 2024:Telugu hate speech detection using LSTM. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics,.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, 5:135–146.

Bharathi Raja Chakravarthi. 2023. Detection of homophobia and transphobia in Youtube comments. *International Journal of Data Science and Analytics*, pages 1–20.

Bharathi Raja Chakravarthi, Adeep Hande, Rahul Ponnusamy, Prasanna Kumar Kumaresan, and Ruba Priyadharshini. 2022a. How can we detect Homophobia and Transphobia? Experiments in a multilingual code-mixed setting for social media governance. *International Journal of Information Management Data Insights*, 2(2):100119.

Bharathi Raja Chakravarthi, Manoj Balaji Jagadeeshan, Vasanth Palanikumar, and Ruba Priyadharshini. 2023a. Offensive language identification in Dravidian languages using MPNet and CNN. *International Journal of Information Management Data Insights*, 3(1):100151.

Bharathi Raja Chakravarthi, Anand Kumar M, John P McCrae, B Premjith, KP Soman, and Thomas Mandl. 2020. Overview of the track on Hasoc-Offensive Language Identification-DravidianCodeMix. In *FIRE (Working notes)*, pages 112–120.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Shubanker Banerjee, Manoj Balaji Jagadeeshan, Prasanna Kumar Kumaresan, Rahul Ponnusamy, Sean Benhur, and John Philip McCrae. 2023b. Detecting abusive comments at a fine-grained level in a low-resource language. *Natural Language Processing Journal*, 3:100006.

Bharathi Raja Chakravarthi, Ruba Priyadharshini, Thenmozhi Durairaj, John McCrae, Paul Buitelaar, Prasanna Kumaresan, and Rahul Ponnusamy. 2022b. Overview of the Shared Task on Homophobia and Transphobia Detection in Social Media Comments. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 369–377, Dublin, Ireland. Association for Computational Linguistics.

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

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised Cross-lingual Representation Learning at Scale. *arXiv preprint arXiv:1911.02116*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

Salman Farsi, Asrarul Hoque Eusha, Jawad Hossain, Shawly Ahsan, Avishek Das, and Mohammed Moshiul Hoque. 2024. CUET_Binary_Hackers@DravidianLangTech 2024: 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*, Malta. European Chapter of the Association for Computational Linguistics,.

Raviraj Joshi. 2022. L3Cube-HindBERT and DevBERT: Pre-Trained Bert Transformer models for Devanagari based Hindi and Marathi Languages. *arXiv preprint arXiv:2211.11418*.

Divyanshu Kakwani, Anoop Kunchukuttan, Satish Golla, Gokul N.C., Avik Bhattacharyya, Mitesh M. Khapra, and Pratyush Kumar. 2020. IndicNLPSuite: Monolingual Corpora, Evaluation Benchmarks and Pre-trained Multilingual Language Models for Indian Languages. In *Findings of EMNLP*.

Selam Abitte Kanta, Grigori Sidorov, and Alexander Grigori. 2024. Selam@DravidianLangTech 2024:Identifying Hate Speech and Offensive Language. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics,.

Simran Khanuja, Diksha Bansal, Sarvesh Mehtani, Savya Khosla, Atreyee Dey, Balaji Gopalan, Dilip Kumar Margam, Pooja Aggarwal, Rajiv Teja Nagipogu, Shachi Dave, et al. 2021. MuRIL: Multilingual Representations for Indian Languages. *arXiv preprint arXiv:2103.10730*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Manavi KK, Sonali, Gauthamraj, Kavya G, Asha Hegde, and H L Shashirekha. 2024. 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*, Malta. European Chapter of the Association for Computational Linguistics,.

Prasanna Kumar Kumaresan, Premjith, Ratnasingam Sakuntharaj, Sajeetha Thavareesan, Subalalitha Navaneethakrishnan, Anand Kumar Madasamy, Bharathi Raja Chakravarthi, and John P McCrae. 2021. Findings of shared task on offensive language identification in Tamil and Malayalam. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 16–18.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yash Madhani, Sushane Parthan, Priyanka A. Bedekar, Ruchi Khapra, Vivek Seshadri, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Aksharantar: Towards building open transliteration tools for the next billion users. *ArXiv*, abs/2205.03018.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. *arXiv preprint arXiv:2012.10289*.

SN Prasanth, R Aswin Raj, P Adhithan, B Premjith, and Soman Kp. 2022. CEN-Tamil@ DravidianLangTech-ACL2022: Abusive comment detection in Tamil using TF-IDF and random kitchen sink algorithm. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 70–74.

B Premjith, KP Soman, and M Anand Kumar. 2018. A deep learning approach for Malayalam morphological analysis at character level. *Procedia computer science*, 132:47–54.

B Premjith, V Sowmya, Bharathi Raja Chakravarthi, Rajeswari Natarajan, K Nandhini, Abirami Murugappan, B Bharathi, M Kaushik, Prasanth Sn, et al. 2023. Findings of the shared task on multimodal abusive language detection and sentiment analysis in Tamil and Malayalam. In *Proceedings of the Third Workshop on Speech and Language Technologies for Dravidian Languages*, pages 72–79.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Subalalitha Chinnaudayar Navaneethakrishnan, Thenmozhi Durairaj, Malliga Subramanian, Kogilavani Shanmugavadivel, Siddhanth U Hegde, and Prasanna Kumar Kumaresan. 2022. Findings of the shared task on Abusive Comment Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages. Association for Computational Linguistics*.

Ratnavel Rajalakshmi, Saptharishee M, Hareesh Teja S, Gabriel Joshua R, and Varsini SR. 2024. DLRG@DravidianLangTech2024: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*, Malta. European Chapter of the Association for Computational Linguistics,.

Chava Srinivasa Sai, Rangoori Vinay Kumar, Sunil Saumya, and Shankar Biradar. 2024. 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*, Malta. European Chapter of the Association for Computational Linguistics,.

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*.

Zuhair Hasan Shaik, Sai Kartheek Reddy Kasu, Sunil Saumya, and Shankar Biradar. 2024. 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*, Malta. European Chapter of the Association for Computational Linguistics,.

Malliga Subramanian, Rahul Ponnusamy, Sean Benhur, Kogilavani Shanmugavadivel, Adhithiya Ganesan, Deepti Ravi, Gowtham Krishnan Shanmugasundaram, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2022. Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer. *Computer Speech & Language*, 76:101404.

Nafisa Tabassum, Mosabbir Hossain Khan, Shawly Ahsan, Jawad Hossain, and Mohammed Moshiul Hoque. 2024. Sandalphon@DravidianLangTech-EACL2024: Hate and Offensive Language Detection in Telugu Code-mixed Text using Transliteration-Augmentation. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics,.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, et al. 2023. Zephyr: Direct Distillation of LM Alignment. *arXiv preprint arXiv:2310.16944*.

Fida Ullah, Muhammad Tayyab Zamir, Muhammad Arif, M Ahmad, E Felipe-Riveron, and A Gelbukh. 2024. Fida@DravidianLangTech 2024: A Novel Approach to Hate Speech Detection Using Distilbert-base-multilingual-cased. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics,.

Gautham Vadakkekara Suresh, Bharathi Raja Chakravarthi, and John Philip McCrae. 2021. Meta-learning for offensive language detection in code-mixed texts. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, pages 58–66.

Guan Wang, Sijie Cheng, Xianyuan Zhan, Xiangang Li, Sen Song, and Yang Liu. 2023. Openchat: Advancing open-source language models with mixed-quality data. *arXiv preprint arXiv:2309.11235*.

M.T Zamir, M.S Tash, Z Ahani, A Gelbukh, and G Sidorov. 2024. Lidoma@DravidianLangTech 2024: Identifying Hate Speech in Telugu Code-Mixed Texts: A BERT Multilingual Approach. In *Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages*, Malta. European Chapter of the Association for Computational Linguistics,.