# MUCS@DravidianLangTech2023: Leveraging Learning Models to Identify Abusive Comments in Code-mixed Dravidian Languages

**Asha Hegde**[a], **Kavya G**[b], **Sharal Coelho**[c],
**Hosahalli Lakshmaiah Shashirekha**[d]
Department of Computer Science, Mangalore University, Mangalore, India
{[a]hegdekasha,[b]kavyamujk,[c]sharalmucs}@gmail.com
[e]hlsrekha@mangaloreuniversity.ac.in

## Abstract

Abusive language detection in user-generated online content has become a pressing concern due to its negative impact on users and challenges for policy makers. Online platforms are faced with the task of moderating abusive content to mitigate societal harm and foster inclusivity. Despite numerous methods developed for automated detection of abusive language, the problem continues to persist. This ongoing challenge necessitates further research and development to enhance the effectiveness of abusive content detection systems and implement proactive measures to create safer and more respectful online space. To address the automatic detection of abusive languages in social media platforms, this paper describes the models submitted by our team - MUCS, to the shared task "Abusive Comment Detection in Tamil and Telugu" at DravidianLangTech - in Recent Advances in Natural Language Processing (RANLP) 2023. This shared task addresses the abusive comment detection in code-mixed Tamil and Telugu texts, that includes the comments in both native script and romanized script, and romanized Tamil (RTamil) text. Two distinct models: i) AbusiveML - a model implemented utilizing Linear Support Vector Classifier (LinearSVC) algorithm fed with Term Frequency - Inverse Document Frequency (TF-IDF) of n-grams of words and character sequences within word boundary (char_wb) both in the range (1, 3) and ii) AbusiveTL - a Transfer Learning (TL) approach with three different Bidirectional Encoder Representations from Transformers (BERT) models, for three datasets along with random oversampling to deal with data imbalance, are submitted to the shared task for detecting abusive language in the given code-mixed texts. The AbusiveTL model fared well among these two models, with macro F1 scores of 0.46, 0.74, and 0.49 securing 1st, 1st, and 4th rank for code-mixed Tamil, Telugu, and RTamil texts respectively.

## 1 Introduction

Abusive language encompasses the use of words to insult, demean, or harm others, often through vulgar or profane language, and can include sexism, misogyny, and other forms of discrimination (Mandl et al., 2020; Subramanian et al., 2022; Chinnaudayar Navaneethakrishnan et al., 2023; Chakravarthi et al., 2023a,b). It may include words that provokes or aggravates an individual or a group of people. The phrase "abusive language" is also used synonymously with phrases like "offensive language" and "hate speech" (Hegde et al., 2021b). Over the past few years, the prevalence of offensive behavior targeting individuals, groups, or entire communities on social media platforms has significantly increased (Balouchzahi et al., 2021b; Hande et al., 2020; Chakravarthi et al., 2022a,b; Chakravarthi, 2023). This rise is creating negative impact such as, cyber-bullying, usage of offensive language, hate speech, and triggering content etc., on the well-being of online users. Hence, such negative content should be removed from the social media to keep online platforms healthy (Chakravarthi, 2022b; Kumaresan et al., 2022; Chakravarthi, 2022a).

Despite the efforts of social media companies to combat offensive/abusive language, the problem continues to escalate due to the limitations of existing algorithms used for detecting such content (Balouchzahi and Shashirekha, 2020; Bharathi and Agnusimmaculate Silvia, 2021; Bharathi and Varsha, 2022; Swaminathan et al., 2022a). These algorithms often fail to grasp the nuances of subjectivity and context that are crucial in accurately identifying abusive language. For instance, a single message might seem innocuous when taken out of context, but within a thread of previous conversations, it can reveal a pattern of abusive behavior. Similarly, certain phrases or words may have dif-



*Third Workshop on Speech and Language Technologies for Dravidian Languages*,
pages 266–274, Varna, Bulgaria, Sep 7, 2023.
https://doi.org/10.26615/978-954-452-085-4_039

ferent meanings depending on the context, making it challenging for algorithms to accurately assess their intent. This complexity poses difficulties for human reviewers who have to navigate through vast amounts of content. Hence, achieving an effective and comprehensive solution to detect abusive language on online platforms require advancements in Natural Language Processing (NLP), Machine Learning (ML) techniques and the ability to capture context in a more nuanced manner (Balouchzahi et al., 2021a). It is a complex problem that necessitates ongoing research, collaboration between experts,
mechanisms to create safer online environments.

One of the challenges in addressing abusive language on social media is the prevalence of code -mixed data where regional languages like Tamil, Kannada, Malayalam, etc., are combined with English at various levels such as sub-word, word, or sentence (Hegde and Shashirekha, 2022a). This linguistic diversity makes it difficult for abusive comment detection algorithms to accurately identify and categorize the offensive content. Further, the use of internet slangs, abbreviations, words in short forms, words from other languages, and emojis complicates the issue. Lack of annotated datasets specifically in low-resource languages like Tamil and Telugu pose an additional hurdle in developing effective abusive content detection algorithms for these languages (Ravikiran et al., 2022). Bridging this gap requires efforts to gather and annotate data in these languages to train models that can better understand and detect abusive content in diverse linguistic settings.

"Abusive Comment Detection in Tamil and Telugu-DravidianLangTech@RANLP - 2023[1]" shared task encourages researchers to develop models to identify whether the given code-mixed Tamil and Telugu texts and RTamil texts is abusive or not (Priyadharshini et al., 2023). Code-mixed Tamil comments are distributed into nine classes (None-of-the-above, Misandry, Counter-speech, Misogyny, Xenophobia, Hope-Speech, Homophobia, Transphobic, Not-Tamil), Telugu comments into two classes (Non-Hate, Hate), and RTamil comments into eight classes (None-of-the-above, Misandry, Counter-speech, Xenophobia, Hope-Speech, Misogyny, Homophobia, Transphobic), in the dataset provided by the shared task organizers.

To address the challenges of this shared task, in this paper, we - team MUCS, describe the two classification models: i) AbusiveML model utilizing LinearSVC fed with TF-IDF of n-grams of words and char_wb both in the range (1, 3) and ii) AbusiveTL - a TL model trained using three different versions of BERT (Tamil BERT, Telugu BERT, and Distilled Multilingual BERT (DistilmBERT)).

The rest of the paper is arranged as follows: a review of related work is included in Section 2 and the methodology is discussed in Section 3. Experiments and results are described in Section 4 followed by concluding the paper with future work in Section 5.

## 2 Related work

Abusive comments are statements that offend a person or a group of people. These comments are directed at people who belong to certain nationality, gender, caste, race, sexuality, etc. The objective of abusive content detection is to find abusive speech on social media platforms, such as hate speech, derogatory language, misogyny, and racism. The description of some works that are carried out to perform a similar task is given below:

S N et al. (2022) presented Support Vector Machine (SVM) classifier for abusive comment detection in code-mixed Tamil and RTamil texts. They used TF-IDF with char_wb features in the range (1, 5) along with Random Kitchen Sink (RKS) algorithm to create feature vectors to train SVM classifier. Their proposed model obtained macro F1 scores of 0.32 and 0.25 for code-mixed Tamil and RTamil texts respectively. Palanikumar et al. (2022) proposed ML models (Light Gradient-boosting Machine (LGBM), Categorical Boosting (Catboost), Random Forest (RF), SVM and Multinomial Naive Bayes (MNB)) on fine-grained abusive detection in Tamil. To increase the size of the minority class in the dataset, they transliterated the given code-mixed dataset and combined it with the dataset. ML models are trained with TF-IDF of char_wb n-grams and MURIL - a pretrained BERT model. The proposed ML model trained with MURIL outperformed other models with macro average F1 score of 0.290 and weighted F1 score of 0.590.

Swaminathan et al. (2022b) proposed the ML models (SVM, MultiLayer Perceptron (MLP), and k-Nearest Neighbours Classifier (k-NN)) to classify abusive content in RTamil code-mixed text. In their study, they combined language-agnostic sentence embeddings with the TF-IDF of word

[1]https://codalab.lisn.upsaclay.fr/competitions/11096

vectors to train SVM classifier and obtained an accuracy of 0.520 and macro F1 score 0.54. Nayel and Shashirekha (2019) described the ML models (SVM, Linear Classifier, MLP) for binary classification and, multi-class classification to detect the type of offensive content in three languages (English, German, and Hindi). For both binary and multi-class classification, SVM classifier trained with TF-IDF of word n-grams in the range (1, 2) exhibited macro F1 scores of 0.66, 0.75, 0.46 and 0.42, 0.47, 0.23 for English, Hindi and German languages respectively. Balouchzahi et al. (2021c) submitted two distinct models: COOLIEnsemble - an ensemble model of MLP, eXtreme Gradient Boosting (XGboost) and Logistic Regression (LR) trained using term frequencies and COOLI-Keras - a Deep Learning (DL) classifier, to identify offensive language in code-mixed Kannada-English, Malayalam-English, and RTamil texts. Out of the two models, COOLI-Ensemble model outperformed the other model with weighted F1 scores of 0.97, 0.75, and 0.69 for Malayalam-English, RTamil, and Kannada-English respectively.

Hegde and Shashirekha (2022b) proposed Dynamic Meta Embedding (DME) based Long Short Term Memory (LSTM) classifier to perform sentiment analysis and homophobia detection as Task A (Malayalam and Kannada) and Task B (Tamil, English, RTamil) respectively. Their proposed methodology exhibited macro F1 scores of 0.61, and 0.44 for Malayalam, and Kannada respectively in Task A and for Task B their models obtained macro F1 scores of 0.74 and 0.58 for English and RTamil languages respectively. Das et al. (2021) explored three learning models (XGboost, LGBM, mBERT) for abusive and threatening content detection in Urdu. They trained XGboost and LGBM classifiers using pre-trained Urdu laser embeddings. Further, they fine-tuned mBERT and dehatebert-mono-arabic pretrained models for abusive and threatening content detection in Urdu. Their fine-tuned mBert models outperformed the other models with macro F1 scores of 0.88 and 0.54 for abusive and threatening content detection respectively. Balouchzahi and Shashirekha (2020) proposed three distinct models to identify hate speech in English, German, and Hindi languages. They implemented i) ensemble of ML classifiers (RFC, LR, and SVM) trained with TF-IDF of word n-gram in the range (1, 2) and character n-grams in the range (1, 5), ii) TL based classifier using Universal Language Model Fine-tuning (ULMFiT) model, and iii) a hybrid model which is an ensemble of ML (i)) and TL (ii)) models. The ensembled ML classifier obtained macro F1 score of 0.5044 for German and hybrid model obtained macro F1 score of 0.5182 for Hindi.

From the above related work, it is found that among ML, DL, and TL models, TL models outperformed the other models indicating the efficiency of the TL models in detecting abusive content on social media. Though there are several models to identify abusive content in social media text, there is still scope for developing models for low-resource languages like Tamil and Telugu as these languages are not much explored in the realm of code-mixed content.

## 3 Methodology

The objective of this work is to identify abusive comments in code-mixed Tamil, Telugu, and RTamil texts. This is achieved by proposing two distinct models, AbusiveML and AbusiveTL. Detailed description of the models are given below:

### 3.1 Preprocessing

Preprocessing the raw text is an important initial step in text processing to enhance the performance of the learning models. During preprocessing, punctuation, numerical data, hashtags, user mention and stopwords are removed. English stopwords list available in Natural Language Tool Kit (NLTK) [2] and Tamil[3] and Telugu[4] stopwords lists available in github repository are used as reference to remove the stopwords.

### 3.2 Models Construction

The framework of AbusiveML and AbusiveTL are visualized in Figures 1 and 2. AbusiveML uses LinearSVC classifer and AbusiveTL uses transformer based classifier - ClassificationModel. Model descriptions are as follows:

#### 3.2.1 AbusiveML

n-grams are widely used in text processing projects due to their ease of implementation and scalability. By increasing the 'n' value up to a certain level, a model can capture larger contexts and store more

[2]https://pythonspot.com/nltk-stop-words/

[3]https://gist.github.com/arulrajnet/e82a5a331f78a5cc9b6d372df13a919c

[4]https://github.com/Xangis/extra/_stopwords/blob/master/telugu

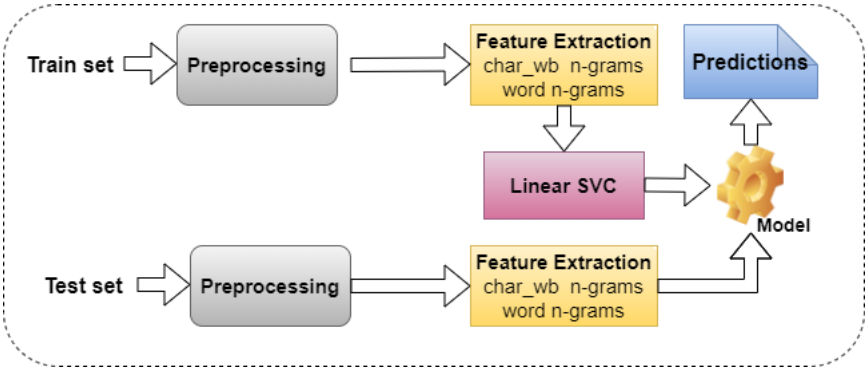



Figure 1: The framework of the AbusiveML model

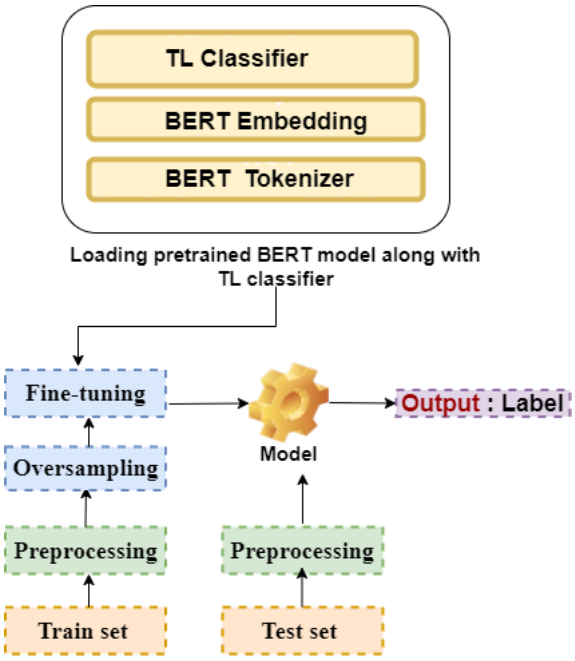



Figure 2: The framework of the AbusiveTL model

information about word/character sequences, enabling a better understanding of the relationships between words. This space-time trade-off is well-understood, allowing text processing experiments to scale up efficiently by adjusting the 'n' value based on the requirements of the task at hand viz. for simpler language tasks (e.g. autocomplete suggestions) smaller values of n (1, 2) is used, whereas, for complex tasks (e.g. text generation) larger values of n (3, 6) is used (Hegde and Shashirekha, 2021). However, larger value of 'n' introduces sparsity and increases the complexity of the learning algorithms.

n-grams of words and char_wb, both in the range (1, 3) extracted from the texts are vectorized using TfidfVectorizer[5] to train LinearSVC model. The

[5]https://scikit-learn.org/stable/modules/generated/sklearn.feature_extraction.text.TfidfVectorizer.html

| Hyperparameters | Values |
|---|---|
| penalty | l2 |
| C | 1.0 |
| class_weight | balanced |
| max_iter | max_iter |
| random_state | 100 |
| loss | squared_hinge |

Table 1: Hyperparameters and their values used in LinearSVC algorithm

hypermeters and their values used in LinearSVC model are shown in Table 1. The hyperparameters which are not mentioned in Table 1 are used with their default values.

In LinearSVC, setting the hyperparameter 'class_weight' to 'balanced' enables automatic adjustment of class weights based on their frequencies, effectively addressing data imbalance without the need for manual intervention.

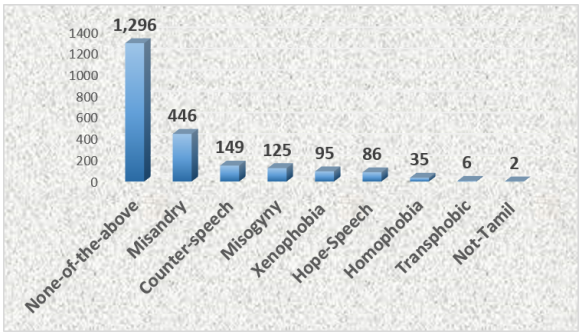
### 3.2.2 AbusiveTL

TL is an ML technique that leverages knowledge gained from one task to improve the performance of a related but different task (Hegde et al., 2021a). It involves using pretrained models as a starting point and fine-tuning them for a specific task or domain (Hegde et al., 2022). In the proposed AbusiveTL model, random oversampling[6] - an oversampling technique which increases the instances in the minority class by replicating the synthetic samples, is used before fine-tuning the pretrained models. DistilmBERT[7], Tamil BERT[8], and Telugu

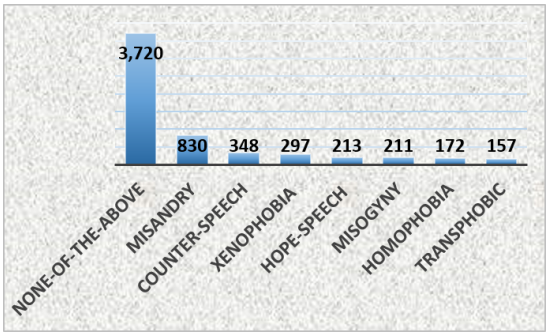
[6]https://imbalanced-learn.org/stable/over_sampling.html

[7]https://huggingface.co/distilbert-base-multilingual-cased

[8]https://huggingface.co/l3cube-pune/tamil-bert




(a) Tamil




(b) RTamil

Figure 3: Classwise distribution of code-mixed Tamil and RTamil datasets

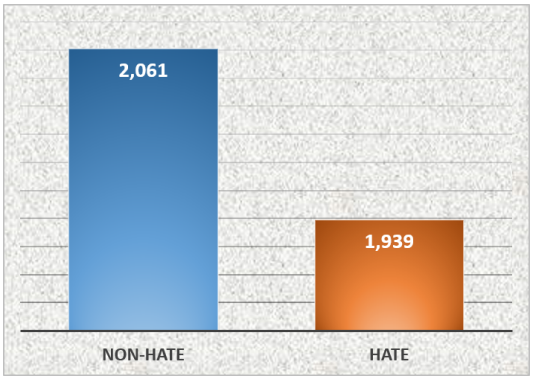



Figure 4: Classwise distribution of Telugu dataset

| Hyperparameters | Values |
|---|---|
| Layers | 6 |
| Dimension | 768 |
| Attention heads | 12 |
| Learning Rate | 2e-5 |
| Batch Size | 32 |
| Maximum Sequence Length | 128 |
| Dropout | 0.3 |

Table 2: Hyperparameters and their values used in AbusiveTL model

BERT[9] from huggingface library[10] are used to load the corresponding pretrained BERT versions which includes pretrained BERT tokenizer, embeddings and a TL classifier. Eventually this model is used to fine-tune the labeled Train set. Hyperparameters and their values used to implement AbusiveTL model are shown in Table 2. The hyperparameters which are not mentioned in Table 2 are used with their default values. The steps involved in fine-tuning the pretrained BERT models are given below:

- Tokenization - input text is passed through BERT's positional encoding-based tokenizer, which segments the text into individual tokens and adds positional information
- BERT encoder - tokens are transformed into contextualized embeddings using BERT encoder that helps to capture the contextual information and semantic representations
- Training - contextualized embeddings are fed into the model's classifiers viz. transformers based classifier for training

Prediction is carried out by the transformers based classifier (ClassificationModel).

## 4 Experiments and Results

Train, Development, and Test sets are provided by the shared task organizers (Priyadharshini et al., 2022) for abusive language detection in code-mixed Tamil and Telugu along with RTamil texts. Multiple experiments are carried out, incorporating different resampling techniques (Synthetic Minority Over-sampling TEchnique (SMOTE), random oversampling, and downsampling), feature combinations (pretrained word vectors, character count, and word count), and classifiers (LR, LinearSVC, NB, and MLP). The models that exhibited considerably good performances on the Development set were subsequently evaluated on the Test set. Figures 3 (a), 3 (b), and 4 show the label distribution in code-mixed Tamil, RTamil, and Telugu datasets respectively.

The predictions of the proposed models are evaluated by the organizers of the shared task based on macro F1 score and performance of the proposed models on Test and Development sets are shown in Table 3. As illustrated in Table 3 AbusiveTL

[9]https://huggingface.co/l3cube-pune/telugu-bert
[10]https://huggingface.co/docs/hub/models-libraries

model outperformed the other model with macro F1 scores of 0.74, 0.46, and 0.49 securing 1st, 1st, and 4th rank in the shared task for Telugu, Tamil, and RTamil, Test sets respectively.

In spite of using data imbalance handling mechanisms, for Tamil and RTamil texts the macro F1 scores are still less. This may be due to the overlapping feature distributions in the Train set. Further, adding class_weight='balanced' as a hyperparameter to a LinearSVC model can help to address the class imbalance by assigning higher weights to the minority class during training. While it generally helps to improve the performance, the macro F1 scores might decrease even after using this technique in certain scenarios, such as, data complexity, loss information, and features used. Further, as random oversampling technique increase the instances in minority classes by duplicating samples, this can lead the model to become overly focused on the minority class, potentially causing overfitting. This means the model might perform exceptionally well on the Train set but fail to generalize to new, unseen data. Table 4 shows the misclassifications for Telugu and Tamil comments along with their English translations, actual labels, predicted labels (obtained from AbusiveTL models for Tamil, Telugu, and RTamil Test sets) and remarks. From Table 4, it is clear that removing stopwords and digits may also lead to misclassification in addition to rare words and wrong annotation. This underscores the importance of a balanced preprocessing approach that carefully considers the impact of each step on the overall classification performance, as eliminating stopwords and digits might inadvertently remove context and information necessary for classification. Figures 5 (a), 5 (b), and 6 illustrate the comparison of macro F1 scores of all the participating teams for code-mixed Tamil and RTamil and Telugu texts respectively.

## 5 Conclusion

This paper describes the models submitted by our team - MUCS, to "Abusive Comment Detection in Tamil and Telugu" shared task at DravidianLangTech@RANLP 2023, to identify abusive content in code-mixed Tamil, Telugu, and RTamil texts. Two models: i) AbusiveML model that utilizes LinearSVC algorithm fed with TF-IDF of n-grams of words and char_wb both in the range (1, 3) and ii) AbusiveTL model fine-tuned on oversampled Train set with three different BERT models (for three different languages), are proposed to detect abusive comments in the input text. AbusiveTL models outperformed the other models with macro F1 scores of 0.74, 0.46, and 0.49 securing 1st, 1st, and 4th rank in the shared task for Telugu, Tamil, and RTamil texts respectively. Efficient resampling techniques for handling imbalanced data with effective feature extraction will be explored further.

## References


F Balouchzahi, S Bashang, G Sidorov, and HL Shashirekha. 2021a. CoMaTa OLI-Code-mixed Malayalam and Tamil Offensive Language Identification. In *Working Notes of FIRE 2021-Forum for Information Retrieval Evaluation (Online). CEUR*.

Fazlourrahman Balouchzahi, BK Aparna, and HL Shashirekha. 2021b. MUCS@ DravidianLangTech-EACL2021: COOLI-Code-Mixing Offensive Language Identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329.

Fazlourrahman Balouchzahi, Aparna B K, and H L Shashirekha. 2021c. MUCS@DravidianLangTech-EACL2021:COOLI-Code-Mixing Offensive Language Identification. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 323–329, Kyiv. Association for Computational Linguistics.

Fazlourrahman Balouchzahi and HL Shashirekha. 2020. LAs for HASOC-Learning Approaches for Hate Speech and Offensive Content Identification. In *FIRE (Working Notes)*, pages 145–151.

B Bharathi and A Agnusimmaculate Silvia. 2021. SSNCSE_NLP@DravidianLangTech-EACL2021: Offensive language identification on multilingual code mixing text. In *Proceedings of the First Workshop on Speech and Language Technologies for Dravidian Languages*, pages 313–318, Kyiv. Association for Computational Linguistics.

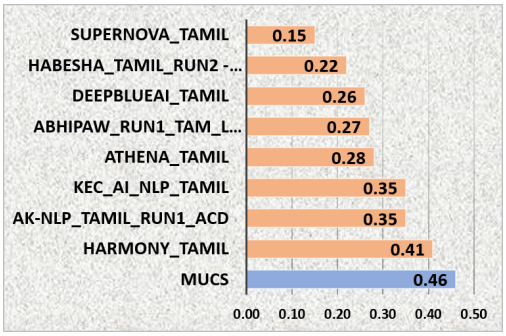
B Bharathi and Josephine Varsha. 2022. SSNCSE NLP@TamilNLP-ACL2022: Transformer based approach for detection of abusive comment for Tamil language. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 158–164, Dublin, Ireland. Association for Computational Linguistics.

Bharathi Raja Chakravarthi. 2022a. Hope speech detection in Youtube comments. *Social Network Analysis and Mining*, 12(1):75.

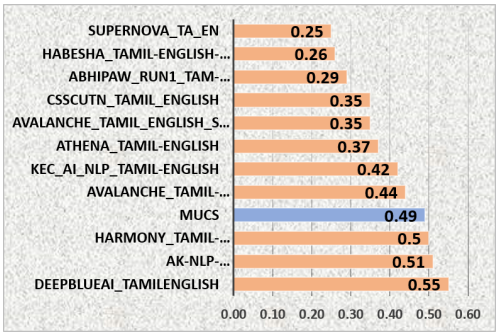
Bharathi Raja Chakravarthi. 2022b. Multilingual hope speech detection in English and Dravidian languages. *International Journal of Data Science and Analytics*, 14(4):389–406.

| Model | Language | Development set | | Test set | |
|---|---|---|---|---|---|
| | | With imbalanced data | With balanced data | With imbalanced data | With balanced data |
| AbusiveML | Telugu | 0.61 | 0.65 | 0.66 | **0.70** |
| | Tamil | 0.31 | 0.39 | 0.30 | 0.32 |
| | RTamil | 0.41 | 0.46 | 0.38 | 0.43 |
| AbusiveTL | Telugu | 0.73 | 0.77 | 0.65 | **0.74** |
| | Tamil | 0.08 | 0.39 | 0.07 | 0.46 |
| | RTamil | 0.15 | 0.51 | 0.16 | 0.49 |

Table 3: Performance of the proposed models with imbalanced and balanced datasets




(a) Tamil




(b) RTamil

Figure 5: Comparison of macro F1 scores of the participating teams in the shared task for code-mixed Tamil and RTamil datasets

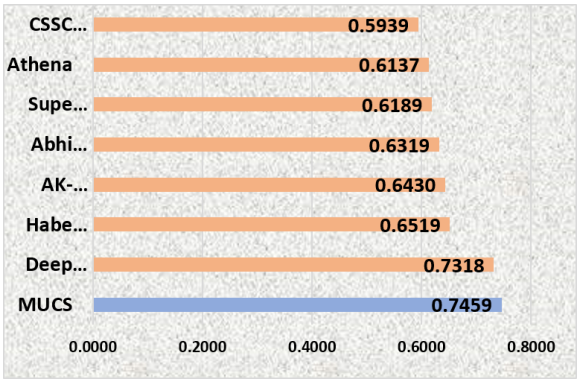



Figure 6: Comparison of macro F1 scores of the participating teams in the shared task for code-mixed Telugu dataset

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, 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.

Subalalitha Chinnaudayar Navaneethakrishnan, Bharathi Raja Chakravarthi, Kogilavani Shanmugavadivel, Malliga Subramanian, Prasanna Kumar Kumaresan, Bharathi, Lavanya Sambath Kumar, and Rahul Ponnusamy. 2023. Findings of shared task on sentiment analysis and homophobia detection

| Language | Comment | English Translations | Actual Label | Predicted Label | Remarks |
|---|---|---|---|---|---|
| Telugu | ఓటమి ఎంతో నేర్పుకోవ చ్చు దాని కిక్కే వేరు ఉంటుంది | Defeat can learn a lot, but it's different | non-hate | hate | The word "defeat" carries a negative connotation and may be due to this the comment is classified as 'hate'. |
| | అదే 420 పరిపాలన | Same 420 administration | hate | non-hate | '420' is a slang term that is often used in the negative tone and it is removed during preprocessing (usually numeric information will be removed). The remaining words has nothing to do with 'hate' class and hence the comment is classified as 'non-hate'. |
| Tamil | அந்த தமிழன் ஒரு சாதிவெறி | That Tamil is a caste fanatic | None | Xenophobia | The comment is incorrectly annotated as 'None' because the terms 'caste' and 'fanatic' indicates the class 'Xenophobia'. |
| | வாழ்த்துக் கள் h ராஜா ஜி | Congratulations h Raja G | None | Hope-Speech | In this comment, the characters 'h' and 'G' will be removed during preprocessing and because the word 'Raja' is a noun, it may not have a representation. Further, the term 'congratulations' is associated with 'Hope-speech' class in the train set and hence it is classified as hope. |

Table 4: Samples of misclassification for code-mixed Telugu and Tamil texts

of Youtube comments in code-mixed Dravidian languages. In *Proceedings of the 14th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '22, page 18–21, New York, NY, USA. Association for Computing Machinery.

Mithun Das, Somnath Banerjee, and Punyajoy Saha. 2021. Abusive and Threatening Language Detection in Urdu using Boosting based and Bert based Models: A Comparative Approach. In *arXiv preprint arXiv:2111.14830*.

Adeep Hande, Ruba Priyadharshini, and Bharathi Raja Chakravarthi. 2020. KanCMD: Kannada CodeMixed Dataset for Sentiment Analysis and Offensive Language Detection. In *Proceedings of the Third Workshop on Computational Modeling of People's Opinions, Personality, and Emotion's in Social Media*, pages 54–63.

Asha Hegde, Mudoor Devadas Anusha, Sharal Coelho, and Hosahalli Lakshmaiah Shashirekha. 2021a. MUM at ComMA@ICON: Multilingual Gender Biased and Communal Language Identification using Supervised Learning Approaches. In *Proceedings of the 18th International Conference on Natural Language Processing: Shared Task on Multilingual Gender Biased and Communal Language Identification*, pages 64–69, NIT Silchar. NLP Association of India (NLPAI).

Asha Hegde, Mudoor Devadas Anusha, and Hosahalli Lakshmaiah Shashirekha. 2021b. Ensemble Based Machine Learning Models for Hate Speech and Offensive Content Identification. In *Forum for Information Retrieval Evaluation (Working Notes)(FIRE), CEUR-WS. org*.

Asha Hegde, Sharal Coelho, Ahmad Elyas Dashti, and Hosahalli Shashirekha. 2022. MUCS@ Text-LT-EDI@ ACL 2022: Detecting Sign of Depression from Social Media Text using Supervised Learning Approach. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 312–316.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2021. Urdu Fake News Detection Using Ensemble of Machine Learning Models.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022a. Learning Models for Emotion Analysis and Threatening Language Detection in Urdu Tweets.

Asha Hegde and Hosahalli Lakshmaiah Shashirekha. 2022b. Leveraging Dynamic Meta Embedding for Sentiment Analysis and Detection of Homophobic/Transphobic Content in Code-mixed Dravidian Languages.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Elizabeth Sherly, Sangeetha Sivanesan, and Bharathi Raja Chakravarthi. 2022. Transformer based hope speech comment classification in code-mixed text. In *International Conference on Speech and Language Technologies for Low-resource Languages*, pages 120–137. Springer.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the

Hasoc Track at Fire 2020: Hate speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for information retrieval evaluation*, pages 29–32.

Hamada Nayel and H. Shashirekha. 2019. DEEP at HASOC2019 : A Machine Learning Framework for Hate Speech and Offensive Language Detection. pages 336–343.

Vasanth Palanikumar, Sean Benhur, Adeep Hande, and Bharathi Raja Chakravarthi. 2022. DE-ABUSE@TamilNLP-ACL 2022: Transliteration as Data Augmentation for Abuse Detection in Tamil. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 33–38, Dublin, Ireland. Association for Computational Linguistics.

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, Dublin, Ireland. Association for Computational Linguistics.

Ruba Priyadharshini, Bharathi Raja Chakravarthi, Malliga S, SUBALALITHA CN, Kogilavani S V, Premjith B, Abirami Murugappan, and Prasanna Kumar Kumaresan. 2023. 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.

Manikandan Ravikiran, Bharathi Raja Chakravarthi, Anand Kumar Madasamy, Sangeetha Sivanesan, Ratnavel Rajalakshmi, Sajeetha Thavareesan, Rahul Ponnusamy, and Shankar Mahadevan. 2022. Findings of the Shared Task on Offensive Span Identification from Code-Mixed Tamil-English Comments. In *arXiv preprint arXiv:2205.06118*.

Prasanth S N, R Aswin Raj, Adhithan P, Premjith B, 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, Dublin, Ireland. 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.

Krithika Swaminathan, Bharathi B, Gayathri G L, and Hrishik Sampath. 2022a. SSNCSE_NLP@LT-EDI-ACL2022: Homophobia/transphobia detection in multiple languages using SVM classifiers and BERT-based transformers. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 239–244, Dublin, Ireland. Association for Computational Linguistics.

Krithika Swaminathan, Divyasri K, Gayathri G L, Thenmozhi Durairaj, and Bharathi B. 2022b. PANDAS@Abusive Comment Detection in Tamil Code-Mixed Data Using Custom Embeddings with LaBSE. In *Proceedings of the Second Workshop on Speech and Language Technologies for Dravidian Languages*, pages 112–119, Dublin, Ireland. Association for Computational Linguistics.