

Offensive Text Detection in Code-Mixed Dravidian Languages towards Marginalized Groups and Women

Team Members:


Joshua Mahadevan – 106120047

Lokkamithran – 106120061

Mubeena – 106120071

Guide:

Dr. C. Oswald,
Assistant Professor,
CSE Department,
NIT Tiruchirappalli



Introduction

- The intersection of technology, language, and social issues in code mixed Dravidian languages has highlighted the surge in offensive content online, marginalizing groups.
- Such offensive contents could be seen in social media like Instagram, Facebook, Reddit and YouTube comments etc.
- The need for effective mechanisms to identify and mitigate such text is critical for creating a safer digital environment that promotes inclusivity and counters discrimination.
- Our problem statement deals with recognizing offensive content existing on social media towards women and marginalized groups, with our focus on code-mixed Dravidian text.

Literature Survey

SI. No	Title	Authors	Publication	Methodology	Demerits
1.	Ceasing hate with MoH: Hate Speech Detection in Hindi-English Code-Switched Language [1]	Arushi Sharma, Anubha Kabrab, Minni Jainc	Elsevier, January 2022	Employs a 'MoH' pipeline, including language identification, Roman to Devanagiri script transliteration and fine-tuned mBERT and MuRIL models.	This methodology does not work for Dravidian languages.
2.	Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments [2]	Hiren Madhua, Shrey Sataparab, Sandip Modhac, Thomas Mandid, Prasenjit Majumderb	Elsevier, April 2023	The pipeline combines SentBERT with an LSTM. An alternative KNN-based pipeline with SentBERT also achieves noteworthy results.	Contains contextual tags only for Hindi-English code-mixed data.

Literature Survey

SI. No	Title	Authors	Publication	Methodology	Demerits
3.	Minority Positive Sampling for Switching Points – an Anecdote for the Code-Mixing Language Modelling [3]	Arindam Chatterjee, Bodla Vineeth Guptha, Parul Chopra, Amitava Das	ACL Anthology, May 2020	Experiments with statistical language modelling reveal challenges at language-switching points. To address this, they introduce minority positive sampling to improve performances.	Minority Positive Sampling is not tested against Dravidian code-mixed data for better performances.
4.	Towards Offensive Language Identification for Dravidian languages [4]	Hiren Madhua, Shrey Sataparab, Sandip Modhac, Thomas Mandid, Prasenjit Majumderb	ACL Anthology, April 2021	The study employs zero-shot and few-shot learning using XLM-RoBERTa and mBERT. The models, including pre-trained, fine-tuned versions of XLM-RoBERTa, leverage transfer learning to achieve better performance.	It only focuses on binary classification of hate speech.

Research Gaps

- **Focus on Binary Classification:** Most existing research in offensive speech detection simplifies the problem by treating it as a binary classification task: offensive or not offensive .This approach fails to capture the nuances and complexities of offensive language, particularly when it comes to targeting specific marginalized groups within a community.
- **Neglect of Marginalized Groups:** By predominantly focusing on hate speech that targets a broad audience or generic hate, the current research overlooks the distinct forms of discrimination and abuse faced by marginalized groups such as different races, religions, sexual orientations, and disabilities.
- **Lack of Attention to Gender-Based Hate:** Another critical blind spot in existing research is the limited attention given to hate speech specifically targeted at women . Gender-based hate speech is a pervasive issue online and offline, yet it often receives inadequate consideration in offensive speech detection studies.

Problem Statement and Objectives

Problem statement: Annotating dataset according to the defined classes, detecting offensive text present in Code-Mixed Tamil Language by initially performing Binary Classification which predicts if the comment is offensive or not, followed by Multi-Label Classification which classifies into one of the 7 categories if detected as offensive and analysing the performance.

Objectives:

- Develop an annotated dataset with respect to the labels we define for targeted groups and women and host it in public domain.
- Create an binary classification Deep Learning model to predict if a statement is offensive.
- Create a multi-label classification Machine Learning model to classify which class the statement belongs to.
- Evaluate and benchmark models performance.

High level block diagram of Proposed Methodology

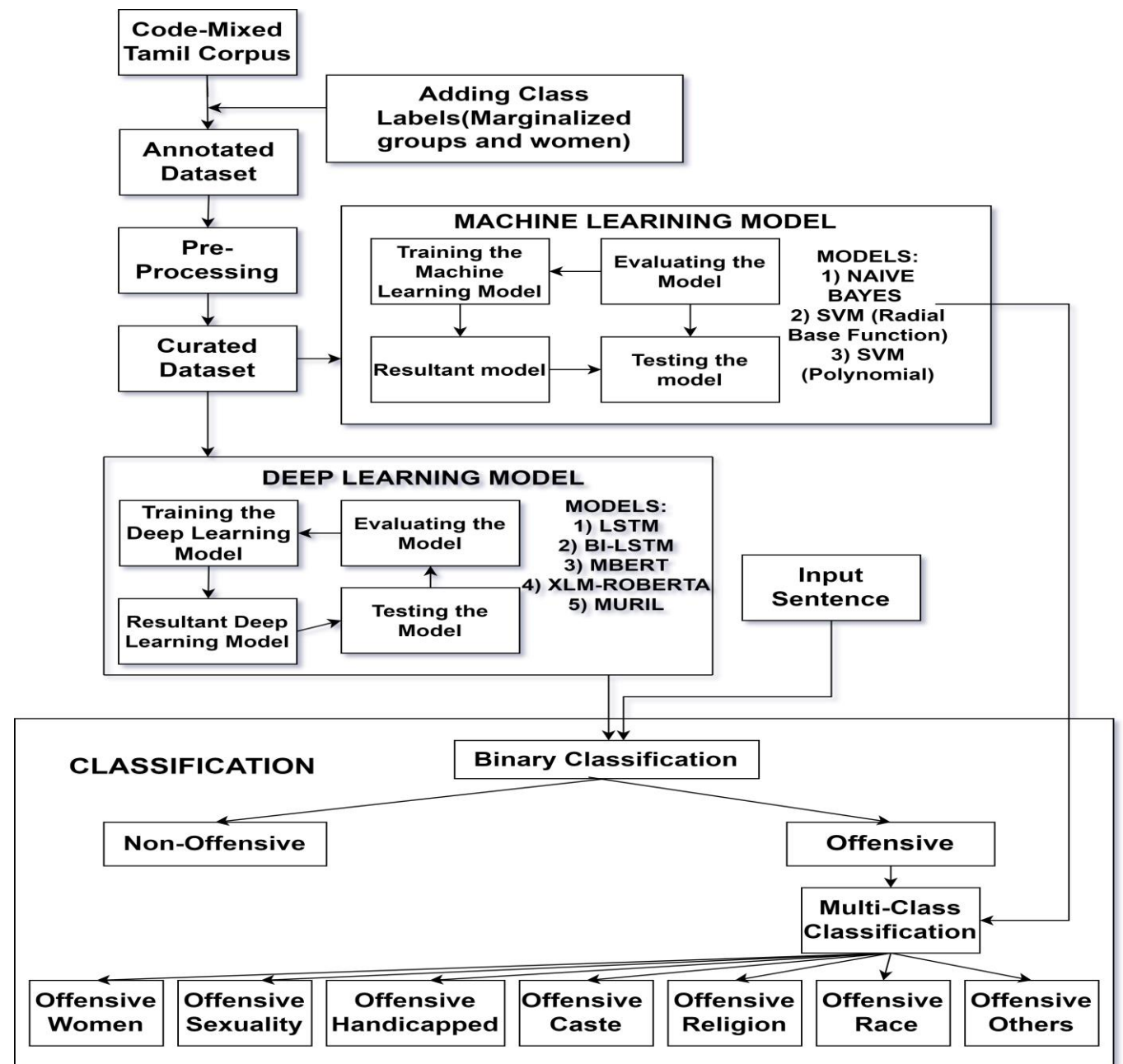


Fig 1. Block diagram of proposed architecture

Novelty

- Contrary to the existed approaches, the proposed methodology aims towards classifying Code-Mixed text into a range of Offensive Categories instead of a Binary classes.
- For Codemix text in Dravidian languages this is the first attempt to detect offense towards Women and Marginalized groups. The model detects the depth of offense through recognizing the category of people it is targeted towards.
- To overcome the limitations in data, the model was built as a combination of Deep Learning and Machine Learning model making it predict better with much lesser computational resources.

**Workflow
diagrams:
Sample Input
Text Classification**

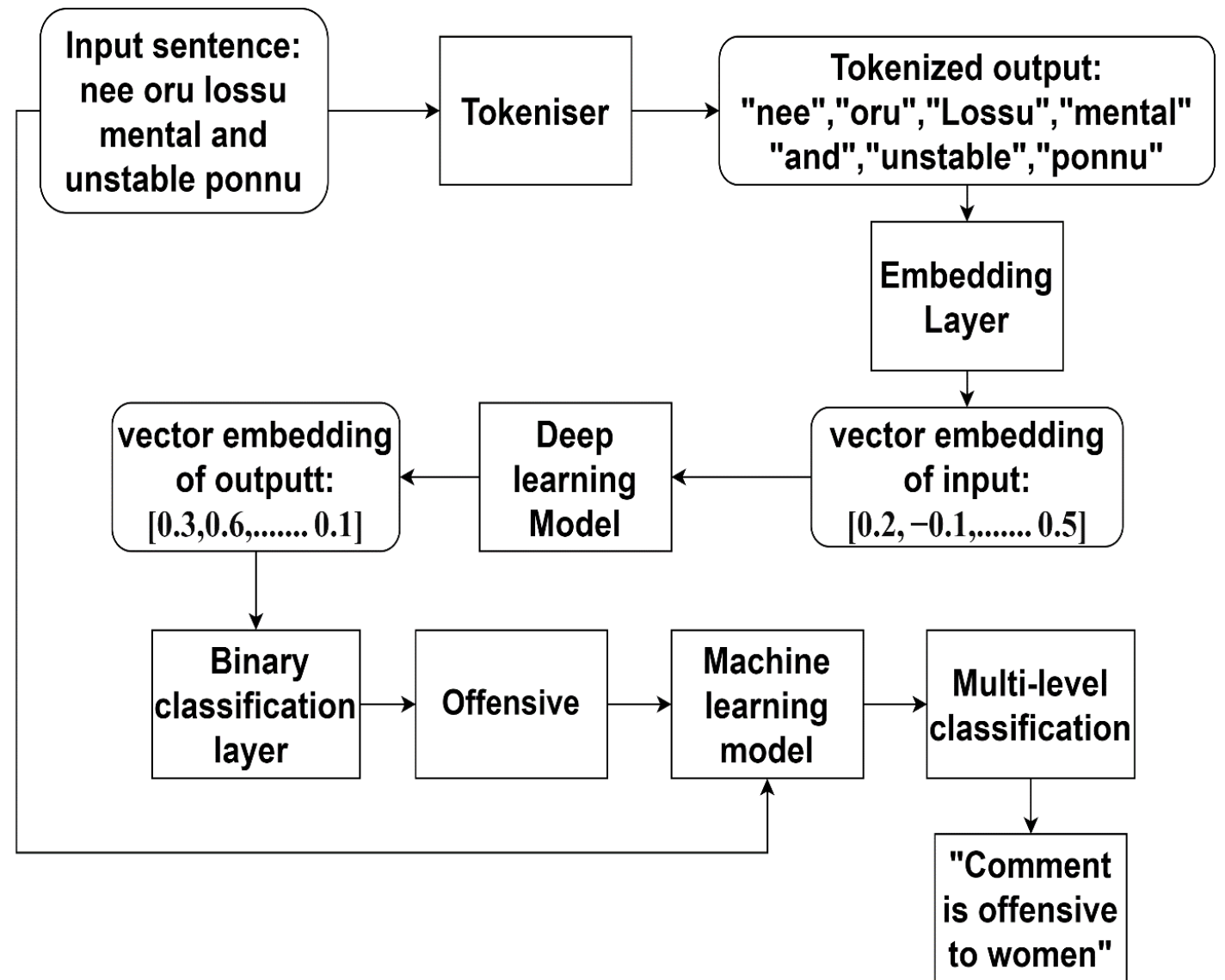


Fig 2. Sample input text classification for Deep Learning and Machine Learning models

Algorithms

- Deep Learning Model
 - Importing libraries and dataset.
 - Pre-Processing and Encoding (Tokenization).
 - Train-test split.
 - Model building (import pre-trained model / from scratch).
 - Training the model.
 - Testing / Calculation of evaluation metrics.
- Machine Learning Model
 - Importing libraries and dataset.
 - Pre-Processing and splitting dependent and independent variables.
 - Encoding and Vectorization of variables.
 - Up-sampling data using SMOTE.
 - Fitting the model on training data.
 - Testing / Calculation of evaluation metrics.

Implementation/Simulation Environment

- The dataset was annotated manually from a base dataset [5].
- The following models were implemented using machine learning and deep learning techniques using python in the Google Collaboratory environment:
 1. LSTM+ Machine Learning (SVM and Naïve Bayes)
 2. Bi-LSTM + Machine Learning (SVM and Naïve Bayes)
 3. MuRIL +Machine Learning (SVM and Naïve Bayes)
 4. XLM-RoBERTa +Machine Learning (SVM and Naïve Bayes)
 5. mBERT +Machine Learning (SVM and Naïve Bayes)

Performance of Deep Learning models

Table 1. Comparison of evaluation metrics for various DL models

DL model / Metric	Accuracy	Precision	Recall	F1 Score
LSTM	0.608	0.669	0.71	0.689
Bi-LSTM	0.604	0.627	0.697	0.66
MuRIL	0.587*	0.615*	0.932*	0.74*
mBERT	0.586*	0.605*	0.954*	0.739*
XLM-RoBERTa	0.588*	0.606*	0.955*	0.74*

Values denoted with (*) have been obtained through k-fold cross validation with k=5, others have been tested on 20% of the dataset

Performance of Machine Learning models

Table 2. Comparison of evaluation metrics for various ML models

ML model / Metric	Accuracy	Precision	Recall	F1 Score
Naïve-Bayes	0.89	0.895	0.912	0.895
Support Vector Machine (RBF)	0.861*	0.808*	0.86*	0.814*
Support Vector Machine (Polynomial)	0.85	0.92	0.85	0.81

Values denoted with (*) have been obtained through k-fold cross validation with k=5, others have been tested on 20% of the dataset

Performance of DL and ML models (combined)

Table 3. Comparison of accuracy for different combinations of DL and ML models (tested on 20% of the dataset)

DL model / ML model	Naive Bayes	SVM(RBF)	SVM(Poly)
LSTM	0.541	0.523	0.516
Bi-LSTM	0.537	0.520	0.513
XLM-RoBERTa	0.522	0.505	0.499
mBERT	0.522	0.505	0.498
MURIL	0.522	0.505	0.498

Performance of the system under several use cases

Sentences / Model Combinations	LSTM + SVM(Poly) / SVM(RBF) / Naïve bayes	Bi-LSTM + SVM(Poly) / SVM(RBF) / Naïve bayes	MURIL + SVM(Poly) / SVM(RBF) / Naïve bayes	XLM-Roberta + SVM(Poly) / SVM(RBF) / Naïve bayes	MBERT + SVM(Poly) / SVM(RBF) / Naïve bayes
nee laam uyir vaala ve kudathu	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Others / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Other / Offensive Sexuality
indha ponnu waste da	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Women	Offensive Caste / Offensive Women / Offensive Sexuality
matha naatu kaarangala India kulla vidave kudathu	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality	Offensive Caste / Offensive Others / Offensive Sexuality

Comparison with existing solutions

- We compared our results with the base dataset's research paper [5] and arrived at the following conclusions:
 - The paper [5] classifies input sequence as one of four classes: Positive, Negative, Neutral and Mixed_feelings. For this classification task, the Random Forest Classifier provided the highest performance with a Precision of 0.63, Recall of 0.62 and a F1 Score of 0.56.
 - Our best performing Deep Learning models classify offensive text with a Precision of 0.669, Recall of 0.955 and a F1 Score of 0.74
 - Our best performing Machine Learning models classify offensive text with a Precision of 0.92, Recall of 0.91 and a F1 Score of 0.89
- Our models perform better at multi-label classification because we augmented the base dataset with annotation to suit our task.

References

1. Sharma, Arushi, Anubha Kabra, and Minni Jain. "Ceasing hate with moh: Hate speech detection in hindi–english code-switched language." *Information Processing & Management* 59, no. 1 (2022): 102760.
2. Madhu, Hiren, Shrey Satapara, Sandip Modha, Thomas Mandl, and Prasenjit Majumder. "Detecting offensive speech in conversational code-mixed dialogue on social media: A contextual dataset and benchmark experiments." *Expert Systems with Applications* 215 (2023): 119342.
3. Chatterjere, Arindam, Vineeth Guptha, Parul Chopra, and Amitava Das. "Minority positive sampling for switching points-an anecdote for the code-mixing language modeling." In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 6228-6236. 2020
4. Sai, Siva, and Yashvardhan Sharma. "Towards offensive language identification for Dravidian languages." In *Proceedings of the first workshop on speech and language technologies for Dravidian languages*, pp. 18-27. 2021.
5. Chakravarthi, B.R., Priyadharshini, R., Muralidaran, V. et al. DravidianCodeMix: sentiment analysis and offensive language identification dataset for Dravidian languages in code-mixed text. *Lang Resources & Evaluation* 56, 765–806 (2022). <https://doi.org/10.1007/s10579-022-09583-7>