Hateful Speech Detection in Code-Mixed Tweets: Conversational Hindi and Individual Telugu tweets

Presented by :
Ramesh Kumar (21dcs15)
Polagani Manoj (21dcs016)

Supervised by :
Dr. Mohit Kumar

DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING
NATIONAL INSTITUTE OF TECHNOLOGY
HAMIRPUR HAMIRPUR (H.P.)-177005

TABLE OF CONTENT

Topic

Hateful Speech Detection in Code-Mixed Tweets: Conversational Hindi and Individual Telugu tweets

INTRODUCTION

- **The problem of hate speech and offensive language is widespread on social media.**
- **Conversations often mix two or more languages (e.g., Hindi-English, Telugu-English), making it difficult to detect offensive speech.**
- **Code-mixing complicates natural language processing tasks due to the blending of multiple languages in a single sentence.**
- **Offensive content in social media conversations is often context-dependent, further increasing the complexity.**

Motivation

- ❑ **Reducing Online Hate speech**
- ❑ **Promoting Positive Online interaction**
- ❑ **Handling Multilingual Challenges**
- ❑ **Addressing sensitive domains**

Problem statement

- **Offensive speech detection is important for curbing cyberbullying, promoting safe online spaces, and improving moderation on platforms.**
- **Traditional models just works on individual tweets and considers just word embeddings to label Tweet.**
- **In all thread conversations the context of parent tweet is also required to classify current tweet.**
- **Traditional models are often not adapted for code-mixed languages, especially Telugu-English, requiring novel approaches.**


· May 18
Modi Ji COVID situation ko solve karne ke liye ideas maang rahe the
Mera idea hai resignation dedo please...
2.2K
9.7K
61.3K
· May 21
Doctors aur Scientists se manga hai
Chuti¥o se nahi.
Baith niche.
Hate and profanity towards author of source tweet
168
752
2.7K
Replying to @ and @
You totally nailed it, can't stop laughing
2:56 PM · May 21, 2021 · Twitter for Android
1 Retweet
27 Likes
This tweet is a reply to the above comment and it expresses a positive sentiment. But the reply is actually supporting the hate expressed in the comment towards the author of source tweet.

Datasets

Table 4.1: Original Dataset Statistics

Data	Total Conversations	HOF	NOT	Parent Tweets	Avg. Comments
Train	5740	2841	2899	82	46
Test	1348	695	653	16	53

Note: Val refers to Validation data.

Table 4.2: Train-Validation-Test Distribution

Data	Total Conversations	HOF (Hateful/Offensive)	NOT (Neither Hateful nor Offensive)
Train	4592	2273	2319
Val	1148	568	580
Test	1348	695	653

Note: Val refers to Validation data.

Data Distribution for Task1

Table 4.3: Number of Samples

Type	Number
Training Samples	4200
Testing Samples	1000

Table 4.4: Training Dataset Label Distribution

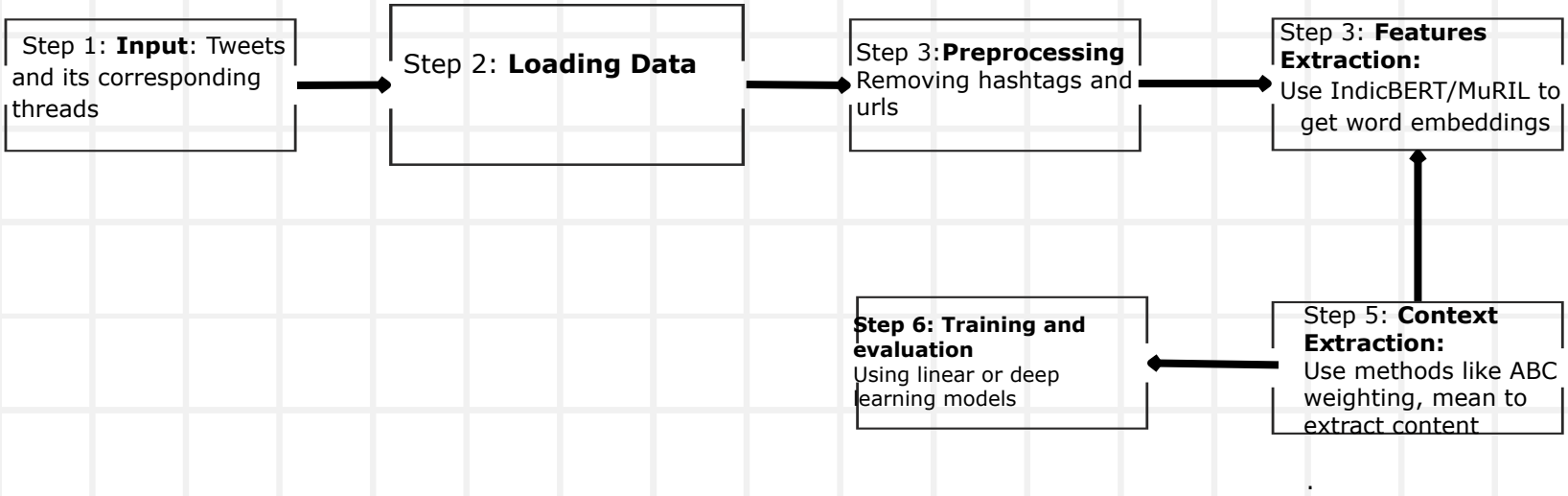
Label	Count
non-hate	2061
hate	1939

Table 4.5: Testing Dataset Label Distribution

Label	Count
non-hate	250
hate	250

Data Distribution for Task2

FLOWCHART


Step 1: **Input**: Tweets and its corresponding threads
Step 2: **Loading Data**
Step 3:**Preprocessing** Removing hashtags and urls
Step 3: **Features Extraction:** Use IndicBERT/MuRIL to get word embeddings
Step 5: **Context Extraction:** Use methods like ABC weighting, mean to extract content
Step 6: Training and evaluation Using linear or deep learning models

.

Embeddings

Following models are used for generating the embeddings from the data

- **BERT**
- **mBERT**
- **IndicBERT**
- **MurilBERT**
- **mDistillBERT**
- **Xmlr**
- **LsBSE**

Context Representation

In this project, several techniques for context representation were employed to effectively process tweets, comments, and replies, which exhibit a hierarchical structure.

1. Concatenation: Combines text at different levels into a single sequence and vectorizes it for consistent input processing.

2. Mean: Averages embeddings at each level to create a document representation, capturing contextual relationships.

3. Sequence: Tokenizes text and concatenates word embeddings into a sequence matrix, standardized with post-padding for neural network input.okeni

4. ABCWeighting: Applies weights to tweet, comment, and reply levels to create a weighted combination of embeddings, adjusting their contribution to the classification task.

Classification Models

1. Support Vector Machines: Finds the best boundary to separate classes, handling both linear and non-linear data.

2. K-Nearest Neighbors: Classifies data based on the majority class of its nearest neighbors.

3. Random Forest: Combines multiple decision trees to improve accuracy and reduce overfitting.

4. Naive Bayes: Applies probability theory to classify data based on the likelihood of features given a class.

5. LSTM (Long Short-Term Memory): A neural network specialized for handling sequential data, capturing long-term dependencies.

6. RNN (Recurrent Neural Network): Processes sequential data by maintaining a hidden state that captures information about previous inputs.

Results

Classifier	Representation	Fusion	F1 Train	F1 Test
KNN	SENTBERT	ABC Weighting	0.67	0.59
SVM	SENTBERT	ABC	0.659	0.62
KNN	Fine-tuned SENTBERT (SoftMax)	ABC	0.75	0.77
KNN	Fine-tuned SENTBERT (Online Contrastive Loss)	ABC Weighting	0.658	0.51
KNN	Fine-tuned Murril	ABC Weighting	0.795	0.792
LSTM	Fine-tuned Murril	Sequence	0.99	0.83
NB	mDistillBERT	Sequence	0.685	0.649

Performance of different models on task 1

Model	Precision	Recall	F1-Score	Accuracy
Xml R	0.769	0.769	0.769	0.827
IndicBERT	0.774	0.773	0.772	0.830
Mbert	0.757	0.755	0.755	0.817
Mdistill Bert	0.765	0.765	0.765	0.824
MuRIL + RNN	0.724	0.709	0.715	0.729
MuRIL + SVM	0.718	0.715	0.715	0.709
MuRIL + KNN	0.710	0.710	0.710	0.715
LaBSE	0.761	0.750	0.750	0.780
RemBERT	0.630	0.620	0.620	0.670

Performance of different models on task 2


Hateful Speech Detection
Enter Tweet Link:
Enter tweet URL here...
Select Language:
Hinglish
Predict
Prediction: Hateful

REFERENCES

- https://www.sciencedirect.com/science/article/abs/pii/S095741742202360 0?fr=RR-9&ref=pdf_download&rr=8cf756d8bfbdb2dd
- https://www.sciencedirect.com/science/article/abs/pii/S0957417422023600
- https://aclanthology.org/S19-2081/
- https://www.nature.com/articles/s41598-024-76632-2
- https://www.sciencedirect.com/science/article/pii/S2772941924000413
- https://link.springer.com/article/10.1007/s13278-024-01264-3
- https://aclanthology.org/W18-4411
- https://dl.acm.org/doi/full/10.1145/3625680
- https://scholar.google.com/scholar?hl=en&as_sdt=0%2C5&as_ylo=2024&q=conversational+tweets+hateful+speech+detection+&btnG=#d=gs_qabs&t=1733803386810&u=%23p%3DHIIA6bjmMMsJ

THANK YOU !