Hope Speech Detection: Identifying Positive Actors in Toxic Discourses

Mohammad Aflah Khan
aflah20082@iiitd.ac.in

Neemesh Yadav
neemesh20529@iiitd.ac.in

Raghav Sahni
raghav20533@iiitd.ac.in

Diksha Sethi
diksha20056@iiitd.ac.in

1. Motivation

Hate Speech on Social Media Platforms has been studied extensively in the past decade. However, all these studies focus on analyzing negativity in English Language for the most part but the problem is not just restricted to harmful content. Through our project we aim to detect Hope Speech which reinforces positivity in online discourse. This problem is relevant as alongside penalizing bad actors we can reward good actors if we can identify them. Some downstream application of our project include using the classifier to curate more data which can be used to train generative models. These models can then be deployed in toxic-online settings to spread positivity.

Overall, we propose building explainable classifiers for Hope Speech Detection which can classify text into 3 classes “Hope Speech”, “Non-Hope Speech” and “Non-English” for English language. Then we plan to use train the same models on Malayalam and Tamil to analyze the results and the shortcomings of directly using models in multi-lingual settings. We also plan to utilize Machine Learning models for this task as that is a relatively unexplored area. The few works on the subject discussed in the next section primarily focus on Deep Learning based methods and show minimal use of ML Techniques.

2. Related Work

Hope Speech Detection is a fairly new and novel task. The first mentions of Hope Speech Detection are found in the paper **HopeEDI: A Multilingual Hope Speech Detection Dataset for Equality, Diversity, and Inclusion** by Bharathi Raja Chakravarthi which defined Hope Speech and curated the first such dataset called **Hope Speech dataset for Equality, Diversity and Inclusion (HopeEDI)**. The dataset contains user-generated comments from YouTube with 28,451, 20,198 and 10,705 comments in English, Tamil and Malayalam. Originally the dataset has 3 classes “Hope Speech”, “Non Hope Speech” and “Non-X” (Non-English/Tamil/Malayalam). It is the first research of its kind to annotate hope speech for equality, diversity and inclusion in a multilingual setting. This task was formulated as the First Workshop on Language Technology For Equality, Diversity, Inclusion (LT-EDI-2021) in EACL 2021 and the papers submitted to the conference mainly focused on building new models and fine tune old models for this task. Some papers which are comparable to State-of-the-Art are **Hope Speech Detection in English YouTube Comments using Deep Learning Techniques** and **Hope Speech Detection for Equality, Diversity, and Inclusion** which use architectures like Bi-LSTMs, GRUs, CNNs, etc. to solve the problem and investigate several other architectures.

3. Timeline

3.1. Week 1-2: Dataset analysis and visualization and preprocessing the data.

3.2. Week 3-8: Training and testing traditional ML models (Logistic Regression, K Means, Decision Trees, SVM, Boosted Classification Trees etc.) alongside suitable Word Vectorization Techniques (TF-IDF, Word2Vec etc.)

3.3. Week 9: Exploring Deep Learning Methods (RNNs, LSTMs, Transformers etc.)

3.4. Week 10: Using model explainability techniques to evaluate models

3.5. Week 11: Combining the results obtained from various algorithms

4. Individual Tasks

All four members will contribute equally to data preprocessing, model training, analysis and the report. For a smoother and efficient work process, we plan to work in pairs. Each pair will take up one of the techniques and contribute to it. Though, we plan to spend one week on each method, the timeline for each might vary.

5. Final Outcome

To build several models and select one that can distinguish Hope Speech from other text confidently.

Further, to test the model on languages other than English (Tamil and Malayalam). We also plan on comparing our ML and DL models to the State-of-the-Art Models and see how it fares against them.