Hate Speech Classification

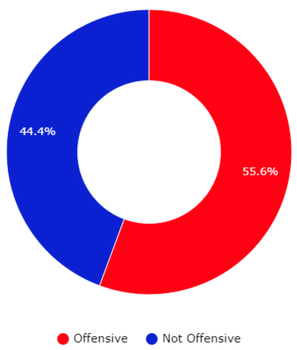
GROUP 48 (Aryan Dhull, Deepanshu, Pranav Aggarwal, Prerak Gupta)

Introduction

Problem Statement: The explosion of social media has made it difficult to monitor hate speech, which targets individuals or groups based on race, religion, etc. Traditional methods rely on manual review by humans, a slow and resource-intensive process. Developing an accurate NLP-based system is important to **flag hateful content** before it spreads, while still allowing free speech.

Motivation: Rise of hate speech online is discouraging, but the fight against it is gaining momentum with Natural Language Processing. Existing models incorporated in Twitter and Facebook **proactively remove harmful content**, even when it's cleverly disguised. The advancements in NLP motivate further development in hate speech detection. .

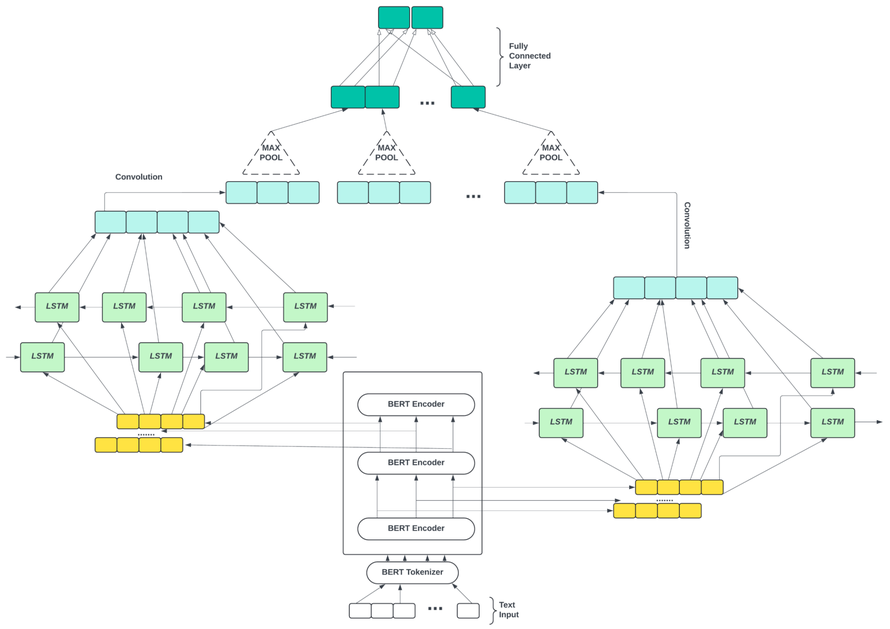
Dataset Description and Preprocessing



Dataset: *Social Bias Frames*
The social bias frames dataset is a valuable collection of posts gathered from various social media platforms, including Twitter, Reddit, and Stormfront. The dataset contains 147,139 entries, each characterized by 19 features.

The project focused on extracting "post" text content and "offensiveYN" labels indicating offensive nature (0 for not offensive, 1 for offensive). Ambiguity was resolved by converting "maybe offensive" (0.5) labels to "offensive" (1). Multiple annotations for posts were addressed through a voting mechanism to assign a representative label. The final dataset comprised 44,756 unique entries after preprocessing, with training, validation, and testing sets containing 35,424, 4,666, and 4,666 entries respectively, ensuring robust model training and evaluation.

Methodology



First, we use a pretrained BERT model for contextualized token embeddings. The output of each hidden layer in the BERT model is passed through a BiLSTM layer, followed by a convolution layer including activation and pooling layers. Finally, these pooled feature maps are passed through fully connected layers for classification. While BiLSTM helps the model capture information in both forward and backward directions, the CNN layer does further finetuning by accessing the local patterns.

Results

Classifier	Accuracy	F1-Score
Random Forest	0.62	0.40
ADABoost	0.60	0.48
SVC	0.61	0.43

These models used TF-IDF (Term Frequency-inverse Document Frequency) vectorizer. The various classifiers, we used gave nearly same results with the accuracy around 60% and F1 score around 0.4.

Model	Accuracy	F1-Score
Model 1	0.69	0.67
Model 2	0.66	0.66
Model 3	**0.73**	**0.72**

Model 3 (Stacked BiLSTM with Convolutional Layers) gave best results with F1 score of 0.72 and accuracy of 73%. Model 1 (BERT Classifier) and Model 2 (Multi-layered CNN) gave nearly same results.

Findings

- Among the TF-IDF based models, including Random Forest, ADABoost, and SVC, demonstrate moderate performance in terms of accuracy and F1 score, with ADABoost achieving the highest F1 score.
- The BERT-based models consistently outperform the TF-IDF models, highlighting the benefits of using contextualized word embeddings for text classification tasks.
- Among the BERT based models, the BiLSTM CNN architecture performs best, indicating its effectiveness in capturing sequential dependencies in the data as well as the ability of CNNs to capture local patterns.
- Further, the usage of embeddings after each hidden layer helps leverage the importance of semantic information extracted at each layer and hierarchical learning.

Future Work

Future work in the direction of this study that can be explored are:

- Model Optimization : Further optimization of hyperparameters could potentially improve the performance of the models using techniques such as Grid Search or Bayesian optimization.
- Transfer Learning : Transfer Learning approaches can be explored, such as fine-tuning a model trained for sentiment classification on hate speech specific data.
- Domain Adaptation : Adversarial training or multitask learning can help the model to perform well even on data from different domains or sources.