Hate Intensity Prediction (HIP)

Team Members: Vinay S, Satya Swaroop Gudipudi, Aman G **Team Number**: 14

Introduction:

Objective: Predict the intensity of hate speech in given textual data, from benign to extremely hateful.
Significance: Understanding the nuanced spectrum of hate to better address it on digital platforms like social media.

Methodology:

Baseline Model:
Refer to the baseline model architecture and implement it. The current baseline model architecture is a combination of Bert and BiLSTM. Bert to handle token representations while BiLSTM capture sequential information.

Model Enhancements:
Explore certain latest techniques to improvise the baseline model performance as below:
Fine-Tuning: Replace the existing baseline architecture Bert embeddings with other pretrained embeddings and fine tune them specific to the hate intensity data if required.
Ensemble Modelling: Experiment with different transformer architectures like Bert, XLNet, GPT etc and perform rank average ensemble where individual model scores are ranked and then the average rank is taken as the final prediction.
Contrastive Learning: Convert the regression problem to multi-class to implement SetFit for contrastive learning. By doing this, the model might learn better representations by distinguishing between various intensity classes.
Prompt engineering with LLMs: Explore various prompt styles to see which gives the most accurate results.

Evaluation: Evaluate the overall effectiveness of our approach by calculating Pearson correlation coefficient(r) between predicted HIP scores and human judgment scores. And r value generally varies from -1 to +1, the higher the value the more correlation exists between two variables, in our case between the variables predicted HIP scores and actual human judgment scores.
There are some other metrics also which can be considered for the task like Cosine similarity, RMSE, F1, Accuracy etc.
Evaluation metrics for fairness detection: At every subgroup level we calculate metrics like disparate metrics, Bias AUCs etc by binning the regression scores.

Challenges:

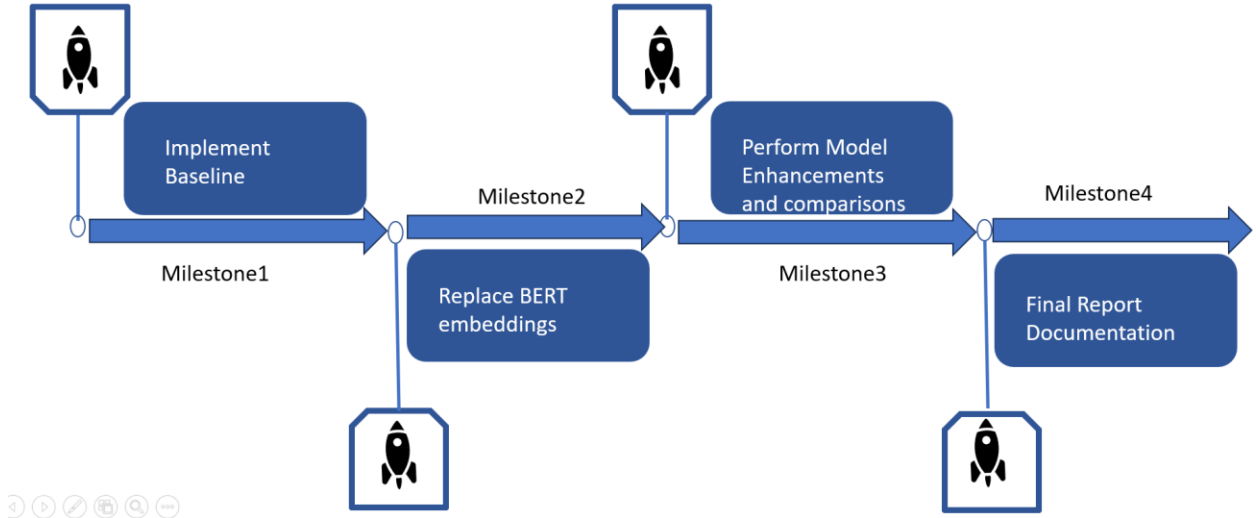
Following are some challenges with developing hate intensity prediction models
Ambiguity: Language is inherently ambiguous. A phrase that seems hateful in one context might be harmless in another.
Bias: Ensuring that the model doesn't inadvertently introduce or perpetuate biases is crucial. This includes biases related to race, gender, religion, etc.

Datasets:

- Hate Norm Gold Dataset
- Hate Speech Data
- Implicit Hate
- Toxic conversations

Timelines:

Milestone 1: Baseline implementation
Milestone 2: Perform model enhancement by replacing BERT with other pretrained embeddings. We plan to have been done with Milestone 2 by our Mid-submission.
Milestone 3: Perform other listed model enhancements and record the individual evaluations.
Milestone 4: Document the final report with findings and observations.



Conclusion:

The approach aims to start with a robust baseline and then methodically enhance its capabilities through advanced techniques. A strong emphasis is placed on not only achieving high accuracy but also ensuring model fairness across different subgroups. By comparing with human judgments, we aim to make the model's predictions more aligned with human perceptions of hate intensity.

References:

- https://arxiv.org/pdf/2206.04007v1.pdf
- https://ieeexplore.ieee.org/document/9679052
- https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
- Efficient Estimation of Word Representation in vector space
- MTEB: Massive Text Embedding Benchmark: Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers [For dataset and embedding reference]