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 either mean voting or 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.

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.

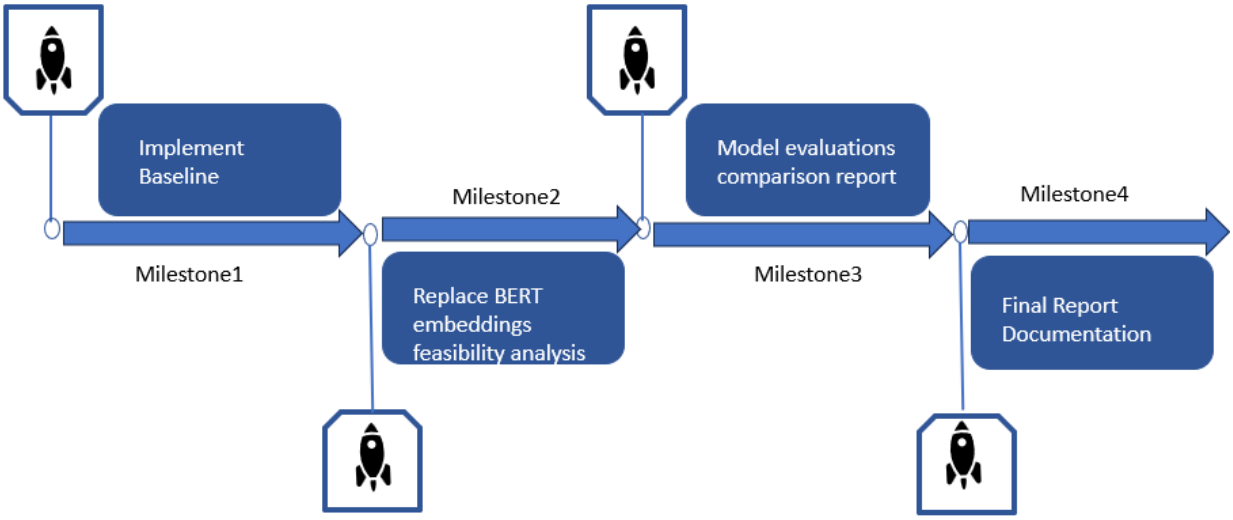
Datasets:

- Hate Norm Gold Dataset
- Hate Speech Data

- Implicit Hate
- Toxic conversations

Timelines:

Milestone 1: Explore data and implement baseline
Milestone 2: Perform model enhancement by replacing BERT with other pretrained embeddings. We plan to attempt this approach and perform feasibility analysis by Milestone 2 by our Mid-submission.(**15th October, 2023**)
Milestone 3: Perform some of the other listed model enhancements in model enhancements section above and record the individual evaluations along with hyper parameter tuning.
Milestone 4: Document the final report with findings and observations. (**19th November, 2023**)



Conclusion:

The approach aims to start with a robust baseline and then methodically enhance its capabilities through advanced techniques. Our strategy emphasizes the importance of acquiring accurate and meaningful word representations. This not only addresses the inherent challenges posed by hate intensity prediction but also ensures that our model captures nuances in context and sentiment. And by comparing with human judgments, we aim to make the model's predictions more aligned with human perceptions of hate intensity.

References:

- NACL HIP module for hate intensity prediction [Sarah Masud et al., 2022]
- https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge
- Efficient Estimation of Word Representation in vector space
- Efficient Few-Shot Learning Without Prompts
- https://ieeexplore.ieee.org/document/9679052
- MTEB: Massive Text Embedding Benchmark: Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers [For dataset and embedding reference]