Hate Intensity Prediction

for Social Media Text

Team 14: Retrievers

G. Satya Swaroop, G. Aman and S. Vinay

Introduction

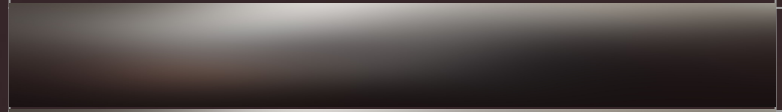
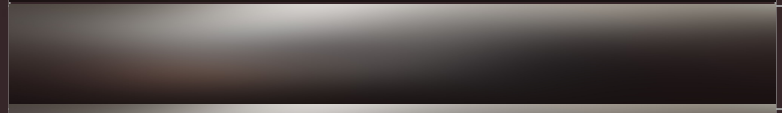
Objective
Predict the intensity of hate speech in given textual data, from neutral to extremely hateful.

Significance
Understanding the nuanced spectrum of hate to better address it on digital platforms like social media. Proper hate prediction can be used to rephrase the hate phrases in a non profane way.

Challenges
Ambiguity and Bias

Datasets

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

Sentence	Score
Arghhhhh I want to kick in the television set right now, @user you d*******le r*t #***	5
	6
	8
	9
	5

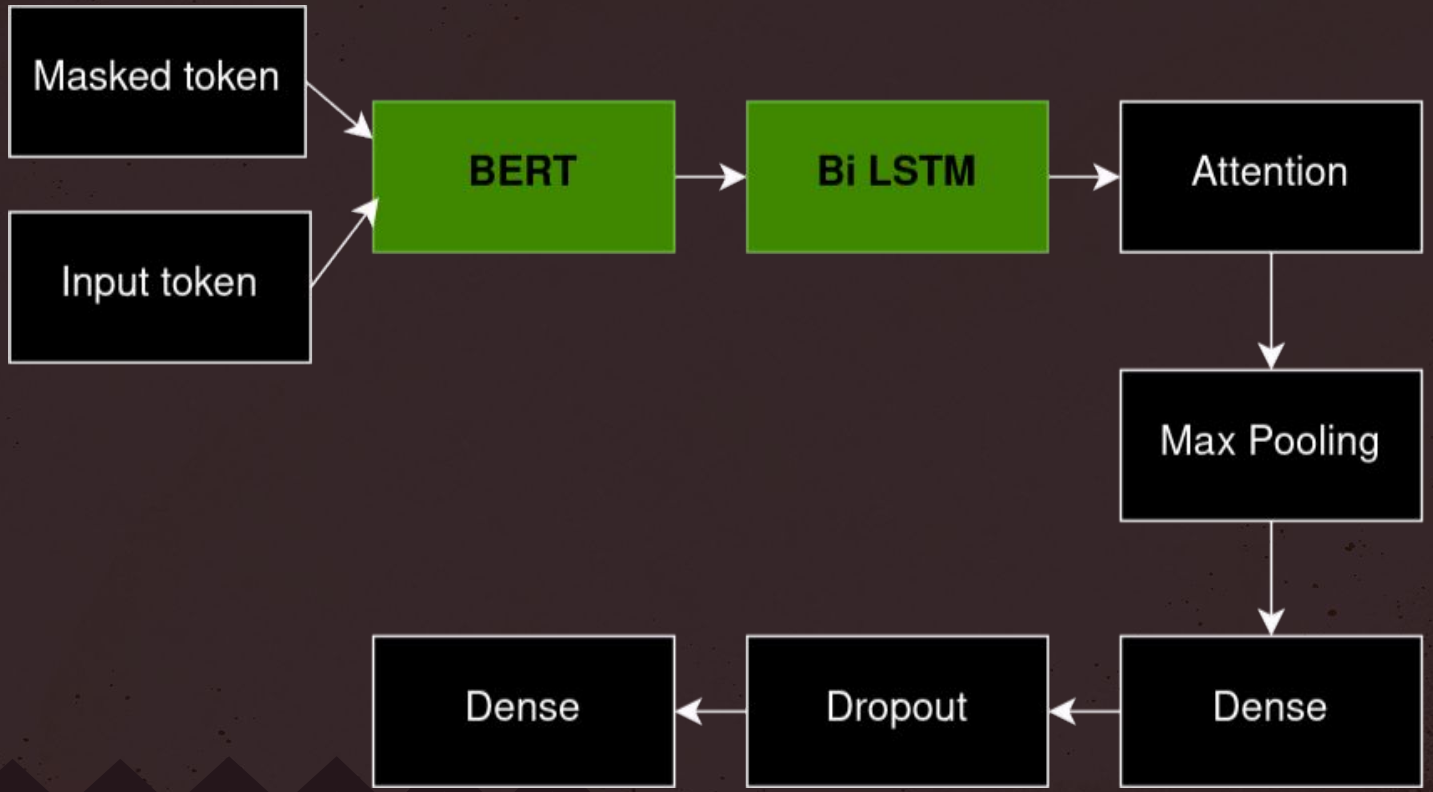
Evaluation

Between the predicted HIP scores and the human judgement scores

- Pearson correlation
- Cosine similarity
- RMSE values

Methodology

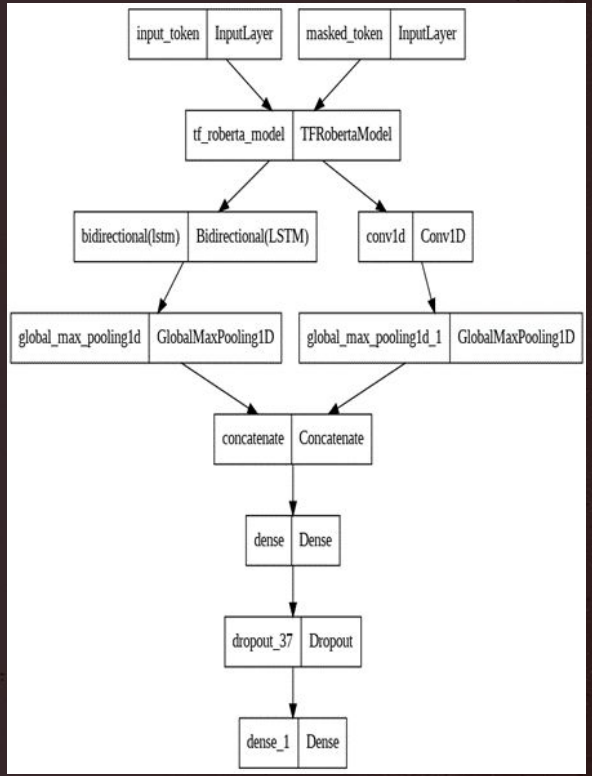
Baseline Model Architecture



Fusion Architectures

We referred to various fusion architectures to handle similar problems like [10] and further applied various fusion based neural architectures that integrates diverse layers to capture rich representation of text for the given task.

The conv1d extracts local and position-invariant features from the embeddings, similar to how convolutional layers in image processing detect features in small regions of an image.



Model enhancements

	Model architecture	Pearson	Cosine	RMSE
0	Only BERT	0.4787	0.9511	1.791
1	BERT + BiLSTM (paper)	0.766	0.973	0.136
	BERT + BiLSTM (our testing)	0.787	0.975	1.3947
2	**RoBERTa + BiLSTM**	**0.818**	**0.98**	**1.1846**
3	**RoBERTa + RCNN**	**0.830**	**0.9799**	**1.1892**
4	**RoBERTa BiLSTM CNN**	**0.844**	**0.981**	**1.1517**
5	**XLnet + BiLSTM**	**0.47**	**0.948**	**2.19**

Stacked meta ensemble model

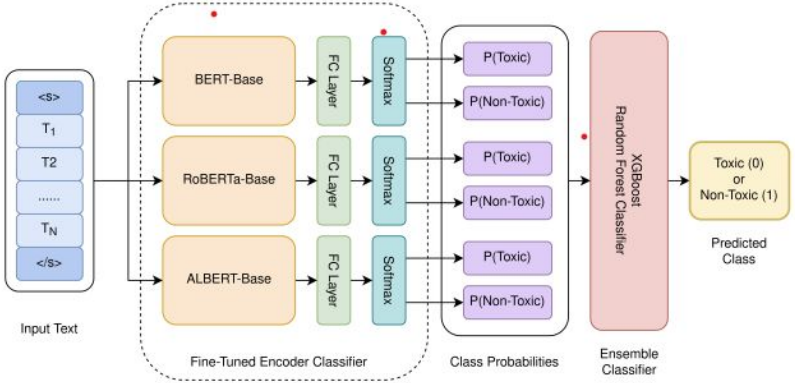


Figure 2: Schematic overview of the architecture of our model.

Reference architecture for ensemble modelling on similar task

Model architecture	Pearson	Cosine	RMSE
Stacked ensemble	0.853	0.983	1.08

Novelty

1. SVO Relative Positional Encoding -> Custom transformer
2. SVO relative positional encoding -> MLP Fusion
3. Exponential positional encoding

1. SVO Relative positional encoding

Hate speech example:
Arghhhhh I want to kick in the television set right now, @user you d*******le r*t #***

Most of the hate speech is characterized by ungrammatical sentences and a loss of semantic adherence to the subject.

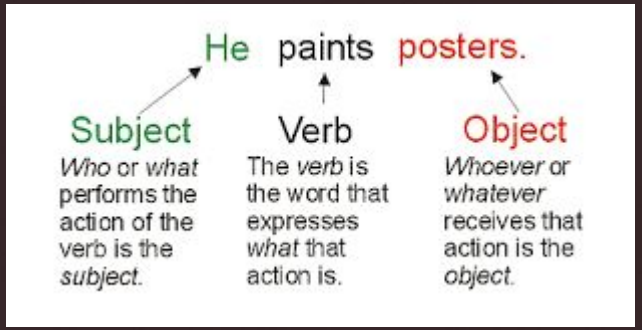
We propose to build a Subject-Verb-Object relative positioning embedding based upon these grammatical inaccuracies found in hate speech online.

These SVO relative positioning embeddings can be used to supplement our better discriminative purposes.

1. SVO relative positional encoding

SVO Relative Position Encoding and Profanity Encoding

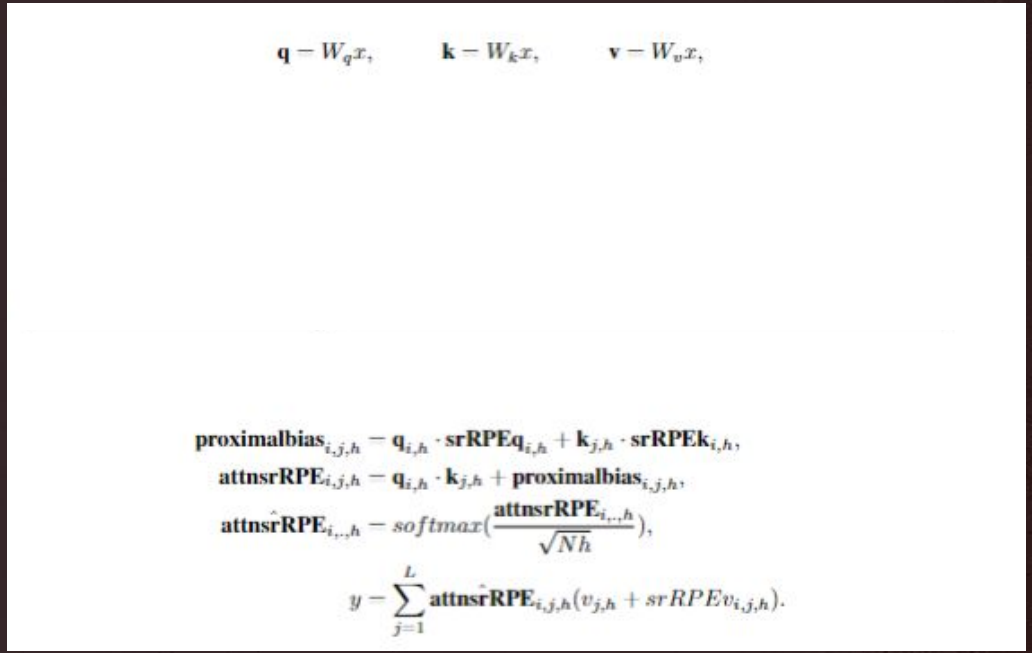
1. In a sentence, we will extract the Subject Verb Object tokens, and construct a neural implicit from their relative positions
2. We will categorize each word as profane or not
3. Then we will construct our own Hate Language Multi Headed Self-Attention (HLG-MSA)



	‘He’	‘paints’	‘posters’
Subject	1	0	0
Verb	0	1	0
Object	0	0	1

1. Custom Transformer head

1. We will have a learnable lookup table for each concatenated SVO Relative Positional Encoding for a Key, Query, Value token to be accompanied with the input token text
2. Then using the semantic embedding of BERT, we will construct our own Transformer models with HLG-MSA blocks
3. If the transformed head does not generalize well we may instead use a RNN framework



$$\mathbf{q} = W_q x, \qquad \mathbf{k} = W_k x, \qquad \mathbf{v} = W_v x,$$

$$\mathbf{proximalbias}_{i,j,h} = \mathbf{q}_{i,h} \cdot \mathbf{srRPEq}_{i,h} + \mathbf{k}_{j,h} \cdot \mathbf{srRPEk}_{i,h},$$

$$\mathbf{attnsrRPE}_{i,j,h} = \mathbf{q}_{i,h} \cdot \mathbf{k}_{j,h} + \mathbf{proximalbias}_{i,j,h},$$

$$\hat{\mathbf{attnsrRPE}}_{i,.,h} = softmax(\frac{\mathbf{attnsrRPE}_{i,.,h}}{\sqrt{Nh}}),$$

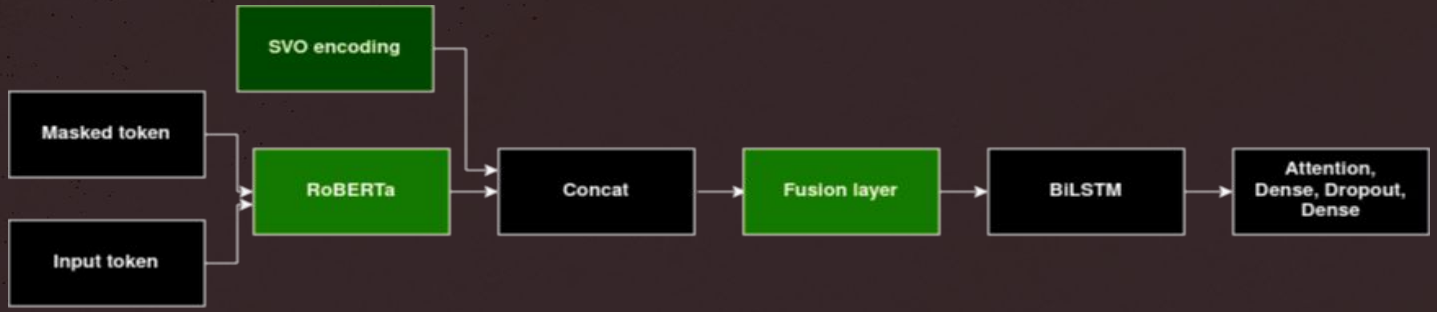
$$y = \sum_{j=1}^{L} \hat{\mathbf{attnsrRPE}}_{i,j,h}(v_{j,h} + srRPEv_{i,j,h}).$$

https://openaccess.thecvf.com/content/ICCV2021/papers/Wu_Rethinking_and_Improving_Relative_Position_Encoding_for_Vision_Transformer_ICCV_2021_paper.pdf

1. SVO + Custom Transformer head

Model architecture	Pearson	Cosine	RMSE
RoBERTA + custom SVO enc injected transformer head	0.1145	0.9343	2.3045

2. SVO encoding fusion with RoBERTA



Model architecture	Pearson	Cosine	RMSE
RoBERTA + BiLSTM	0.818	0.98	1.1846
RoBERTA + SVO enc MLP fusion + BiLSTM	0.8368	0.9809	1.1266

3. Exponential positional encoding

Decaying Exponential Positional Encoding

The positional encoding designed such that later positions in a sequence get a higher weight, and this weight decays exponentially towards the beginning of the sequence. This could be treated as a task-specific design choice for the problem of detecting irony, where the end of the sentence might hold more importance.

Remarks:

- The BERT model within our custom architecture is partially fine-tuned for the task of hate intensity prediction. The first three layers of the BERT model are frozen to preserve foundational language representations, while the subsequent layers are updated during training to adapt to the nuances of our specific task. We enhance the model's capability to understand the significance of token positions in the sequence by incorporating a custom exponential positional encoding. This encoding potentially places greater emphasis on tokens appearing towards the end of the input sequence. During fine-tuning, the trainable layers of the BERT model, along with the BiLSTM and dense layers, adjust their weights to minimize the task-specific loss function, thereby calibrating the model for precise hate intensity assessment.
- This fine tuned bert performs relatively on par with the bert variant recorded in the paper.

3. Exponential positional encoding

	Model architecture	Pearson	Cosine	RMSE
1	Hybrid BERT BiLSTM position enc	0.469	0.949	1.9

All Results

	Model architecture	Pearson	Cosine	RMSE
0	Baseline	0.787	0.975	1.3947
1	**Stacked ensemble**	**0.853**	**0.983**	**1.08**
2	Hybrid BERT BiLSTM position enc	0.469	0.949	1.9
5	RoBERTA + custom SVO enc injected transformer head	0.1145	0.9343	2.3045
6	**RoBERTA + SVO enc MLP fusion**	**0.8368**	**0.9809**	**1.1266**

Hyperparameter trials with optuna

	number	value	datetime_start	datetime_complete	duration	params_dense_dropout	params_dense_units	params_lstm_dropout	params_lstm_units	state
0	0	1.705589	2023-11-20 14:21:59.678938	2023-11-20 14:24:50.716875	0 days 00:02:51.037937	0.123092	128	0.242514	512	COMPLETE
1	1	1.869526	2023-11-20 14:24:50.718600	2023-11-20 14:27:37.980608	0 days 00:02:47.262008	0.388297	32	0.100739	512	COMPLETE
2	2	1.857027	2023-11-20 14:27:37.985563	2023-11-20 14:30:25.054431	0 days 00:02:47.068868	0.492689	128	0.144967	512	COMPLETE
3	3	1.962702	2023-11-20 14:30:25.056499	2023-11-20 14:33:16.394910	0 days 00:02:51.338411	0.292910	32	0.274939	512	COMPLETE
4	4	1.816723	2023-11-20 14:33:16.396903	2023-11-20 14:35:48.417365	0 days 00:02:32.020462	0.271246	128	0.152795	512	COMPLETE
5	5	2.302576	2023-11-20 14:35:48.419210	2023-11-20 14:38:13.692055	0 days 00:02:25.272845	0.232407	128	0.491553	128	COMPLETE
6	6	1.878226	2023-11-20 14:38:13.695147	2023-11-20 14:40:58.740070	0 days 00:02:45.044923	0.130036	32	0.315010	64	COMPLETE
7	7	1.732225	2023-11-20 14:40:58.744247	2023-11-20 14:43:50.461545	0 days 00:02:51.717298	0.110690	64	0.215035	512	COMPLETE
8	8	2.185067	2023-11-20 14:43:50.463997	2023-11-20 14:46:40.254193	0 days 00:02:49.790196	0.461438	128	0.303577	128	COMPLETE
9	9	2.065182	2023-11-20 14:46:40.256642	2023-11-20 14:49:31.582869	0 days 00:02:51.326227	0.415360	32	0.334122	256	COMPLETE

Achievements and Learnings

- We are able to significantly improve the baseline performance by changing the embedding from Bert to Roberta representations.
- We have referred to a similar problem that is “Sarcasm detection”[10] and have implemented similar architecture that is Roberta + RCNN for the current problem statement “Hate Intensity prediction” and results are improving by small fractions.
- We explored finetuning with various transformer architectures like Bert, Roberta and XLNet and a stacked ensemble of these models is outperforming the baseline.
- Our Novel SVO Encoding fusion with Roberta is also performing better than the baseline model results we obtained.
- Our Bert variant finetuning with novel exponential task specific encoding is relatively performing Bert variant mentioned in the reference paper

Further improvements

"After closely studying the problem, we identified some potential improvements worth exploring:

- Data Cleaning: Clean the label errors with state of the art confident learning algorithms that estimate label noises.
- Incorporate different pooling strategies, like average pooling or attentive pooling, which may capture different aspects of the sequence.
- Decoding with LLM-based Models: It might be intriguing to explore how LLM-based models perform in decoding for regression tasks.
- Hyperparameter Tuning: We also recommend further tuning of hyperparameters, especially when experimenting with different model embeddings.“

Appendix

[1] NACL HIP module for hate intensity prediction [Sarah Masud et al., 2022]
[2] https://arxiv.org/abs/2209.11055
[3] https://aclanthology.org/N19-1423.pdf
[4] https://papers.nips.cc/paper_files/paper/2019/hash/dc6a7e655d7e5840e66733e9ee67cc69-Abstract.html
[5] Efficient Estimation of Word Representation in vector space
[6] Efficient Few-Shot Learning Without Prompts
[7] https://ieeexplore.ieee.org/document/9679052
[8] MTEB: Massive Text Embedding Benchmark: Niklas Muennighoff, Nouamane Tazi, Loïc Magne, Nils Reimers
[9] https://openaccess.thecvf.com/content/ICCV2021/papers/Wu_Rethinking_and_Improving_Relative_Position_Encoding_for_Vision_Transformer_ICCV_2021_paper.pdf
[10] A transformer based approach to Irony and Sarcasm detection