
Hate Intensity Prediction for social media text

Team 14: Satya Swaroop G, Aman G, Vinay S

Overview

In the endeavor to assess and mitigate hate speech on digital platforms such as social media, our objective is to develop a scoring system ranging from 1 to 10, reflecting the spectrum from benign expressions to highly hateful ones. This initiative aims to comprehend the nuanced nature of hate speech, acknowledging the inherent ambiguity in language where a seemingly harmful phrase in one context may prove innocuous in another. An essential challenge lies in tackling biases to ensure that the model neither unintentionally introduces nor perpetuates prejudices related to race, gender, religion, and other factors. The ultimate goal is to leverage accurate hate prediction for the purpose of rephrasing offensive phrases in a manner that is non-profane yet effectively addresses the issue at hand.

We have identified benchmark datasets for which baselines[1] are available and performed exploratory data analysis to identify any patterns and applied tokenization by ensuring we filter noise in the data and finally we experimented with various model architectures by incrementally updating the baseline architecture. The predicted scores from the trained model are evaluated with metrics like Pearson, Cosine Sim, RMSE against baseline results.

In our exploration of enhancing the baseline architecture, we delved into the integration of RoBERTa with various modifications, such as incorporating BiLSTM, RCNN, and a combination of both (BiLSTM RCNN). And we built a stacked ensemble of these approaches.

Additionally, we pursued a novel approach by introducing a **Subject-Verb-Object (SVO) relative binary positional encoding** specifically tailored for hate sentences. To implement this, we experimented with two distinct methods: developing our own Hate Language Multi-Headed Self-Attention (**HLG-MSA**) transformer by injecting the SVO encoding, and integrating the SVO encoding directly into RoBERTa embeddings through a fusion layer. Notably, the latter method demonstrated significant success, highlighting its promise as a viable direction for further exploration.

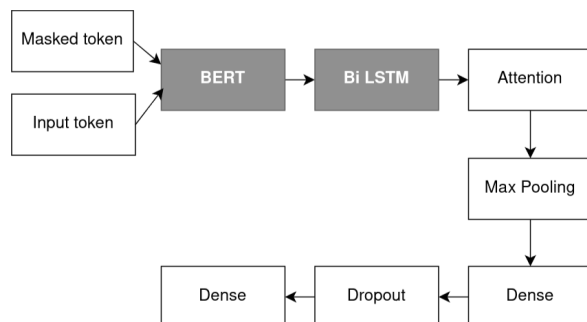
Furthermore, our other novelty extends to the creation of a hybrid BERT model featuring **exponential positional encoding**. This specialized design places increased emphasis on tokens situated near the end of a given sequence. This could be treated as a task-specific design choice for the problem of detecting hate, where the end of the sentence might hold more importance.

Dataset

[Link to dataset](#)

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

Baseline



Various Layers used in baseline and experiments as enhancements:

Input Layer: Takes input_ids and input_masks as inputs.

BERT Embedding Layer: Transforms the input IDs into embeddings.

Bi-directional LSTM: Processes the embeddings and captures sequential information from both directions.

Attention Layer: (Optional) Applies self-attention to the LSTM outputs, focusing on different parts of the sequence.

Global Max Pooling: Reduces dimensionality by retaining max values from LSTM/Attention outputs.

Conv1D: A 1D Convolution layer applies filters on the BERT embeddings, potentially capturing local patterns or n-gram features from the embeddings.

Concatenation: The max-pooled outputs of both the LSTM and Conv1D layers are concatenated. This action effectively merges the features learned from both paths.

Dense Layer: Fully connected layer that can learn representations from the previous layer.

Dropout: Reduces overfitting by dropping out nodes during training.
 Output Layer: Produces the final prediction of hate intensity.

Model Advancements

We have experimented with modifying the baseline architecture, by replacing the BERT embeddings with RoBERTa. Here are the results:

	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

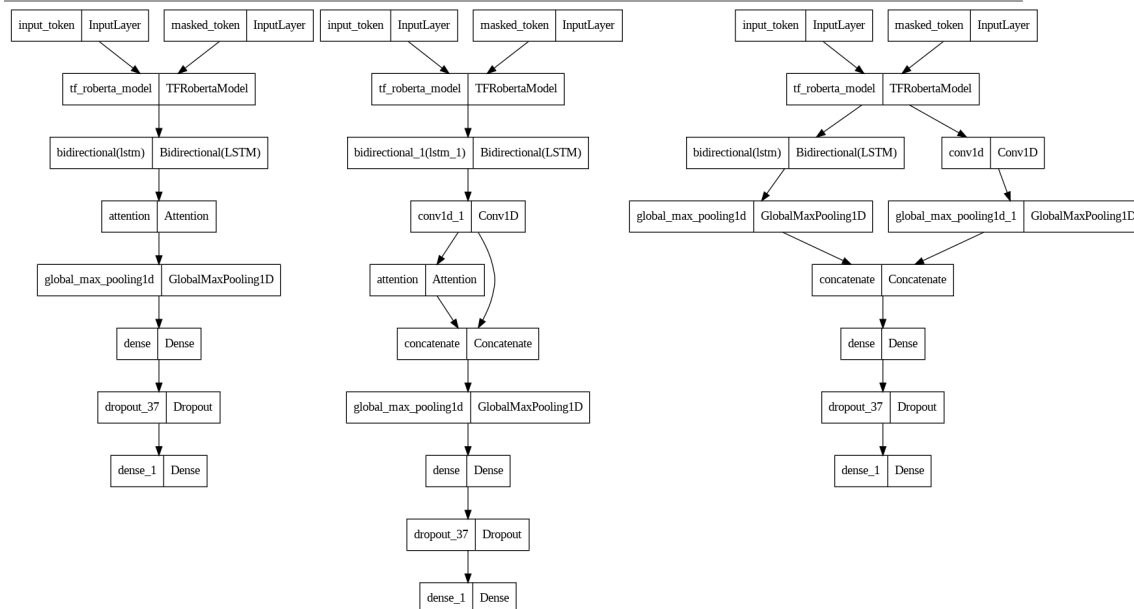


Fig : (from left to right) Roberta + BiLSTM; Roberta + RCNN; Roberta + BiLSTM + CNN

So we will be making a stacked ensemble of our three experiments' models that is Bert, Roberta and XLNet base with random forest regressor bayesian optimized as meta model. The results are as follows:

Model architecture	Pearson	Cosine	RMSE
Stacked ensemble	0.853	0.983	1.08

Novelty

1. SVO Relative positional encoding -> custom transformer (HLG-MSA)

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.

We obtain the SVO binary encodings from the spacy library. In every hate sentence, each token has an associated 'Subject', 'Verb', and 'Object' binary encoding as illustrated below.

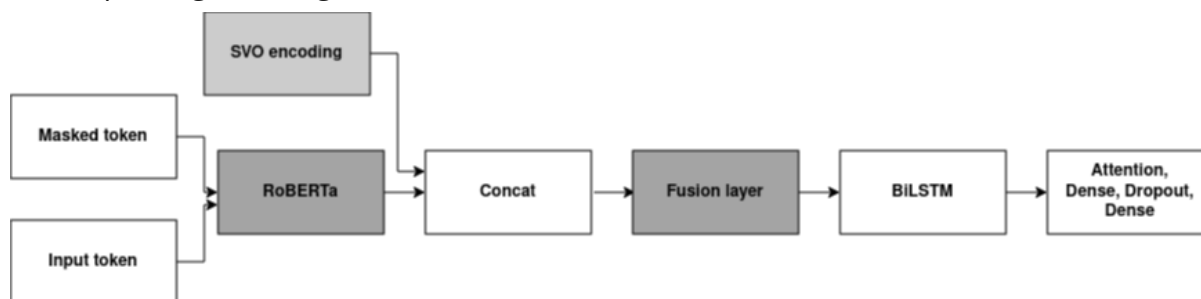
The SVO encoding in a binary encoding, and will be of shape Nx3

Sentence	'Cows'	'eat'	'grass'
Subject	1	0	0
Verb	0	1	0
Object	0	0	1

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)
4. 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
5. Then using the semantic embedding of BERT, we will construct our own Transformer models with HLG-MSA blocks
6. If the transformed head does not generalize well we may instead use a RNN framework

2. SVO Relative positional encoding -> MLP Fusion

For each sentence, to the embedding we obtained using RoBERTa, we shall append the SVO relative positional binary encodings to the end, and then pass it through a fusion layer, before passing it through the BiLSTM. The architecture is as follows:



From the results in the table at the end, we can observe that simply injecting the SVO encodings to the RoBERTa embeddings, then passing the concatenated embeddings to a fusion layer, can improve the performance of the model considerably.

3. Hybrid Bert BiLSTM with exponential positional encoding -> Custom Bert

Decaying Exponential Positional Encoding

The positional encoding is 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 hate, where the end of the sentence might hold more importance.

The way it works is, the bert embeddings are passed to a custom layer, `self.positional_encoding`, which applies exponential positional encoding to the embeddings. The purpose of this layer is to modify the embeddings by adding information about the position of each token in the sequence, potentially giving more weight to tokens that appear later in the sequence, as suggested by the "exponential" aspect of the encoding.

Observation:

The Bert model is fine tuned to detect hate intensity using the custom exponential positional encoding. During this fine-tuning, the weights of all layers in the BERT model can be updated based on the new task-specific objective(hate detection) for the new dataset.

This fine tuned bert performs relatively on par with the bert variant recorded in the paper.

Results

Results from all our advancements and novelties are below:

	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 with exponential positional encoding	0.469	0.949	1.9
3	RoBERTA + custom SVO enc injected transformer head	0.1145	0.9343	2.3045
4	RoBERTA + SVO enc MLP fusion	0.8368	0.9809	1.1266


Hyperparameter tuning

With optuna we performed hyperparameter trials and then further manually tuned as per the trial results achieved.

number	value	datetime_start	datetime_complete	duration	params_dense_dropout	params_dense_units	params_lstm_dropout	params_lstm_units	state
0	0	2023-11-20 14:21:59.678938	2023-11-20 14:24:50.716875	0 days 00:02:51.037937	0.123082	128	0.242514	512	COMPLETE
1	1	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	2023-11-20 14:27:37.985563	2023-11-20 14:30:25.054431	0 days 00:02:47.068868	0.482689	128	0.144967	512	COMPLETE
3	3	2023-11-20 14:30:25.056499	2023-11-20 14:33:16.394910	0 days 00:02:51.338411	0.282910	32	0.274939	512	COMPLETE
4	4	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	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	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	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	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	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

Conclusion

In conclusion, our team has undertaken a comprehensive exploration of hate intensity prediction for social media text, addressing the critical need to identify and mitigate hate speech on digital platforms. Through the development of a scoring system ranging from 1 to 10, our objective was to capture the spectrum of expressions from benign to highly



hateful. The complexity of hate speech, influenced by contextual nuances and potential biases, was acknowledged and carefully navigated throughout our model development. Our initiatives included the integration of RoBERTa with various modifications, such as BiLSTM, RCNN, and a combination of both, leading to a stacked ensemble model that showcased significant improvement. Novelty was introduced through the creation of a Subject-Verb-Object (SVO) relative positional encoding, demonstrating promise in enhancing discriminative capabilities and suggesting its potential utility in future applications. Additionally, a hybrid BERT model with exponential positional encoding was devised, emphasizing the potential importance of sentence endings in hate detection. Our results highlight the effectiveness of our approaches in predicting hate intensity, paving the way for further exploration and refinement in the ongoing effort to combat online hate speech.

References

1. [Hate Intensity Prediction baseline paper](#)
2. [A transformer based approach to Irony and Sarcasm detection](#)
3. [Rethinking and Improving Relative Positional Encoding for Vision Transformer](#)