HateXplain2.0: An Explainable Hate Speech Detection Framework Utilizing Subjective Projection from Contextual Knowledge Space to Disjoint Concept Space

Md Fahim[1], **Md Shihab Shahriar**[2], **Mohammad Sabik Irbaz**[2], **Syed Ishtiaque Ahmed**[3] & **Mohammad Ruhul Amin**[4]

[1]CCDS Lab, Independent University Bangladesh, [2]Islamic University of Technology, [3]University of Toronto, [4]Fordham University

Highlights

- Word level attention scores per class for each word in a sentence. It will help to understand more about model interpretability.
- Increase model explainbility along with the model performance.
- Defining a learnable concept space per class which holds class specific information
- Word level attention scores define from the concept space which are separable from each to another via a loss function.
- A bias score per word can be defined from the attentions from the concept space to measure word level efficiency.
- A novel framework for model explainability and interpretability

Introduction

Toxicity and hate speech detection is one of the most important tasks of NLP in the era of internet. LLMs can play a vital role to detect the toxicity and hate speech. Besides, it also very important to explain the model on how good the model is to detect real-life hate speech. In this project, we introduce a novel framework for detecting hatespeech utilizing the LLMs via concept space projection. We project the contextual word embeddings of a sentence into class specific concept space from which we get class level attention scores for words. The concept is learnable and explainable. We analyse the effectiveness of our model in HateXplain[1] two and three class dataset.

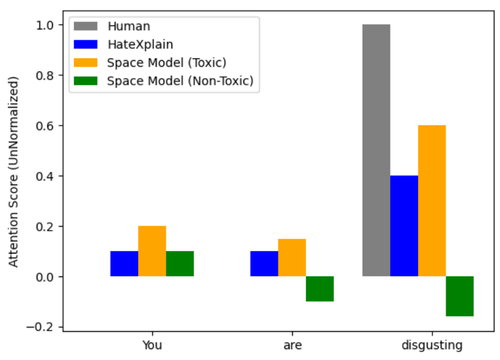

Figure 1:Comparison of Attention Scores of Space model with HateXplain and Human Rationales for a toxic sentence. For binary toxic comment classification, the Space model provides both non-toxic and toxic attention scores for each individual word of a sentence.

In the figure 1, we demonstrated how our model works for a sentence. If there are two class in the dataset, we will get two different attentions per word for each class in a sentence.

Methodology

The paper proposes a framework for hate speech detection that uses BERT to generate contextual word embeddings for a given input sentence and then projects them onto two conceptual spaces, a hate space, and a non-hate space, for classification. The training objective is to learn embeddings that are far apart from each other in the conceptual space for hate and non-hate words, and also prevent the embeddings from converging to the same word embedding within the conceptual space. The model is trained using an inter-space loss and an intra-space loss along with binary cross-entropy loss for classification. A brief explanation of inter-space loss and intra-space loss is given below.

- Inter-space loss: The inter-space loss is used to ensure that the embeddings of hate and non-hate words are far apart from each other in the conceptual space. The loss is calculated as the distance between the mean vectors of the hate and non-hate spaces. The model is trained to minimize this loss, which encourages the embeddings to be well-separated in the conceptual space.
- Intra-space loss: The intra-space loss is used to prevent the embeddings from converging to the same word embedding within the conceptual space. The loss is calculated as the distance between each word embedding and the mean vector of its corresponding conceptual space. The model is trained to minimize this loss, which encourages the embeddings to be diverse within the conceptual space.

Model Architecture

The proposed framework in the paper uses BERT to generate contextual word embeddings for a given input sentence and then projects them onto two conceptual spaces, a hate space, and a nonhate space, for classification.

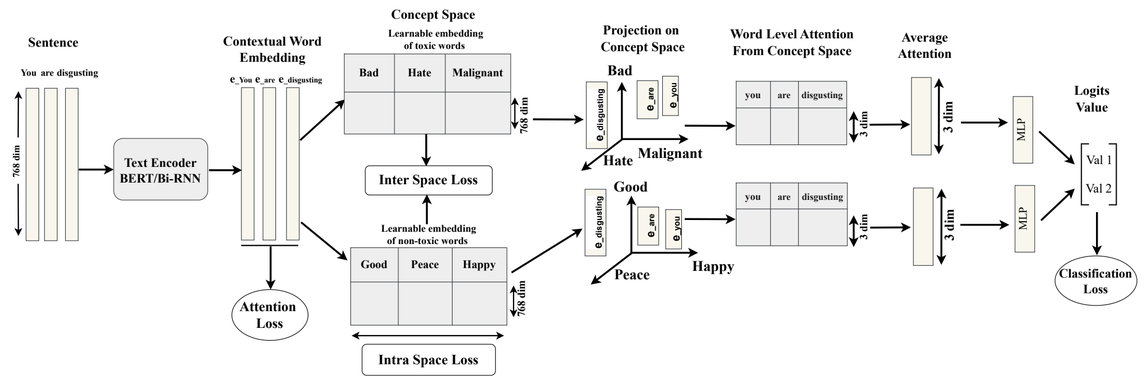

Figure 2:Model Architecture of SpaceModel.

Result Analysis

The paper evaluates the proposed framework on the HateXplain dataset for both 2-class and 3-class scenarios. The performance of the proposed Space Model is compared with the HateXplain model using classification metrics. The results show that the subjective projection of sentence representations onto task-specific conceptual spaces improves the performance metrics of the model.

Model [Token Method]	Performance			Explainability				
				Plausibility			Faithfulness	
	Acc.↑	Macro F1↑	AUROC↑	IOU F1↑	Token F1↑	AUPRC↑	Comp.↑	Suff.↓
BiRNN-HateXplain[LIME]	0.629	0.629	0.805	0.174	0.407	0.685	0.343	-0.075
BiRNN-SpaceModel [LIME]	0.618	0.614	0.781	0.195	0.334	0.568	0.335	0.090
BiRNN-HateXplain [Attn]	0.629	0.629	0.805	0.222	0.506	0.841	0.281	0.039
BiRNN-SpaceModel [Attn]	0.624	0.612	0.786	**0.353**	**0.544**	0.848	0.208	**-0.0025**
BERT-HateXplain [LIME]	0.698	0.687	**0.851**	0.112	0.452	0.722	0.500	0.004
BERT-SpaceModel [LIME]	0.695	0.688	0.812	0.277	0.466	0.729	**0.579**	0.053
BERT-HateXplain [Attn]	0.698	0.687	**0.851**	0.120	0.411	0.626	0.424	0.160
BERT-SpaceModel [Attn]	**0.701**	**0.693**	0.826	0.133	0.515	**0.881**	0.538	0.035

Table 1:Comparison of performance and explainability metrics of our Space Model with HateXplain model on 3 class HateXplain Dataset.

The results shows that the proposed framework achieved better accuracy and explainability compared to the baseline models. On the HateXplain dataset, the proposed model showed at least a 10% improvement in various explainability metrics. The BiRNN models outperformed BERT in explainability metrics, although they tended to generate more false rationales. In terms of token methods, the attention-guided token method outperformed the LIME attention approach in the proposed model.

Analysis of Attention Scores from Concept Space

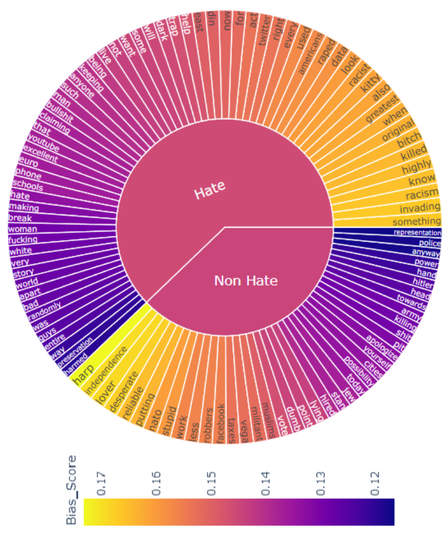

Figure 3:SunBrust Plot of Words Based on the Bias Score. If the bias score of a word is greater than 0 then it is labeled as a Non-Hate word otherwise labeled as a Hate word in the plot.

To get more intuition about the attention scores derived from the concept space, we measured average hate attention scores and average non-hate scores for each word based on their occurrences in different sentences. Then we define a **Bias_Score** for each word as the following:

$$\text{Bias\_Score}_{\omega} = \frac{\sum \text{Non-HateAttn}_{\omega}}{n_{\omega}} - \frac{\sum \text{HateAttn}_{\omega}}{n_{\omega}} \quad (1)$$

where n_{ω} is the no. of occurrence of word ω in test dataset. Following the equation in 1, we refer a word bias to Non-Hate class if its bias score is greater than 0 otherwise the word is biased to Hate class.

Concept Words Analysis

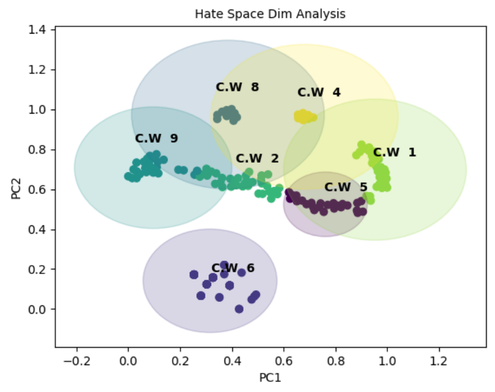

Figure 4:Visualization of Hate Concept Space with Space Dim 10. Here, C.W stands for Concept Words in that space. Within the hate space, the concept words form distinct clusters that partially overlap with each other.

In Figure 4, we present a 2D t-SNE plot illustrating the projected word representations of the sentences' words in the hate space. The plot reveals that the words tend to cluster around specific conceptual word embeddings, providing insights into the types of words represented by these conceptual word embeddings. This analysis offers an explanatory perspective on the hate space and its association with different word categories.
Similar experiment was exceuted for non-hate concept space to analyse the non hate concept dimensions.

2 class Experiments

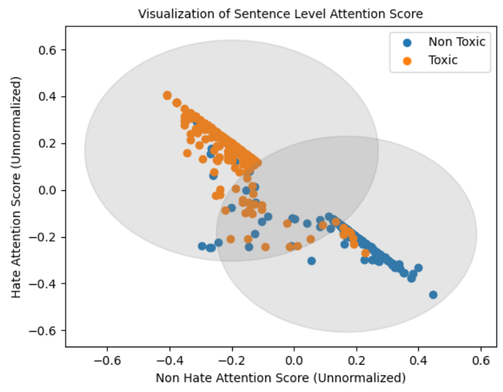

Figure 5:Visualization of Sentence Level Attention Scores (which is the mean of concept words attention scores for a concept space) from the Concept Space of Space Model on 2 class HateXplain dataset.

We also evaluated our model effectiveness in sentence level attentions. Measuring the average attentions of each words, we defined the sentence level attention scores per class. We observed a good clusters between the class labels.
Result for 2 class HateXplain dataset in terms of model performance where we can see we beat the HateXplain model by a good margin.

Model Name	Performance		
	Acc.↑	Macro F1↑	AUROC↑
BERT HateXplain	0.845	0.841	0.882
Space Model + BERT [Attn]	**0.893**	**0.892**	**0.924**

Table 2:Benchmarking on HateXplain Dataset (2 Class). We use the attention-based token method for this experiment. Space model gives around 5% improvement in the model performance than the BERT HateXplain model

Conclusion

The paper proposes a novel approach, Subjective Projection on Conceptual Space named as Space Model, for fine-tuning pre-trained natural language processing models for text classification tasks. The proposed model not only improves performance but also enhances explainability by learning the representations of conceptual words that are specific to a particular class. The results of the experiments suggest that the proposed approach can create a more effective representation for task-specific classes and provide a new method for adapting the knowledge of pre-trained language models. The findings open up opportunities for further research on improving the performance and explainability of downstream tasks through novel fine-tuning strategies. Therefore, the paper concludes that the proposed framework is a promising approach for hate speech detection and can be extended to other text classification tasks. We use LSTM[2] and [3] to extract contextual word embeddings from the sentences.

Findings

- Proposing a novel approach, Subjective Projection on Conceptual Space (Space Model), for fine-tuning pre-trained natural language processing models for text classification tasks.
- Enhancing the explainability of the model by learning the representations of conceptual words that are specific to a particular class.
- Introducing intra-and interspace losses to optimize the training of the operators.

References

[1] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee.
Hatexplain: A benchmark dataset for explainable hate speech detection.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, 2021.

[2] Sepp Hochreiter and Jürgen Schmidhuber.
Long short-term memory.
Neural computation, 9(8):1735–1780, 1997.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova.
Bert: Pre-training of deep bidirectional transformers for language understanding.
arXiv preprint arXiv:1810.04805, 2018.

[4] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov.
Roberta: A robustly optimized bert pretraining approach.
arXiv preprint arXiv:1907.11692, 2019.

[5] Pengcheng He, Jianfeng Gao, and Weizhu Chen.
Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing.
arXiv preprint arXiv:2111.09543, 2021.