Explainable Identification Of Hate Speech Towards Islam Using Graph Neural Networks

Azmine Toushik Wasi
Shahjalal University of Science and Technology, Bangladesh

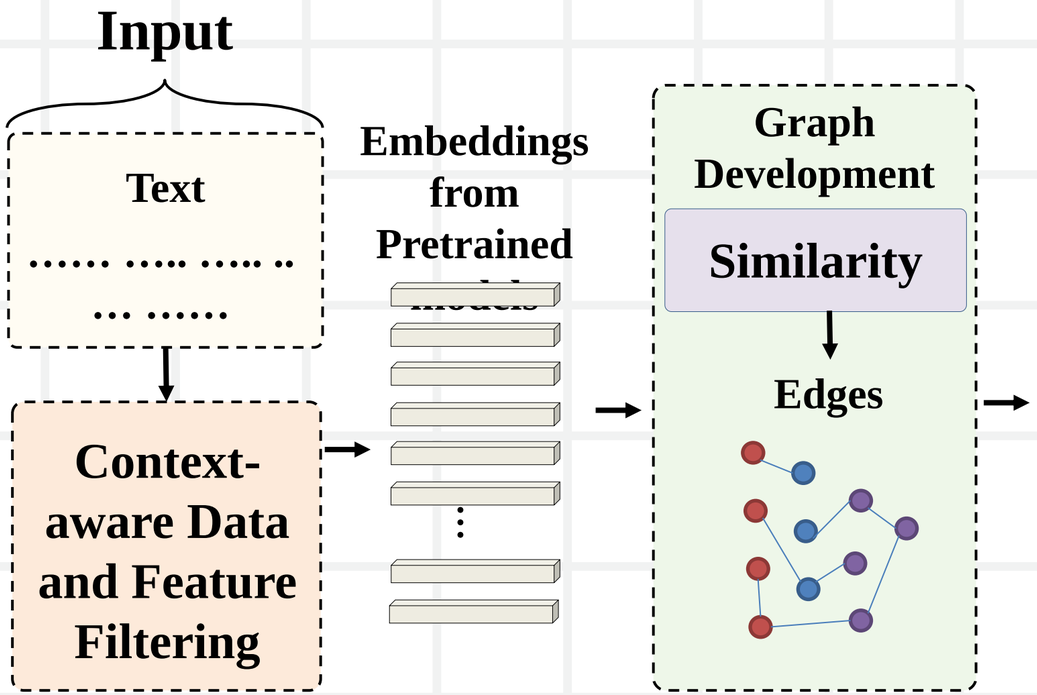
Muslims in ML Workshop
37th Conference on Neural Information Processing Systems (NeurIPS 2023)

MOTIVATION

- **Escalating presence of hate speech specifically targeting Islam or Muslim communities on online discussion platforms is currently a growing concern [1].**
- **GNNs are powerful in analyzing complex, interconnected data as graphs. Using their ability to utilize relations between different datapoints, GNNs have shown tremendous promise in text classification and detection tasks [2].**

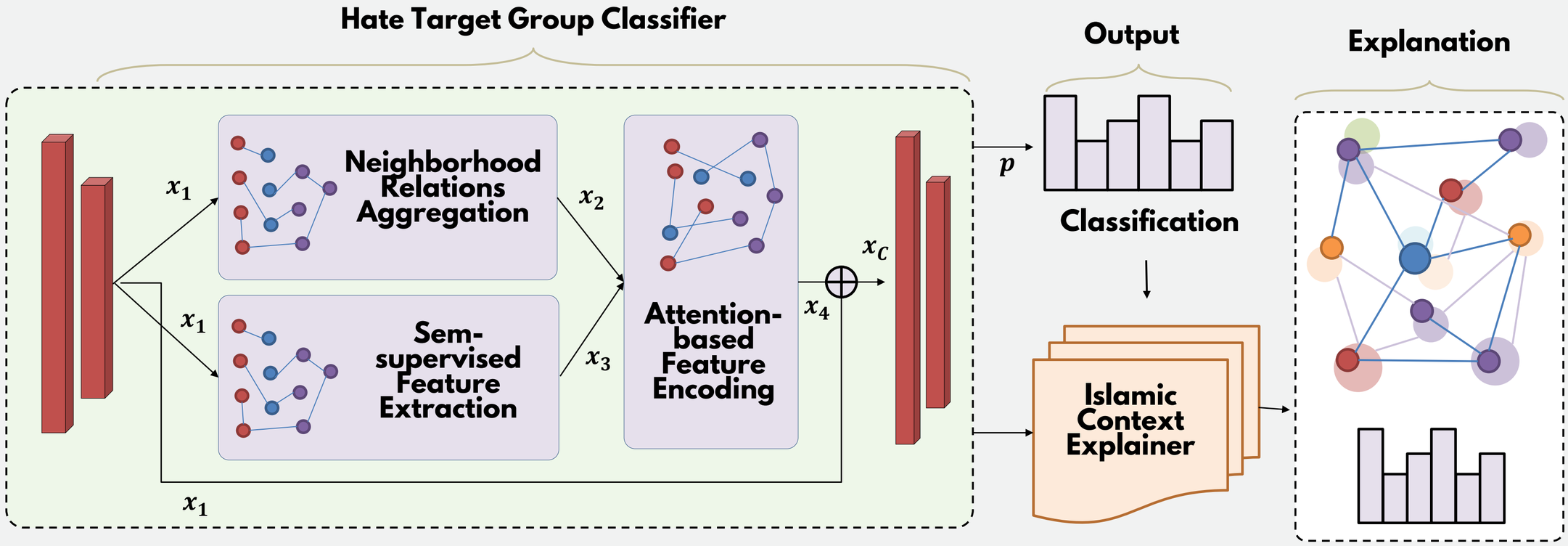
[1] Making muslim the enemy: A transitivity analysis on anti-islam hate speech. Research on Humanities and Social Sciences, 2022
[2] Zhibin Lu, Pan Du, and Jianyun Nie. Vgcn-bert: Augmenting bert with graph embedding for text classification. Advances in Information Retrieval, 12035:369 – 382, 2020.

PRE-PROCESSING



- ✓ Initially, the dataset (HateXplain) is filtered to focus on hate speech targeting Islam.
- ✓ Next, pretrained NLP models is applied to the text to obtain word embeddings.
- ✓ Edges are determined using cosine similarity between embeddings.

NEURAL NET



EXPERIMENTS

Model	Accuracy	Macro F1
CNN-GRU	0.627	0.606
BiRNN	0.595	0.575
BiRNN-HateXplain	0.629	0.629
BERT	0.69	0.674
BERT-HateXplain	0.698	0.687
XG-HSI-BiRNN (Ours)	**0.768**	**0.767**
XG-HSI-BERT (Ours)	**0.791**	**0.797**

Case Study:

How is all that awesome muslim diversity going for you native germans? You have allowed this yourselves. If you do not stand and fight against this. You get what you asked for what you deserve!

↓

Offensive Towards Islam

↓

As per the explainer, the neighboring and self-tokens helped to classify this as offensive to Muslims are:

fight, muslim diversity, brooks, \#\#rish, donald, syrian, schultz, typed.

THANK YOU!

Muslims in ML Workshop
37th Conference on Neural Information Processing Systems (NeurIPS 2023)