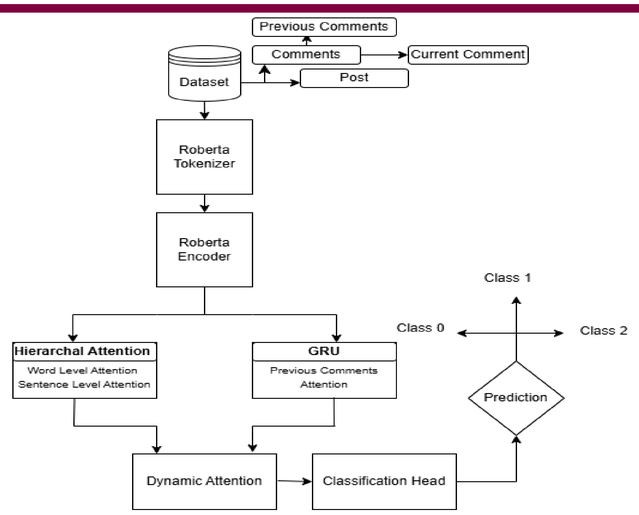
SpecturmNET: An LGBTQ+ Centric Cyberbullying Detection ML Model

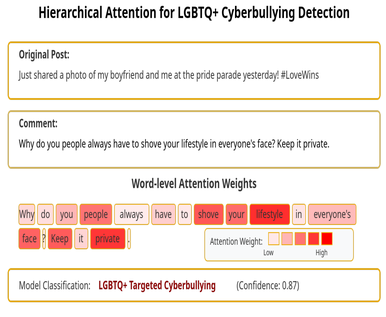
Muhammad Arslan[1], Mohammed Abuhamad[1], Deborah Hall[2], Yasin Silva[1]
1: Loyola University Chicago, 2: Arizona State University

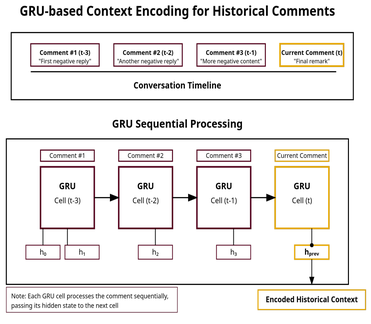
Abstract

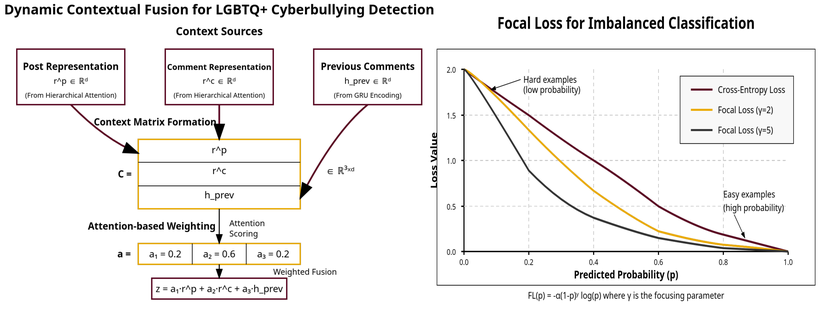
This research introduces SpectrumNET-Full, an advanced transformer-based model specifically designed to detect cyberbullying targeting the LGBTQ+ community in social media texts. Building upon the SpectrumNET-Base model, SpectrumNET-Full incorporates hierarchical attention mechanisms and dynamic contextual fusion to capture both immediate and historical context for more accurate detection of subtle, nuanced forms of online harassment.

Methods









Dataset

Statistics	Value
Total Comments	106,618
Cyberbullying Comments	11,266
LGBTQ+-targeted Comments	1,013
Max Length of Comments	410
Max Length of Cyberbullying Comments	353
Max Length of LGBTQ+-targeted Comments	182
Max Length of Posts	267

Training Hyperparameters

- Learning Rate (LR): 1 × 10−5
- Batch Size: 8
- Number of Epochs: 50
- Loss Function: Focal Loss (γ = 2, α = [1.0, 2.0, 1.0])
- Total Training Steps: (num batches per epoch) × 50
- RoBERTa Freeze Policy: Fully frozen
- Maximum Sequence Length: 512 tokens
- Hidden Dimension: 768
- Dropout Rates Feature Fusion: 0.2, Classifier: 0.3
- Weighted Sampling: Yes (inverse class frequency)

References

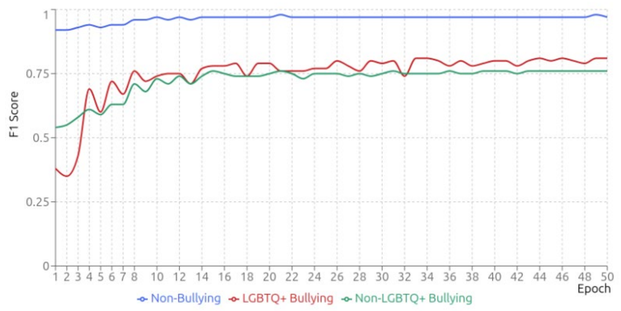
M. Plöderl and P. Tremblay, "Mental health of sexual minorities. a systematic review," International Review of Psychiatry, vol. 27, no. 5, pp. 367–385, 2015, pMID: 26552495. [Online]. Available: https://doi.org/10.3109/09540261.2015.1083949

B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, and J. P. McCrae, "Dataset for identification of homophobia and transophobia in multilingual youtube comments," 2021.

M. Hamlett, G. Powell, Y. N. Silva, and D. Hall, "A labeled dataset for investigating cyberbullying content patterns in instagram," Proceedings of the International AAAI Conference on Web and Social Media, vol. 16, no. 1, pp. 1251–1258, May 2022. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/19376

Results

Class	Accuracy	Precision	Recall	F1	AUROC
Non-Bullying	0.954	0.980	0.953	0.964	0.953
LGBTQ+ Bullying	0.807	0.717	0.807	0.747	
Non-LGBTQ+ Bullying	0.794	0.678	0.795	0.724	



SpecturmNET: F1 Score Per Epoch

Limitations & Future Work

- The underrepresentation of LGBTQ+ bullying instances in the dataset poses challenges. Although techniques like weighted sampling were used, the inherent imbalance may still affect the generalizability of the results.
- RoBERTa restricts the input sequence to a maximum of 512 tokens. When attempting to include previous comments along with the post and current comment, we must either truncate or employ a sliding window strategy. This limitation can result in the loss of important contextual cues, thereby affecting the accuracy of bullying detection.
- Adding multiple comments is inherently challenging. For example, if Comment 1 is correctly classified as bullying while Comment 2 is non-bullying and Comment 3 is also non-bullying, the fusion process might inadvertently aggregate these conflicting signals. As a result, the model may incorrectly label the overall context as bullying due to the influence of earlier, more aggressive inputs.
- The use of a GRU for encoding previous comments assumes a linear progression of context. However, online conversations often exhibit branching or non-linear dynamics. This linear encoding may oversimplify complex conversation structures, leading to a less precise representation of the historical context.
- The additional modules for hierarchical attention and dynamic contextual fusion introduce extra computational overhead and potential training instability. Such complexity might hinder scalability and affect real-time performance, especially when processing large volumes of data.

BullyBlocker

Acknowledgement

This work was supported by National Science Foundation Awards #2227488 and a Google Award for Inclusion Research