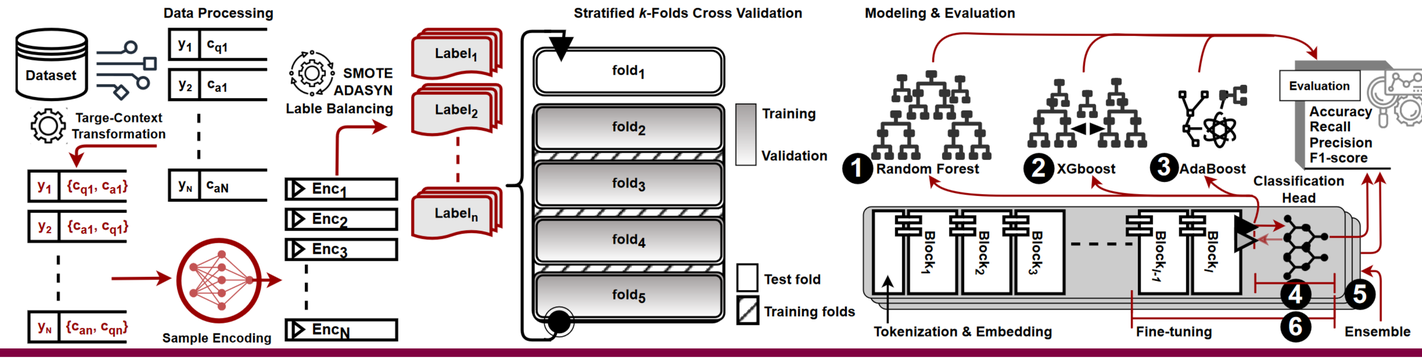
Cyberbullying Role Detection Using LLMs

Manuel Sandoval Madrigal[1], Patrick Furman[1], Mohammed Abuhamad[1], Mujtaba Nazari[1], Juan Zhang[1], Deborah Hall[2], Yasin Silva[1]

1: Loyola University Chicago, 2: Arizona State University

Model Framework



Abstract

Cyberbullying is a global phenomena impacting the mental health of thousands of adolescent. Understanding the dynamics of cyberbullying roles requires large amounts of labeled data on nuanced interactions. This study explores the use of machine learning models to predict the cyberbullying roles, harasser, victim, bystander defender, and bystander assistant [2] in online interactions. Using the AMiCA ASM.fm dataset [3], which consists of113,698 English posts, we evaluated the performance of various models built with four underlying LLMs, i.e., BERT, RoBERTa [1], T5, and GPT-2 for role detection. **A fine-tuned RoBERTa model performed the best, achieving an F1 score of 0.828.**

Methods

- Basic data preprocessing – removal of punctuation, tokenization, and padding shorter comments
- To maintain the connection between the comments, we generate two samples for each Q&A pair in the original dataset, utilizing both labels
- To handle class imbalance, we use ADASYN with its default parameters (neighbors = 15) to oversample the minority samples: Harasser, Victim, Bystander-Defender, and Bystander-Assistant
- Using context-target pair embedding vectors, we fine-tune each LLM to predict the cyberbullying roles of the targets

Data Structure

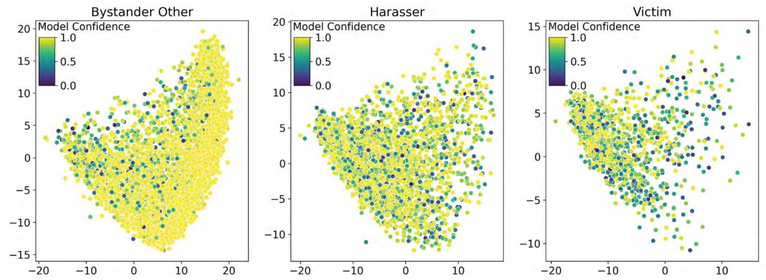
- Posts from users are conversations, each of which consists of one Q&A pair. For our model, we reconstruct the Q&A pair into a context-target embedding vector.
- Each comment is tagged with severity (none, mild, severe) and labeled with role of the author (harasser, victim, bystander defender, bystander assistant). For our purposes, we only focus on the role label.

Comment Text	Role
Q: [awh thats cute that you send yourself messages just for people to think you're not hated]$^{\text{GEN\_INSULT}}$:') looooool you're so gay. [So's your mom]$^{\text{ATTACKING\_RELATIVES}}$:)	Harasser
A: It wasn't me & [don't talk about my mom]$^{\text{ASSERTIVE\_SELF\_DEF}}$ hate in me all you want but you've just past the limit [I am going to find out who you are & I swear you are going to regret it.]$^{\text{THREAT\_BLACKMAIL}}$	Victim

Results From Fine-tuned LLMs

LLM	Accuracy	Recall	Precision	F1	Top-2 F1
BERT	0.721	0.721	0.726	0.723	0.898
RoBERTa	**0.830**	**0.830**	**0.828**	**0.828**	**0.955**
T5	0.682	0.682	0.702	0.688	0.893
GPT2	0.544	0.544	0.533	0.546	0.828

- The fine-tuned RoBERTa [1] model performs well in most situations, but struggles in some surprising situations such as distinguishing between harassers and victims
- An example of model confusion is mistaking harassers for victims and vice versa, but there are several other instances of overlap, such as harassers and bystander assistants, which have very similar behavior in cyberbullying interactions. This can be explained as victims aggressively defending themselves, which can make it appear as the harasser



Future Work

- Collecting a dataset with severity of bullying, the topic of bullying, and improved role labels
- Models purposefully built for identifying bystander interventions (anti-bullying)
- Conversation-level analysis to identify patterns across several posts for a single user

References

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. CoRR, abs/1907.11692.

Salmivalli, C., Lagerspetz, K., Björkqvist, K., Österman, K., & Kaukiainen, A. (1996). Bullying as a group process: Participant roles and their relations to social status within the group. Aggressive Behavior, 22(1), 1–15. https://doi.org/10.1002/(SICI)1098-2337(1996)22:1<1::AID-AB1>3.0.CO;2-T

Van Hee C, Jacobs G, Emmery C, et al. Automatic detection of cyberbullying in social media text. PLoS One. 2018;13(10):e0203794. Published 2018 Oct 8. doi:10.1371/journal.pone.0203794

BullyBlocker
http://bullyblocker.cs.luc.edu/

Acknowledgement

This work was supported by National Science Foundation Awards #2227488 and #1719722, and a Google Award for Inclusion Research