Cyberbullying Role Detection Using LLMs

Manuel Sandoval Madrigal[1] Mohammed Abuhamad[1], Deborah Hall[2], Yasin Silva[1]
1: Loyola University Chicago, 2: Arizona State University

Abstract

Cyberbullying is a global phenomena impacting the mental health of thousands of adolescents. 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: bully, bully assistant, non-aggressive defender, aggressive defender, non-aggressive victim, aggressive victim, and passive bystander [2]. Using an Instagram dataset of 103,371 individually labeled comments, we explored fine-tunning the large language model RoBERTa [1] for role classification.

Methods

Convert each comment into their respective Input Ids and Attention Mask using the pretrained RoBERTa [1] byte-level BPE tokenizer which has a vocabulary of 50K sub-word units.

The dataset is extremely imbalanced, hence we use **weighted random sampling** to deliberately oversample both the difficult and rare classes (e.g., victims and defenders), as opposed to allowing the model to learn the most common and easiest class (passive bystander).

Class Statistics

Roles Majority	Labels	Percent	Imbalance Ratio
Bystander	87878	85.01%	1:1
Bully	7995	7.73%	11:1
Catch all	5285	5.11%	17:1
Agg Defend	871	0.84%	101:1
Non-Agg Defend	737	0.71%	120:1
Non-Agg Victim	373	0.36%	236:1
Agg Victim	154	0.15%	567:1
Bully Assist	78	0.08%	1063:1

Roles Ordered	Labels	Percent	Imbalance Ratio
Bystander	89533	86.61%	1:1
Bully	9153	8.85%	10:1
Agg Defend	1574	1.52%	57:1
Non-Agg Victim	1282	1.24%	70:1
Non-Agg Defend	1260	1.22%	71:1
Agg Victim	356	0.34%	255:1
Bully Assist	213	0.21%	412:1

Example Instagram Comments

Passive Bystander: you should like…make a full image out of this

Non-Aggressive Victim: Height ain't got nothin to do with it! @user_name

Aggressive Victim: Y don't u come the f*** over here and make me @user_name

Non-Aggressive Defender: Don't worry about him he just a Hater @user_name

Aggressive Defender: Just a punk a** Facebook gangster… do ur thing man

Bully: @user_name you're so f***ing gay. Delete your ig.

Bully Assistant: Calm down your mommy bought you all of your sh** @user_name

Experiment Results

Configuration	Accuracy	Precision	Recall	F1
Roles Ordered	87.18%	88.12%	87.18%	87.47%
Grouped Roles Majority	85.67%	86.24%	85.67%	85.80%
Roles Majority	83.95%	85.78%	83.95%	84.66%
Roles Dense Ranked All Ones	81.15%	86.03%	81.15%	83.08%
No Passive Roles Ordered	61.35%	68.98%	61.35%	63.80%
No Passive Grouped Roles Majority	61.15%	64.63%	61.15%	62.23%
No Passive Bystander Roles Majority	59.58%	62.26%	59.58%	60.54%
Roles Dense Ranked	26.38%	32.30%	26.38%	25.76%

Roles Ordered	Precision	Recall	F1
Bystander	95.25%	92.62%	93.91%
Bully	49.50%	67.28%	56.97%
Non Agg Defend	30.03%	28.49%	28.96%
Agg Defender	33.11%	26.24%	28.41%
Non Agg Victim	22.16%	15.91%	18.29%
Bully Assist	17.53%	13.94%	15.26%
Agg Victim	12.33%	8.44%	9.61%

Future Work

Models purposefully built for identifying bystander interventions (anti-bullying)

Modeling role interactions using graphs to identify patterns across Instagram

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

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