Identifying Participant Roles in Cyberbullying Through Hierarchical Attention Networks

Patrick Furman[1], Yasin Silva[1], Deborah Hall[2]

1: Loyola University Chicago, 2: Arizona State University

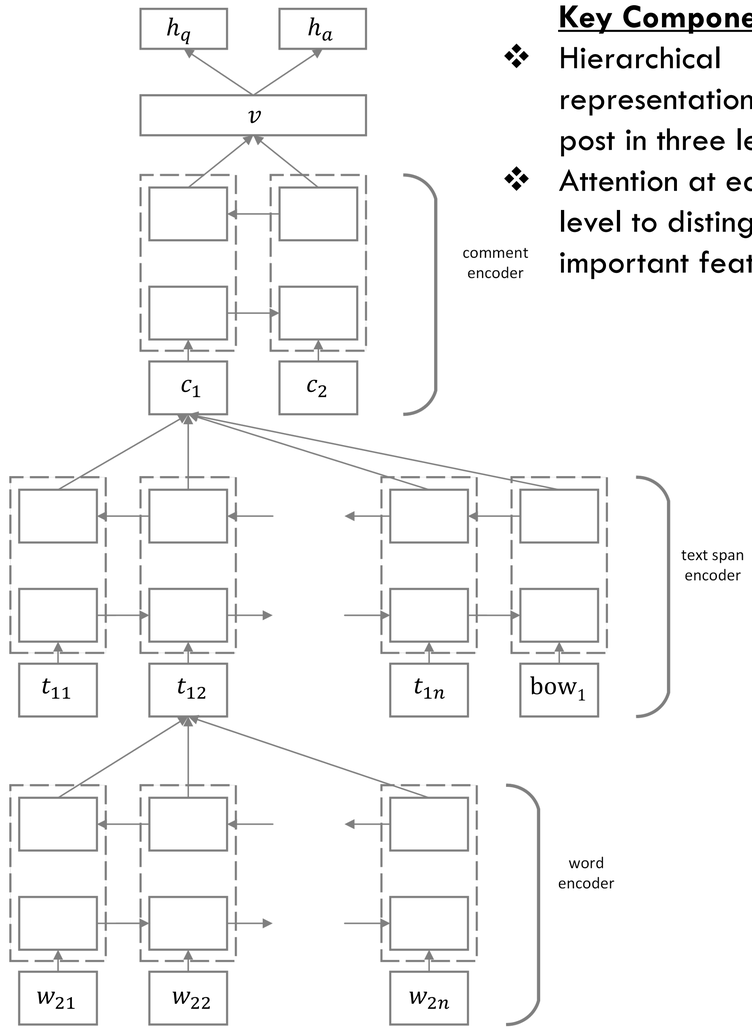
Abstract

Cyberbullying is a widespread form of online harassment with serious negative consequences for victims. In a cyberbullying instance, participants can be classified as harassers, victims, or bystanders. Identifying the roles of participants in cyberbullying instances can facilitate more effective intervention in these instances. We propose a hierarchical attention network to automatically classify the roles of users in cyberbullying conversations on ASKfm, a social media platform where users can ask and answer questions anonymously. Our model combines word, sub-sentence, and sentence-level attention mechanisms to represent the structure of posts on the ASKfm platform and capture relevant features for classification.

Methods

- Basic data preprocessing – removal of punctuation, tokenization, and padding shorter comments
- Word embeddings used from a word2vec model pretrained on Twitter data [1] to capture semantic meaning common to social media and fine-tuned on ASKfm dataset [2]
- Information for a post is broken down into three levels: comments, text spans, and words. These components are used to generate a post-level representation following a HAN framework [3]
- Participant roles are predicted simultaneously based on the post-level representation for the authors of both the question and answer comments

Model Framework



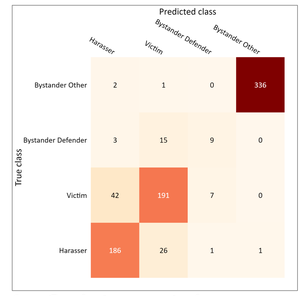
Key Components

- Hierarchical representation of a post in three levels
- Attention at each level to distinguish important features

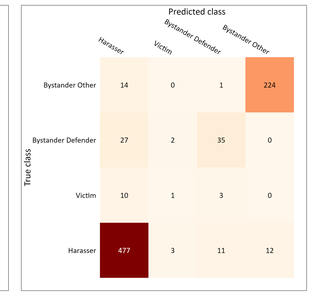
Preliminary Results

- An initial proof-of-concept model achieved 90.7% accuracy on the validation set but it showed some signs of overfitting due to limited data.
- The model performed very well at identifying harassers and unrelated bystanders, but was less confident identifying bystander defenders and victims, particularly within question comments

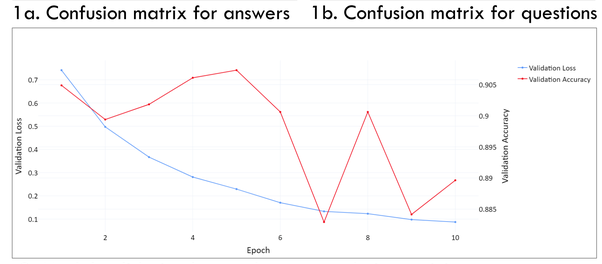
Class	F1 Scores for Questions	F1 Scores for Answers	Combined F1 Scores
Harasser	0.9253	0.8322	0.8972
Victim	0.1000	0.8076	0.7789
Bystander Defender	0.6140	0.4091	0.5570
Bystander Other	0.9432	0.9941	0.9731



1a. Confusion matrix for answers



1b. Confusion matrix for questions



2. Model validation loss and accuracy during training

Data Structure

- Posts for users are grouped into conversations each of which consists of one Q&A pair
  - Each comment is labeled with the degree of cyberbullying harmfulness (none, mild, severe) and for cyberbullying-related comments the role of the author (harasser, victim, bystander defender, bystander assistant) in the interaction is identified
- Each comment in a Q&A pair may contain any number of highlighted text spans (as shown below) identifying what type of language is used
  - Text spans can be labeled multiple times and may overlap

¶ [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]$^{\text{ATTACKING\_RELATIVES}}$:)

¶ 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.]$^{\text{THREAT\_BLACKMAIL}}$

Future Work

- Comparison with alternative transformer-based architectures not reliant on text span labels
- Multi-stage models to include general cyberbullying detection and text span labeling
- Conversation-level analysis to identify patterns across several posts for a single user

References

[1] Godin, Fréderic, et al. "Multimedia lab@ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations." Proceedings of the workshop on noisy user-generated text. 2015.

[2] Van Hee C., Verhoeven B., Lefever E., De Pauw G., Daelemans W. and Hoste V. (2015c). Guidelines for the Fine-Grained Analysis of Cyberbullying, version 1.0. Technical Report LT3 15-01, LT3, Language and Translation Technology Team–Ghent University.

[3] Cheng, Lu, et al. "Hierarchical Attention Networks for Cyberbullying Detection on the Instagram Social Network." Proceedings of the 2019 SIAM International Conference on Data Mining, 2019, pp. 235–243., https://doi.org/10.1137/1.9781611975673.27.

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

Acknowledgement

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