Understanding Cyberbullying on Instagram and Ask.fm via Social Role Detection

Hsien-Te Kao
University of Southern California,
Information Sciences Institute
hsientek@usc.edu

Shen Yan
University of Southern California,
Information Sciences Institute
shenyan@usc.edu

Di Huang
University of Southern California,
Information Sciences Institute
dh_599@usc.edu

Nathan Bartley
University of Southern California,
Information Sciences Institute
nbartley@usc.edu

Homa Hosseinmardi
University of Southern California,
Information Sciences Institute
homahoss@isi.edu

Emilio Ferrara
University of Southern California,
Information Sciences Institute
ferrarae@isi.edu

ABSTRACT

Cyberbullying is a major issue on online social platforms, and can have prolonged negative psychological impact on both the bullies and their targets. Users can be characterized by their involvement in cyberbullying according to different social roles including victim, bully, and victim supporter. In this work, we propose a social role detection framework to understand cyberbullying on online social platforms, and select a dataset that contains users' records on both *Instagram* and *Ask.fm* as a case study. We refine the traditional victim-bully framework by constructing a victim-bully-supporter network on Instagram. These social roles are automatically identified via ego comment networks and linguistic cues of comments. Additionally, we analyze the consistency of users' social role within *Instagram* and compare users' behaviors on *Ask.fm*. Our analysis reveals the inconsistency of social roles both within and across platforms, which suggests social roles in cyberbullying are not invariant by conversation, person, or social platform.

CCS CONCEPTS

• **Networks → Online social networks**; • **Human-centered computing → Empirical studies in collaborative and social computing**.

KEYWORDS

cyberbullying, social role, social networks

ACM Reference Format:
Hsien-Te Kao, Shen Yan, Di Huang, Nathan Bartley, Homa Hosseinmardi, and Emilio Ferrara. 2019. Understanding Cyberbullying on Instagram and Ask.fm via Social Role Detection. In *Companion Proceedings of the 2019 World Wide Web Conference (WWW '19 Companion), May 13–17, 2019, San Francisco, CA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3308560.3316505

1 INTRODUCTION

Cyberbullying is a major problem among teenagers on online social platforms. One in four middle and high school students reported being bullied at least once a week between 2015 and 2016 [6]. These students experienced both emotional (e.g., teasing, inappropriate comments) and social bullying (e.g., spreading rumors, embarrassing comments) on social media. Students who experienced cyberbullying are more likely to develop depression, anxiety, sleep difficulties, have a lower academic achievement, and drop out of school [3]. The bullies suffer as well: students who cyberbullied other students are at increased risk for mental health, behavioral problems, academic problems, and substance use [5]. Cyberbullying has prolonged negative impacts on both bullies and their targets, thus early detection is essential to minimize the potential harms.

There is extensive research on understanding individual behaviors and interactions during cyberbullying. Wegge *et al.* [19] showed victims tend to be cyberbullied by the same perpetrator who bully them offline. Festl and Quandt[7] found perpetrators in offline bullying are likely to bully other online users, whereas victims in offline bullying may not be victims in cyberbullying. Bastiaensens *et al.* [2] discovered bystanders will help the victims if the incident is severe, but they will also bully the victims when they are good friends with the perpetrator. These studies viewed cyberbullying as individual incidents rather than patterns of behaviors that might carry over between different incidents, different avenues of communication within a platform, and different platforms altogether. In addition, these studies focused heavily on victim-bully interactions that oversimplify cyberbullying interactions when there are other roles involved, like victim supporter.

To better understand individual behaviors on online social platforms, user interactions can be expressed by the networks and different interaction patterns can be decomposed into different social roles. Social role detection has been used to classify different types of user interactions. Akar and Mardikyan [1] reviewed various identified user roles on social platforms including Wikipedia, Reddit, and Twitter. Buntain and Golbeck [4] discovered the "answer-person" and "discussion-person" roles on Reddit and the consistency of these social roles across sub-communities of Reddit. Lumbreras *et al.* [13] developed a social role detection algorithm for coexistence roles in the community based on growth models for trees.

In this work, we aim to utilize social role detection to understand cyberbullying on a complex social network. We select a dataset [9]

WWW '19 Companion, May 13–17, 2019, San Francisco, CA, USA

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316505

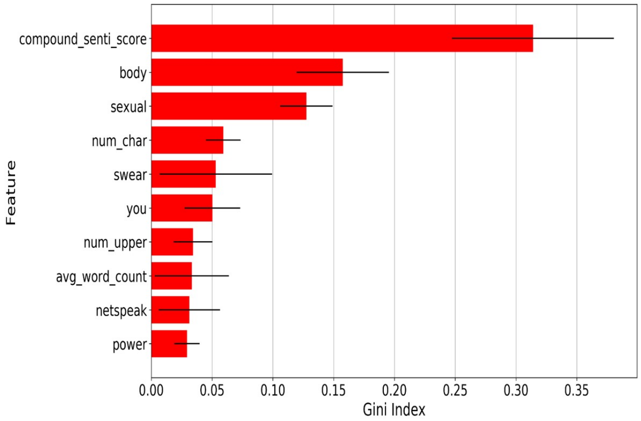

Figure 1: Top 10 important features.

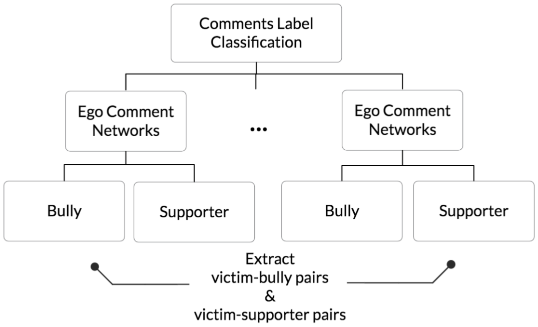

Figure 2: Pipeline of social role detection method.

that consists of users on such Instagram and Ask.fm to understand cyberbullying on online social platforms. Both platforms were ranked among the top five cyberbullying platforms in 2013 [18], and recent reporting shows Instagram continues to be a top platform [12]. In addition to the traditional victim-bully framework in cyberbullying, we further introduce the victim-supporter role based on bullying literature [17]. There are two types of supporting behavior, positive supporting comments that side with the victim and negative aggression toward the bullies, and our work focuses on the latter case. We propose a social role detection framework for victim-bully-supporter network that answers the following three questions: *(i)* whether the victim, bully, and supporter social roles exist on Instagram and Ask.fm, (ii) if it is possible to identify these roles in an automated fashion, and, (iii) whether a user's social roles and behaviors are consistent within and across online social platforms.

2 RELATED WORK

2.1 Social role detection

There are numerous studies about role detection on social platforms, such as Reddit. Buntain *et al.* [4] set out to identify the "answer-person" social role on Reddit in an automated fashion using Decision Trees for classification. To do this, they constructed ego-networks of users on a collection of subreddits and hand-labeled those users who conformed to the "answer-person" role they were looking for. Then, they used various network metrics as features for their classifier, whose performance suggested that network features may be very predictive for this specific role. The presented work differs in that it makes use of both linguistic and network features to examine multiple roles.

Similarly, Kou *et al.* [11] set out to study knowledge production about user experience work in the specific subreddit "/r/userexperience". They used a mixed methods approach consisting of statistical tests and qualitative content analysis to analyze the roles of the top users by activity in their dataset. This role analysis resulted in five distinct roles that gave insight into how learning and knowledge distribution happens within a community of user experience practitioners. The presented work differs methodologically in that roles are fixed and analyzed in a more automated fashion.

Another role detection study was done by Akar *et al.* [1] on the Turkish forum *incisozluk.com.tr*. The authors collected the data, constructed bipartite graph representations of the users and the topics they posted in, and projected that to look only at the users network. They then used the network properties of each member and a developed questionnaire to describe the different structural roles they discovered in the data. Importantly, they discovered that certain behaviors and perceptions seemed to be moderated by the social role of the user: for instance, perceived critical mass of a topic is a more prominent factor for visitors and passive members of a community, but not as much for content generators.

2.2 Cyberbullying detection

Detection of cyberbullying is a difficult problem that is hard to define as it manifests in different ways on different online social networks. Hosseinmardi *et al.* [9] acknowledged this and collected their dataset consisting of common users between Instagram and Ask.fm to characterize how much of an effect different platforms might have on how cyberbullying manifests. They performed various statistical tests on the number of positive and negative words found in posts across both platforms, as well as analysis of network-related features in their data. For instance, they found correlations between being positive on one network and being positive on the other network, and similarly for being negative. However, given limitations of their methods they do not label comments as bullying or not nor do they explore social roles contained in their data. Zhao *et al.* [21] proposed an automatic representation learning method for cyberbullying detection named embeddings enhanced Bag-of-Words model (EBoW). EBoW combines traditional linguistic features and bullying features based on insulting words with word embeddings. However, their detection method only considers the language properties but cannot capture the interactions between users.

3 DATASET

We use a dataset [8, 9] that contains user records on both Instagram and Ask.fm collected from August 2011 to June 2014. It consists of 14,063 users' information on Instagram and 8,696 of these users'

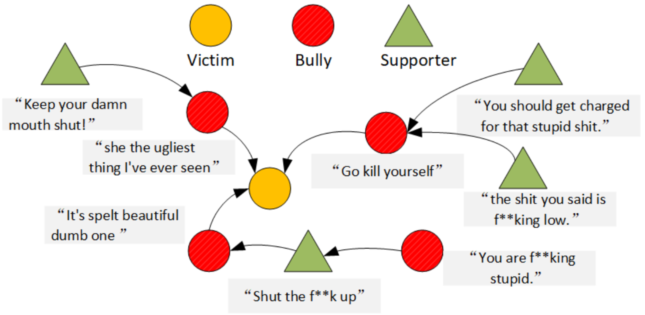
Victim
Bully
Supporter
“Keep your damn mouth shut!”
“she the ugliest thing I've ever seen”
“Go kill yourself”
“You should get charged for that stupid shit.”
“the shit you said is f**king low.”
“It's spelt beautiful dumb one ”
“Shut the f**k up”
“You are f**king stupid.”

Figure 3: Example of aggressive comments in a Instagram media session. Yellow – victim, red – bully, green – supporter.

Ask.fm information who also have an Ask.fm account. The Instagram data include 1,738,850 media sessions and 6,816,844 comments. These media sessions are in some form of media (e.g., video, picture) posted by the user to the user's account. The user has the opportunity to supply a caption to their own media when they submit the media, at which point it is either presented to the user's followers or the broader public if it is a public account. This allows for other users to comment on the media session. The Ask.fm data include 11,047,124 question-answer pairs. A user can send a question anonymously or non-anonymously to another user profile on Ask.fm. We separate the dataset into two sets of unique questions, one with 3,144,000 non-anonymous questions and one with 5,702,324 anonymous questions, and each question receives 1.7 answers on average.

4 AGGRESSIVE COMMENT CLASSIFIER

We labeled 13,310 Instagram comments as either aggressive or not in order to train a classifier to label the Instagram comments. We divided the comments into four different sets randomly, each of which was labeled by one of the four members of our team using the same set of instructions.

Each comment is given either a label 1 as aggressive or label 0 as not aggressive. In the end, there are about 12% comments labeled as aggressive and the rest are labeled as not aggressive. The labeled Instagram comments are chosen from top 200 media sessions ranked by the number of insulting words which appeared in the comments of the media session, the percentage of negative comments, and the absolute number of negative comments on the media session. The original insulting words were drawn from a aggressive words dictionary[1]. We expanded this dictionary with functionally similar terms from our corpus using the Gensim implementation of word2vec [14, 16]. We have a total of 739 insulting words after expansion. After this, we extract 83 linguistic features from each comment: 73 features from LIWC2015 categories [15], the compound sentiment score from VADER [10], and 7 basic linguistic features (e.g., the number of upper case words). We use LIWC2015 categories to provide a higher-level linguistic pattern extraction, the VADER compound sentiment score because it considers emojis and punctuation, and basic linguistic features to capture linguistic difference between comments. Considering the class imbalance and the feature properties, we train a binary random forest classifier with non-linear HSIC (Hilbert-Schmidt Independence Criterion) Lasso feature selection [20] and parameter tuning to label the remaining comments.

Table 1: Model performance for minority class (aggressive comment) with class imbalance 88-12 for nested 10-fold cross-validation.

Precision	Recall	F1 Score	False Positive
0.40	0.35	0.37	0.04

In Figure 1, we can see that our binary random forest classifier is labeling the Instagram comment as aggressive or not based on sentiment score, linguistic properties, and context. Compound sentiment score is the strongest feature because aggressive comments have higher negative sentiment score compared to non-aggressive comments. The classifier is labeling based on sexual, body, swear, and power words that target someone in addition to large sentence with a lot of upper case words and internet language. The selected word categories are often used in cyberbullying, and long sentence and upper case word are signs for strong negative emotion. In Table 1, we show the binary random forest classifier performance with accuracy 0.89. We apply the trained classification model to label all 6.8 million Instagram comments.

5 SOCIAL ROLE DETECTION

In this work, we focus our analysis on three different social roles in cyberbullying: victim, bully, and supporter. Figure 2 describes the pipeline of our social role detection method. Social roles are identified based on the aggressive label of comments and each user's ego comments networks. Figure 3 shows a typical example of a real bullying event in our dataset. Bullies posted aggressive comments to the author of the media session (victim), and supporters tried to defend the victim by attacking the bullies. Based on the observations above, we define users' social roles by their interaction patterns:

- **Victim:** users who receive aggressive comments repeatedly.
- **Bully:** users who send aggressive comments repeatedly.
- **Supporter:** users who send aggressive comments to bullies.

5.1 Ego network

In order to identify victim-bully and victim-supporter pairs, we build ego comment networks for each user. We only include the aggressive comments we classified from Section §4. An ego comment network $G_i(V_i, E_i, W_i)$ for user i is built based on all aggressive comments of i's media sessions (Figure 5 shows an ego comment network example for a particular user). V_i are all the users who commented on i's media sessions, including i. E_i represent the set of all edges e_{jk} between any user j and k, when there at least one aggressive comment from user j toward user k, and $w_{jk} \in W_i$ indicates number of media sessions with aggressive comments from user j to k, e.g., the edge $e_{jk} : (j \rightarrow k, 3)$ means that j has sent aggressive comments to i in 3 media sessions. In order to capture the recurrent nature of cyberbullying behaviors, any user j who has sent aggressive comments in more than one media sessions

[1] https://www.noswearing.com/dictionary

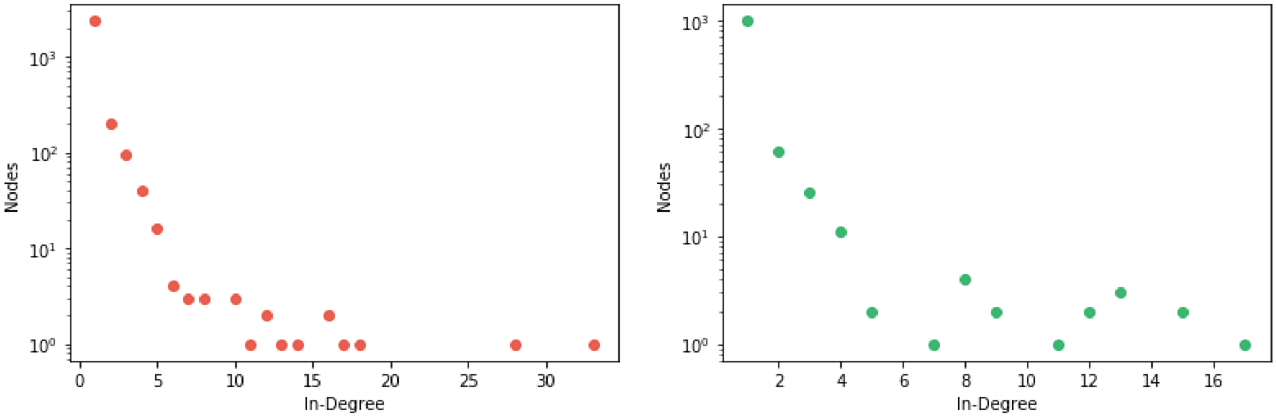

Figure 4: Number of nodes as function of in-degree for victim-bully/victim-supporter pair networks.

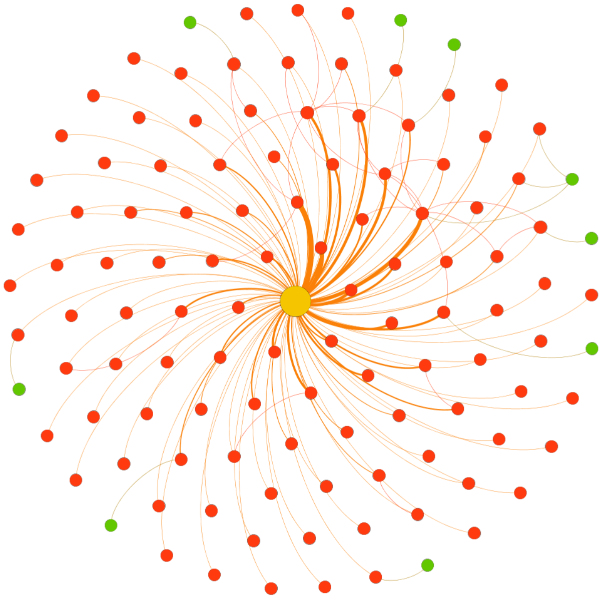

Figure 5: Example of ego comment network. yellow – victim, red – bully, green – supporter.

($w_{j,i} > 1$) towards user i, is considered as potential bully. The supporters are the users who have sent aggressive comments to the bullies identified above.

After filtering out the one-time aggressive behaviors, we identify 1,800 victim-bully pairs and 712 victim-supporter pairs from Instagram networks. Figure 4 shows the in-degree distribution of the victim-bully pair network, where in-degree values represent the number of bullies the user has. Comparing with the in-degree distribution in Figure 6, the number of bullies significantly decreased, which indicates that most bullying behaviors are temporal and occasional. Based on the victim-bully pair network, 13 users are bullied by more than 10 people, and the maximum number of bullies is 33. While in terms of the out-degree (i.e., number of users that he/she has bullied), only 2 users bullied more than one people repeatedly. Similar patterns also appear in the victim-supporter pair networks. Most users only has one supporter, and the users who has been identified as supporter only support one specific user. Only 9 users have more than 10 supporters.

5.2 Aggregated user network

Based on each ego bullying network, we assign users to one of the three social roles: victim, bully, or supporter. However, those individual-based social roles may differ across different interactions. In order to analyze a user's role consistency, we create an weighted aggregate victim-bully-supporter network $G(V, E, W)$ based on all aggressive comments on Instagram. Here, each node represents a user involved in Instagram media sessions. A directed edge e_{ij} from user i to user j represents there is at least one aggressive comment sent by i to j. Furthermore, we weight the directed link from i to j by counting the number of aggressive comments made by i towards j (i.e., w_{ij}). The case would either be that j was the owner of the media session where i commented or j replies to i in other users' media sessions.

Consequently, we generate a graph consisting of 441,523 nodes and 466,309 edges. Figure 6 shows the in-degree/out-degree distribution of the victim-bully-supporter network. The maximum in-degree value is 8,445, which means a user received aggressive comments from 8,445 users. The maximum out-degree value is 487, which means a user has sent aggressive comments to 487 users. For node i, if its in-degree d_i^{in} is much larger than its out-degree d_i^{out}, i is a global victim. When both d_i^{in} and the sum of the weights of incoming edges are larger, the bullying behaviors towards user i is more severe. On the other hand, if out-degree d_i^{out} is much larger than in-degree d_i^{in}, then user i is regarded as a global bully. Moreover, i could be a hater who bullied numerous users when

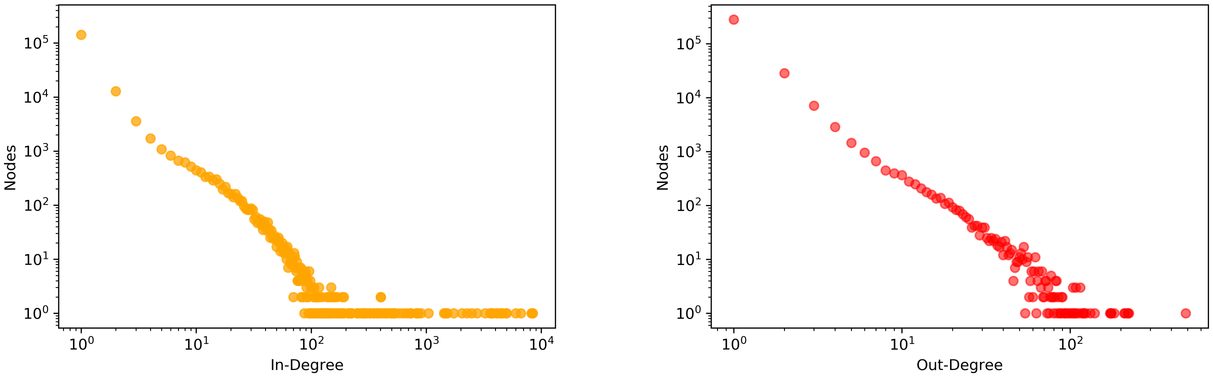

Figure 6: Number of nodes as function of in-degree/out-degree for victim-bully-supporter network. The in-degree value represents the number of bullies the user received; the out-degree value indicates the number of users he/she bullied before.

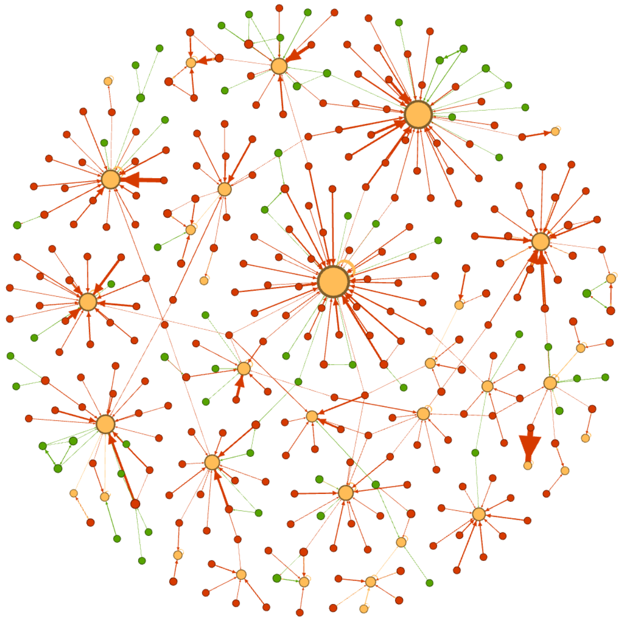

Figure 7: Victim-bully-supporter network (subgraph). Yellow – victim, red – bully, green – supporter.

d_i^{out} is relatively large. In addition, i is a global supporter only when his/her links were directed to global bullies.

By comparing the global social role and individual-based social roles for each user, we find that the three social roles have different consistency across media sessions. Figure 7 is a subgraph based on the largest component of the victim-bully-supporter network. The role assigned to a node is the majority of its individual-based roles through different media sessions. As Figure 7 shows, victims and bullies are more consistent between the roles identified based on ego networks and the aggregate network. However, some supporters also bullied victims on the aggregate network, which shows the inconsistency of the social role supporter across media sessions.

5.3 Consistency in Ask.fm

We rank the users on Instagram by the total number of aggressive comments that they posted on Instagram and look at these top-ranked users' Ask.fm data to examine how consistent the roles are between platforms. We select all the aggressive comments on Instagram that these users were involved in and identify the original author, the commenter, and anyone who is mentioned in the comment. From the top 1,000 ranked Instagram users, we identify 27 bully-victim pairs where the top-ranked user is the bully and the victim was either mentioned or had interaction with the bully on both platforms (61 pairs are present when we look at all the users). Interestingly, none of the roles in these pairs seemed to be consistent, as the identified bullies were always responding to their own media sessions on Instagram, suggesting they were inviting or expecting comments in their media sessions. The corresponding question/answer pairs on Ask.fm for these bully-victim pairs were all benign. The language in some of Instagram comments seems to suggest a playful aggression.

6 CONCLUSIONS AND FUTURE WORK

Cyberbullying has been growing on online social platforms and generates massive negative user experiences in addition to causing prolonged negative psychological impacts on both bully and their targets. Numerous researchers have studied cyberbullying in specific contexts (e.g., conversation, topic, relationship, social platform) using a constrained victim-bully framework. We here proposed a fine-grained social role detection method automatically discovering victim-bully-supporter interactions on online social platforms to understand cyberbullying. We chose a dataset that has the records from common users on both Instagram and Ask.fm as our case study. Comparing the consistency of users' social roles within Instagram and users' behaviors on Ask.fm, our work revealed that bullies and victims are more likely to have consistent social roles within Instagram, while supporters show more complex interaction patterns. Users' behavior patterns on Ask.fm also deviate from their social roles identified in Instagram. These observations show the consistency of social roles is specific to a conversion topic or person. Therefore, it is difficult to classify users' behaviors solely based on their social role in cyberbullying detected in a specific conversation, relationship, or social platform. This new insight is essential in designing safer online social ecosystems with adaptive cyberbullying social role detection.

This work mainly focused on the detection of pre-defined social roles in cyberbullying scenarios. For future work, we are considering to extend our methods to the discovering different behavior patterns as their social roles. We plan to leverage entity recognition and relation extraction techniques to discover the relationships between users from the comments' content, to enable capturing more complex behavior patterns and give us more insights into cyberbullying events.

REFERENCES

[1] Ezgi Akar and Sona Mardikyan. 2018. User Roles and Contribution Patterns in Online Communities: A Managerial Perspective. *SAGE Open* 8, 3 (2018), 2158244018794773.

[2] Sara Bastiaensens, Heidi Vandebosch, Karolien Poels, Katrien Van Cleemput, Ann Desmet, and Ilse De Bourdeaudhuij. 2014. Cyberbullying on social network sites. An experimental study into bystanders' behavioural intentions to help the victim or reinforce the bully. *Computers in Human Behavior* 31 (2014), 259–271.

[3] Sheri Bauman, Russell B Toomey, and Jenny L Walker. 2013. Associations among bullying, cyberbullying, and suicide in high school students. *Journal of Adolescence* 36, 2 (2013), 341–350.

[4] Cody Buntain and Jennifer Golbeck. 2014. Identifying social roles in reddit using network structure. In *Proceedings of the 23rd International Conference on World Wide Web*. ACM, 615–620.

[5] Centers for Disease Control and Prevention. 2017. Youth Risk Behavior Surveillance. https://www.cdc.gov/healthyyouth/data/yrbs/pdf/2017/ss6708.pdf. (2017).

[6] Melissa Diliberti, Michael Jackson, and Jana Kemp. 2017. Crime, Violence, Discipline, and Safety in US Public Schools: Findings from the School Survey on Crime and Safety: 2015-16. First Look. NCES 2017-122. *National Center for Education Statistics* (2017).

[7] Ruth Festl and Thorsten Quandt. 2013. Social relations and cyberbullying: The influence of individual and structural attributes on victimization and perpetration via the internet. *Human communication research* 39, 1 (2013), 101–126.

[8] Homa Hosseinmardi, Amir Ghasemianlangroodi, Richard Han, Qin Lv, and Shivakant Mishra. 2014. Towards understanding cyberbullying behavior in a semi-anonymous social network. In *Advances in Social Networks Analysis and Mining (ASONAM), 2014 IEEE/ACM International Conference on*. IEEE, 244–252.

[9] Homa Hosseinmardi, Shaosong Li, Zhili Yang, Qin Lv, Richard Han, Rahat Ibn Rafiq, and Shivakant Mishra. 2014. A comparison of common users across instagram and ask. fm to better understand cyberbullying. In *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*. IEEE, 355–362.

[10] C.J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International Conference on Weblogs and Social Media*. International Conference on Weblogs and Social Media, 216–225.

[11] Yubo Kou, Colin Gray, Austin Toombs, and Robin Adams. 2018. Knowledge Production and Social Roles in an Online Community of Emerging Occupation: A Study of User Experience Practitioners on Reddit. (2018).

[12] Taylor Lorenz. 2018. "Teens Are Being Bullied 'Constantly' on Instagram". *The Atlantic* (10 2018).

[13] Alberto Lumbreras, Bertrand Jouve, Julien Velcin, and Marie Guégan. 2017. Role detection in online forums based on growth models for trees. *Social Network Analysis and Mining* 7, 1 (2017), 49.

[14] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.

[15] James W Pennebaker, Ryan L Boyd, Kayla Jordan, and Kate Blackburn. 2015. The development and psychometric properties of LIWC2015. (2015).

[16] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50.

[17] Stopbullying. 2017. The Roles Kids Play in Bullying. https://www.stopbullying.gov/what-is-bullying/roles-kids-play/index.html. (2017).

[18] Ditch the Label. 2013. The Annual Cyberbullying Survey. (2013).

[19] Denis Wegge, Heidi Vandebosch, and Steven Eggermont. 2014. Who bullies whom online: A social network analysis of cyberbullying in a school context.

[20] Makoto Yamada, Wittawat Jitkrittum, Leonid Sigal, Eric P Xing, and Masashi Sugiyama. 2014. High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation* 26, 1 (2014), 185–207.

[21] Rui Zhao, Anna Zhou, and Kezhi Mao. 2016. Automatic detection of cyberbullying on social networks based on bullying features. In *Proceedings of the 17th international conference on distributed computing and networking*. ACM, 43.