Lightning Talk–Towards Robust Detection of Cyberbullying in Social Media

Mengfan Yao
University at Albany, State University of New York
Department of Computer Science
myao@albany

Charalampos Chelmis
University at Albany, State University of New York
Department of Computer Science
cchelmis@albany.edu

Daphney–Stavroula Zois
University at Albany, State University of New York
Electrical and Computer Engineering Department
dzois@albany.edu

ABSTRACT

The potentially detrimental effects of cyberbullying have led to the development of numerous automated, data–driven approaches, with emphasis on classification accuracy. Cyberbullying, as a form of abusive online behavior, although not well–defined, is a repetitive process, i.e., a sequence of aggressive messages sent from a bully to a victim over a period of time with the intent to harm the victim.

Existing work has focused on harassment (i.e., using profanity to classify toxic comments independently) as an indicator of cyberbullying, disregarding the repetitive nature of this harassing process. However, raising a cyberbullying alert immediately after an aggressive comment is detected can lead to a high number of false positives. At the same time, two key practical challenges remain unaddressed: (i) timeliness: the state–of–the–art relies on a fixed set of features learned during training for offline detection (i.e., after all correspondence has become available), hindering the ability to respond in a timely manner (i.e., as soon as possible) to cyberbullying events. (ii) scalabilty: the scalability of existing methods to the staggering rates at which content is generated (e.g., 95 million photos and videos are shared on Instagram per day[1]) has largely remained unaddressed.

In my lightning talk, I will introduce *CONcISE*, a novel approach for timely and accurate Cyberbullying detectiON on Instagram media SEssions, that has been accepted for presentation at the main conference [1]. Specifically, I will present a novel two–stage online approach (illustrated in Figure 1) designed to reduce the time to raise a cyberbullying alert by (i) sequentially examining comments as they become available over time, and (ii) minimizing the number of feature evaluations necessary for a decision to be made for each comment. By formalizing the problem as a sequential hypothesis testing problem, a novel algorithm has been developed that satisfies four key properties: *accuracy*, *repetitiveness*, *timeliness*, and *efficiency*.

Extensive experiments on a real–world Instagram dataset with ∼ 4M users and ∼ 10M comments demonstrate the effectiveness of the proposed approach with respect to accuracy, timeliness, efficiency, and robustness, and show that it consistently outperforms the stat–of–the–art, often by a considerable margin.

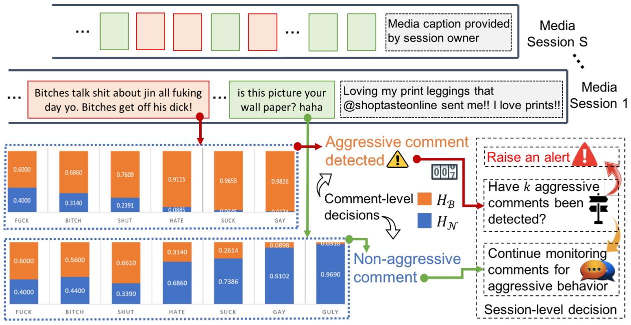

Figure 1: Overview of the proposed approach. Given a set of media objects along with their corresponding captions and hashtags, and set of comments, and an alert threshold, CONcISE examines comments as they become available over time and raises an alert only after the number of comment–level detections surpasses the threshold. The posterior probability evolution of an aggressive (upper) and non–aggressive (lower) comment as more features are examined is provided for illustration purposes. Note that the number of features used to make a decision in each case differs.

CCS CONCEPTS

• **Information systems** → *Collaborative and social computing systems and tools*; *Social networking sites*; *Data mining*; *World Wide Web*; *Social networks*; • **Computing methodologies** → *Machine learning*.

KEYWORDS

Classification; cyberharassment; optimization; sequential selection; social networks

ACM Reference Format:
Mengfan Yao, Charalampos Chelmis, and Daphney–Stavroula Zois. 2019. Lightning Talk–Towards Robust Detection of Cyberbullying in Social Media. 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, 2 pages. https://doi.org/10.1145/3308560.3316474

[1] 33 Mind–Boggling Instagram Stats & Facts for 2018: https://www.wordstream.com/blog/ws/2017/04/20/instagram-statistics

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

ACM ISBN 978-1-4503-6675-5/19/05.
https://doi.org/10.1145/3308560.3316474

REFERENCES

[1] Mengfan Yao, Charalampos Chelmis, and Daphney-Stavroula Zois. 2019. Cyberbullying Ends Here: Towards Robust Detection of Cyberbullying in Social Media. In *Proceedings of the 2019 World Wide Web Conference (WWW '19)*. International World Wide Web Conferences Steering Committee, 7 pages. https://doi.org/10.1145/3308558.3313462