Prevention of Cyberbullying Using Machine Learning: A Review

TWLA Perera#, TL Weerawardane, RGC Upeksha

Department of Computer Science, Faculty of Computing General Sir John Kotelawala Defence University, Sri Lanka

Abstract. With the advancement of technology, social media has grown to be widely popular. Although social media platforms provide excellent opportunities, they can pose a negative impact on users. Cyberbullying is one such adverse phenomenon which can occur online through social media, forums, or games where users can read, interact with, or exchange content. Cyberbullying and online harassment can impair people's lives causing victims mental and physical distress and often leading to extremes such as suicide. Social media platforms have provided users with options such as flagging, blocking, or reporting the bullies fostering a safer online community. But due to the alarming amount of reported content and users daily, there is a need for automated, data-driven methods in detecting and preventing such harmful activities in social media. It will help foresee potentially dangerous and harmful situations and prevent them from occurring. Machine Learning based approach is used in most existing systems to tackle this problem. There exist several knowledge gaps in detecting and preventing cyberbullying on social media such as the effectiveness of current automated prevention and intervention methods and the impact of anonymity which can make it difficult to identify and hold bullies accountable. This study is a systematic literature review where similar existing systems are explored, examined, and analyzed. The purpose of the study is to identify and examine the features, methods, and limitations present in the systems which will help discover a novel solution to mitigate the adverse effects on people. It was identified that the utilized algorithms in applications have performed in various ways depending on the specific characteristics of the dataset and the problem at hand. Furthermore, most research has been done on text-based hate speech detection and is considering combining textual data with video and images as future work.

Keywords: *Cyberbullying, Machine Learning, social media, social networks*