Detection of Anti-social Behaviour in Online Communities

Martin BORÁK[*]

Slovak University of Technology in Bratislava
Faculty of Informatics and Information Technologies
Ilkovičova 2, 842 16 Bratislava, Slovakia
xborakm@stuba.sk

During the past decade, the Web has changed dramatically. It is no longer just about finding information or private and work related communication. Online communities have been gaining on importance and popularity in places like social networks, CQA (Community Question Answering) systems, online games and news and entertainment portals. Immediate communication with an unlimited number of users on a variety of topics has become part of everyday life for hundreds of millions of people worldwide.

Due to the high number of members of these communities, the content of such communications is rather diverse. From sophisticated, evidence-based arguments on serious political issues in discussion sections of news portals, to informal discussions that are often useless, in comments below YouTube videos. However, there are people everywhere, who try to undermine the course of these communications at any cost. Whether it is by writing meaningless messages, share links to pages that have nothing to do with the discussed topic or unnecessary sarcasm, direct aggressiveness and harsh verbal attacks. Such behaviour is most widespread in sites, where users have the opportunity to act anonymously (YouTube, forums, etc.), but it has recently spread to systems where the contribution is associated with the real name of the author, possibly even a photograph and other personal information (e.g. Facebook). That causes content with higher quality to be surrounded by low quality content and thus makes it less likely to be read by other users. Sites often set policy against such content, since they have a legal responsibility for all content that can be viewed on their pages.

In our work we focus on a very frequent type of anti-social behaviour called "trolling". Trolling has various definitions. Trolls are people who participate in negatively marked online behaviour. They are also referred to as creatures who take pleasure in annoying other people and cause them anger and suffering [1]. They often initially pretend to be regular members of community, but later they try to disrupt it. The identification of trolls is a challenging task because many posts made by users who have

[*] Supervisor: Ivan Srba, Institute of Informatics, Information Systems and Software Engineering

less experience in online communication, or those possibly suffering from mental illness, show signs of troll posts, but without any intention to harm anyone [2].

A few dedicated researchers have already made contributions towards automated detection of anti-social behaviour in the past. Dinakar et al. [3] attempted to detect anti-social behaviour in YouTube comments. Specifically, they focused on comments with offensive or humiliating nature. For classification of such comments they used analysis of textual features and SVM (Support Vector Machines) algorithm. Dadvar et al. [4] focused on a similar thing over a similar dataset from YouTube, but for detection they also used users' history (e.g. previous comments, reputation), what helped them in achieving better results. That gives us a perspective of improving their work even more, if we were able to find other significant features. We have already contacted the authors of both articles, in order to obtain their datasets that we would use in our work.

Our goal is to be able to detect the content created by trolls. Text analysis is not sufficient for us, since trolls are known to often use sarcasm and act like ordinary users in order to provoke discussion, but simultaneously they try to cause conflicts. Therefore, in our work we also want to use sentiment analysis, user history and the analysis of community response. Based on these features and by using an appropriate machine learning algorithm, a model will be developed, which will be able to detect posts created by trolls. We believe that the result of our work has great potential to be useful for sites with user generated content, which could facilitate the work of moderators, who are currently responsible for filtering inappropriate content manually.

References

[1] Cheng, J., Danescu-Niculescu-Mizil, C. & Leskovec, J., Antisocial Behavior in Online Discussion Communities. *Proceedings of the Ninth International Conference on Web and Social Media, 2015, University of Oxford, Oxford, UK, May 26-29, 2015*, 2015, pp. 61-70.

[2] Hardaker, C., Trolling in asynchronous computer-mediated communication: From user discussions to academic definitions. *Journal of Politeness Research. Language, Behaviour, Culture*, 6(2), 2010, pp. 215-242.

[3] Dinakar, K., Reichart, R. & Lieberman, H., Modeling the Detection of Textual Cyberbullying. *Association for the Advancement of Artificial Intelligence*, 2011, pp. 11-17.

[4] Dadvar, M. et al., Improving Cyberbullying Detection with User Context. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2013, pp. 693-696.