CSE 291: Anti Social Computing - WI22
Final Project Report
Group-1: Strategized Counter-Trolling

INTRODUCTION

Trolling is considered one of the most disruptive, unpleasant, and toxic online antisocial behaviors. Many studies have revealed that trolling has a profound negative influence on physical, mental, emotional, and social well-being. [1] Although some people view trolling as light humor or fun online activity [2], in many cases trolling can inevitably lead to aggressive, violent, and inappropriate online content. Our project aims to provide an innovative feature that could be integrated into the comment section of all social media platforms to help the victims counter-troll the trolls, in an effort to mellow down the toxicity before it takes off. We believe that such an approach will help the victims who want to troll their trolls, in a light humorous way, and also bring down the intensity of the initial troller's comment.

BACKGROUND

A study into the psychology of malicious trolling revealed that trolls are motivated by the desire to inflict anger, hurt, or humiliation on their victims [3]. An article about the different ways to respond to online trolling considers humor as a key strategy because it could throw the trolls off their game [4]. We observed that many Twitter accounts used counter-trolling humor to actively respond to their trolls while maintaining respect, and harmony. To date, there has not been much work in literature considering the affordances of a counter-trolling option in social media platforms. To further understand if people would accept such a feature, we conducted an online survey [5], 18 respondents out of 22 agreed that they think a support system is needed to tackle trolling. 95% also agreed with the observation that online trolling has become more offensive in recent years. This has led us to believe that counter-trolling, with light humor, could be an effective approach to weigh down hurtful trolling comments before they take off. This approach also does not violate the goals of free speech and expression. We design a useful feature that would allow users to select a strategy for their response to trolling comments.

IMPLEMENTATION

The implementation involves a full-stack application including a user-facing interface (implemented using React.js) that mimics the comment section of an online social media platform (for instance, Reddit). Users/authors input some comments relating to some topic, and the users engaging with that comment (in the form of replies) can choose to use the counter-troll feature to respond to trolls. Given the user selects the counter-troll feature to respond to a troll, they choose one of the following strategies that shall determine the form of their response:

- Response in the form of a Joke
- Response in the form of a Meme
- Response in the form of a Quote
- Response in the form of the phrase – 'TOXIC TOXIC TOXIC'

Once the user selects a strategy to use for a response, a Flask API call is made to the back-end system (implemented using python) which predicts if the troll's post/comment is toxic and generates a response from a static database of Joke(s), Meme(s), and Quote(s) that has been mined from several online sources (brainyquote, goodreads, quotestats, etc.). For prediction of an input comment as toxic, Google's Perspective API is consumed, which uses machine learning to produce a score based on the perceived impact that the phrase may have in an active conversation. The scores produced by the API are provided for several different attributes including severe toxicity, insult, profanity, etc. Internally, the API uses a multilingual BERT model that is trained on the comments from several online sources (Wikipedia/The New York Times for instance). In case a user selects some strategy to counter-troll a comment/post which isn't toxic, the feature returns a phrase stating 'NOT TOXIC!' against that comment/post. This ensures that the feature isn't misused in a way that could induce toxic content online.

CONCLUSION

We believe that counter-trolling, using humor, will deflect troll behaviors online, preventing it from gaining traction and aggravating the toxic content. Users of a social networking platform who aren't confident in tackling online trolls can use our counter-troll comment feature to support themselves. We believe this will create a welcoming space in social platforms, promoting non-toxic content online and allowing all users to be expressive.

FUTURE SCOPE

- For the scope of this project, the input comments are assumed to have been typed in English, but the Perspective API provides multilingual support and hence the counter-troll feature can be extended to support non-English languages as well.
- Presently the backend script uses the Perspective API such that the scores/predictions produced by the API are provided only for the 'severe-toxicity' attribute. However, the API can predict scores for several different attributes including insult, profanity, identity attack, flirtation, etc., and hence the counter-troll feature can be extended to incorporate those attributes as well.
- The counter-troll feature presently accepts only text as an input, i.e., if a user tries to troll someone with a meme or an image, the feature is incapable of extracting text from images at present, and hence in addition to the Perspective API, the backend script can be extended to include functionality for extracting text from images.
- From the view of false negatives, if a user finds a comment/troll as toxic but the API fails to predict it as toxic, the Perspective API offers a 'SuggestCommentScore' method, where a user can suggest a better score for a comment. The counter-troll feature can incorporate this method to handle false negatives.

REFERENCES

[1] https://www.fitday.com/fitness-articles/fitness/how-online-trolling-affects-your-health.html

[2] Phillips, Whitney. This is why we can't have nice things: Mapping the relationship between online trolling and mainstream culture. Mit Press, 2015.

[3] Buckels, Erin E., et al. "Internet trolling and everyday sadism: Parallel effects on pain perception and moral judgment." *Journal of personality* 87.2 (2019): 328-340.

[4] https://www.searchenginejournal.com/defeat-online-trolls/323439/#close

[5] Link to our survey questions

[6] Github link: https://github.com/techno96/Trobo