Toxic Tweet Classification with Natural Language Processing and Machine Learning Techniques

(Technical Paper)

The Impact of Cyberbullying and Social Harassment on Video Game Streamers

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering

Tanapol Kosolwattana
Fall 2019

On my honor as a University Student, I have neither given nor received unauthorized aid on this assignment as defined by the Honor Guidelines for Thesis-Related Assignments

Signature _____ Date 12/5/2019
Tanapol Kosolwattana

Approved _____ Date 12/06/2019
Rich Nguyen, Department of Computer Science

Approved _____ Date 12/11/2019
Kent Wayland, Department of Engineering and Society

General Research Problem

How can artificial intelligence create a hate-free online community on social media?

In the online world, social platforms are created to serve the freedom of users to post messages or share their thoughts regardless of their gender, age, or race and to make healthy, hate-free online communities (Dhakal, 2019). However, when a conflict between users emerges, the posts are filled up with negative words or phrases, often leading to cyber harassment, such as cyberbullying or hate speech, and different platforms deal with this problem in different ways. Therefore, there should be some discourse analysis to understand the cause of conflicts and how different streaming or social media platforms handle the situation.

Recent research has experimented with the methods of classifying toxic contents using Information Retrieval techniques (Gaydhani, 2018). However, these methods did not take into account the sentiment analysis between words, which might result in lacking insights about the wording relationship. In my technical research, I will create the alternative methods for classifying toxic content by implementing the natural language processing toolkit called GloVe: Global Vector for word Representation (Pennington, 2014) to perform sentiment analysis. Then, I will use various machine learning models to classify the tweet datasets. After the training process, I will compare the performance of each model to see which model has the best accuracy in classifying tweets.

A growing social problem called “Cyberbullying” becomes a common problem in online communities since cyberbullying takes place on any online channels where people share, view, or participate in the content (What Is Cyberbullying, 2012). One of the online communities that are affected most by these actions is the gaming community. According to the data from the Cyberbullying Research Center, 11% of gamers are victims of online harassment compared to

8% for non-gamers (Petrov, 2019). Recently, there was an incident in which a Japanese streamer got harassed by a group of Thai viewers who spread lots of obscene and inappropriate words on her streaming chat. Even though she posted on Twitter about how these harassments damaged her emotionally and psychologically (core63_mc, 2019), there were still no actions from the streaming platform against these viewers. Building on this example, the STS research will focus on the impact of cyberbullying on gaming, especially the harassment on game streaming platforms and will identify how the platforms handle the situation regarding the existing policies, counter-measures, and the details of actions. I will also conduct some survey experiments to interview UVA students about how they experience a situation regarding cyberbullying and how they handle it when they play the game. In the end, I will propose some recommendations to determine which terms and policies should be adopted so that the game streaming platforms will create a better streaming experience for streamers and viewers and improve the quality of cyberbullying protection on the system.

Toxic Tweet Classification with Natural Language Processing and Machines Learning Techniques

How can the effective toxic tweet classification be created using natural language processing and machine learning techniques?

In the technical research, I will investigate the category of toxic tweets on Twitter to understand how online communities should benefit from creating policies against hostility in tweets. In the data cleaning step, I will convert each tweet dataset into the vector that contains only the essential content of the tweet. (i.e., I will remove the stop words and the punctuation on the tweet.) Then, for the preprocessing step, I will apply the lemmatization and stemming process to generate the root for inflected words so that the words are unified as one root for the

meaning. For example, after lemmatizing and stemming the set of words “Playing”, “Plays”, and “Played”, the program will generate the root “Play” (Jabeen, 2018). I will integrate the word embedding toolkit called The Global Vector for word Representation (GloVe) to find the relationship between words to create vector representations. Then, I will use the result of GloVe embedding to convert each tweet into a feature vector and classify them using various machine learning classifiers, including the Support Vector Model (SVM) with linear kernel, and Decision Tree Classifier to label which category that each particular tweet belongs to. After each training session, I will tune some parameters in each classifier so that it will give the best performance in each training iteration.

After the classification process, I will compare a performance from each classifier to see which classifier obtains the best accuracy for tweet classification. Also, I will interpret the result from the classification to order which hostile tweets appear most on the Twitter community so that it will help Twitter to create more effective policies against this type of abusive tweets.

The Impact of Cyberbullying and Social Harassment on Video Game Streamers

How do different game streaming platforms handle incidents of cyberbullying and social harassment in video game streaming communities?

Introduction

The gaming community is one of the places where cyberbullying frequently occurs since the anonymity of players or viewers can disguise their harmful behaviors and actions toward the streamers. Now, there are numerous tactics employed by online viewers to harass streamers, including chatting with abusive language, cyber harassment, sexual harassment of female gamers, objectification, game phishing scams, toxicity, and impersonation. Sometimes, it is hard

to differentiate whether the chat is a troll or just a funny one, which might lead to miscommunication between streamers and users. These problems not only create deleterious impression toward new and existing users, but also make negative impacts which lead to a psychological and emotional wound on people in the gaming community. Therefore, I will investigate how game streaming platforms receive the cases and how they handle them.

Background and Theoretical Framework

There are several essential vectors for cyberbullying and harassment. Cell phones are a conventional means by which teenagers are cyberbullied. According to the data from the Pew Research Center, cell phone usage has been increasing since 2004 by 45%, and 75% of 12 to 17-year-olds teenagers owned cell phones (Donegan, 2012). Even though many parents purchased the phone for their children for protective purposes, they did not realize that this device could potentially be utilized as the tool for cyberbullying (Donegan, 2012). Also, when the internet came into the online world, user anonymity allowed people to communicate with each other besides face-to-face interaction. It could cause some online users to abuse this ability to perform cyberbullying and perform negative actions against the victims, such as threats, insults, and trolling. Therefore, instead of being hate-free and healthy communities, online communities, especially social media, became loaded with hate speech and social harassment.

In recent years, streaming gameplay has become very popular in the online community since it is regarded as an entertainment platform that allows the audiences to engage with their favorite games (Edge, 2013). With the increase of streamers and viewers, sometimes there are chances that viewers have different gameplay preferences and the personalities which might not match the ones on streamers. Therefore, the viewers' opinions on live streaming are very diverse; some of them are positive or neutral while the rest contain negative feelings to the streamers.

These negative formations of viewers' opinions can be in terms of trolling or verbal harassment with abusive language, which hurts the emotion of streamers.

In STS research, the primary groups are streamers, streaming service providers, and viewers of streaming content. I will investigate some specific cases of harassment that are related to game streaming and understand the impact of this harassment on victims and how each platform handles the situation. For the theoretical framework, I will create an evaluation rubric for each game streaming platform's policy by analyzing the effectiveness of the actions and tools that it takes against perpetrators.

Evidence/ Data Collection

I will be studying the cases on several gaming platforms, including Youtube Gaming, Twitch, and Mixer. I will look for several cases where streamers have been consistently harassed by groups of watchers or online haters. The example of the data collection process is the case of Tanya DePass, a content creator on Twitch. She has been consistently harassed and stalked by Twitch users (Grayson, 2019). The data collection will be from Grayson's (2019) interview with Tanya DePass and the media commentary, which narrates the incident of the case. Then, in the content analyzing a process, I will apply these data to understand how Twitch uses its technology to solve her case and her feeling after applying it to her channel. In other cases, I will gather various types of sources such as gaming streams with harassment, documents about platforms' response, media commentaries about incidents, and social groups' responses on discussion forums, particularly from viewers and streamers to analyze narrative data using content or discourse analysis. I will also look at a website on each game streaming platform to get the information about platforms' policies. Then, I will investigate how they handle harassment using policies, counter-measures, or actual practice. I will make a comparison of the policies of

different platforms based on the result of their actions against harassers. This comparison will also use responses from streamers who are victims of harassment as another criterion to demonstrate whether actions are successful in banning harassing viewers and recovering victims' emotions. A question will arise to evaluate whether the policy has some positive impact: What reactions do different groups have regarding the response from the platforms? To fill in this gap, I will create a score-based survey to ask UVA students who are currently active in the game streaming community if they agree to adopt this technology to the social platform. The responses from a survey will also determine the quality of platforms' anti-harassment policies and tools and their efficacy, whether they can be applied to various groups of people who are active in the game streaming community. I will need to get approval from the Institutional Review Board, UVA's board for governing the ethics of human subject research, to conduct the survey. The result of the investigation will be in the form of recommendations for policy adoption or technology implementation, which aims to benefit the platform in tackling the attackers in cyberbullying cases.

Analysis

For each case, I will look at what technology or policy that the platform uses to take action against cyberbullying. For instance, Twitch recently announced new tools to help reduce harassment called AutoMod (The Cybersmile Foundation, 2019). According to the article, developers believe that with the combination of humans and machine learning into this feature, streamers can build the baseline of the appropriate language, which will make the platform a safer place for them to stream (The Cybersmile Foundation, 2019). After the product was launched, there have been some reports that users are satisfied. For example, Little Siha, a streamer with 36,000 followers, and her moderators struggled to clean the chat when she was

regularly called with inappropriate words. When she used the AutoMod, it helped her catch many trigger words even when they were intentionally misspelled to avoid the ban (D'Anastasio, 2016). However, I will also seek out the cases in which AutoMod was effective in ensuring a balanced view of the tool. For example, even though AutoMod can detect common negative words, it may not consider the emotes or the memes as the in-jokes that create the offensive meaning (Lo, 2018). At the end of the evaluation, the technology and policy will be assessed whether other platforms should adopt this implementation or not. In this case, I may conclude that the implementation of the AutoMod can create a safer environment for streaming. However, the social platform which integrates this feature needs to make sure that this technology can solve a problem as much as it can. Other technologies or policy assessments will also be conducted in the same way.

Conclusion

Even though both technical and STS research establish ways to mitigate toxicity on social media, there are still difficulties in controlling trolling and harassment on social media. For the future work, while the STS research focus is based on the victims' perspective of how the cyberbullying affects their life, the further research experiment can be extended in the harassers' perspective which is in term of behavioral studies or harassers' characteristic development studies to understand how the cyberbullying influences their actions and how the community prevents it.

References

core-こあ-. (2019, October 1). To the Thai who gave me a reply. You don't have to feel responsible for bad people because you are Thai. I don't really care about race, such as Thai and Japanese. → [Tweet]. Retrieved December 8, 2019, from @core63_mc website: https://twitter.com/core63_mc/status/1179059713406726145

Cybersmile Foundation (n.d.). Twitch Announce New Tool To Help Reduce Harassment And Abuse – Cybersmile. Retrieved December 8, 2019, from https://www.cybersmile.org/news/twitch-announce-new-tool-to-help-reduce-harassment-and-abuse

D'Anastasio, C. (2016, December 14). Twitch's AutoMod Is Already A Game-Changer, Streamers Say. Retrieved December 8, 2019, from Kotaku website: https://kotaku.com/twitch-s-automod-is-already-a-game-changer-streamers-s-1790109323

Dhakal, A., Kedia, D. (2019). Toxicity Classification On Social Media Platforms. Poster presented in CS 224, Stanford University. Retrieved from https://web.stanford.edu/class/cs224n/posters/15722323.pdf

Donegan, R. (2012). Bullying and Cyberbullying: History, Statistics, Law, Prevention and Analysis. *The Elon Journal of Undergraduate Research in Communications, 3*(1), 33-42. Retrieved from https://www.elon.edu/u/academics/communications/journal/wp-content/uploads/sites/153/2017/06/04DoneganEJSpring12.pdf

Edge, N. (2013). Evolution of the Gaming Experience: Live Video Streaming and the Emergence of a New Web Community. *Elon Journal of Undergraduate Research in Communications*, *4*(2). Retrieved from

http://www.inquiriesjournal.com/articles/821/evolution-of-the-gaming-experience-live-video-streaming-and-the-emergence-of-a-new-web-community

ElSherief, M., Nilizadeh, S., Nguyen, D., Vigna, G., & Belding, E. (2018). Peer to Peer Hate: Hate Speech Instigators and Their Targets. *Twelfth International AAAI Conference on Web and Social Media*. Presented at the Twelfth International AAAI Conference on Web and Social Media. Retrieved from https://www.aaai.org/ocs/index.php/ICWSM/ICWSM18/paper/view/17905

Gaydhani, A., Doma, V., Kendre, S., & Bhagwat, L. (2018). Detecting Hate Speech and Offensive Language on Twitter using Machine Learning: An N-gram and TFIDF based Approach. *ArXiv:1809.08651 [Cs]*. Retrieved from http://arxiv.org/abs/1809.08651

Grayson, N. (2019, January 29). For Streamers Dealing With Stalkers, Twitch's Solutions Fall Short. Retrieved December 8, 2019, from Kotaku website: https://kotaku.com/for-streamers-dealing-with-stalkers-twitchs-solutions-1832063386

Jabeen, H. (2018, October 23). Stemming and Lemmatization in Python. Retrieved December 8, 2019, from DataCamp Community website: https://www.datacamp.com/community/tutorials/stemming-lemmatization-python

Lo, C. (2018). *When All You Have is a Banhammer: The Social and Communicative Work of Volunteer Moderators*. 86. Retrieved from https://cmsw.mit.edu/wp/wp-content/uploads/2018/05/Claudia-Lo-When-All-You-Have-Is-a-Banhammer.pdf

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Petrov, C. (2019, December 28). The Latest Cyberbullying Statistics You Should Know In 2019. Retrieved December 8, 2019, from Tech Jury website: https://techjury.net/stats-about/cyberbullying/

What Is Cyberbullying. (2012, March 7). Retrieved December 8, 2019, from StopBullying.gov website: https://www.stopbullying.gov/cyberbullying/what-is-it/index.html