PREDICTING ESCALATION: PROTECTING ANALYSTS PROGRESS REPORT

Jordan Arnold, Canadian Center for Identity Based Conflict

Purpose of Research

An argument can be made that hateful rhetoric and group versus group conflict has increased in Canada in part due to the lack of identification in the Criminal Code of identity-based soft violence, thereby inadvertently providing perpetrators with the incentive to continue with their activities unpunished (Meyers, 2019). Kelshall has defined these unrecognized acts of hate as soft violence, which includes "actions that stop short of criminally identified violence…and highlight the superiority of one group over another without kinetic impact" (as cited in Kelshall & Meyers, 2019, p. 40). The damage created by soft violence is incalculable, as its harmful effects range from instilling fear within the individual victim or targeted identity-based group, to the polarization of society that damages cohesion within the general public (Kelshall & Meyers, 2019). Furthermore, it might be useful to consider the damage of soft violence on researchers of such content. Therefore, the Predicting Escalation research project first focuses on the impact of analyzing hateful content on the researchers themselves. The following progress report outlines the supportive literature, research challenges, and research findings that have been collated thus far.

Literature Review

Social media and internet platforms are increasingly being used by non-state actors as a tool for radicalization (Perry & Scrivens, 2019; Koehler, 2014, Leuprecht et al., 2010; Neumann, 2013). As a result, researchers that analyze these non-state actors over a long period of time become exposed to hateful rhetoric, and violent psychologically challenging content. The impact on the researcher may be exacerbated by the issue of confidentiality, as researchers may be limited to whom they can discuss the nature of their research with without breaching confidentiality (McCosker, Barnard & Gerber, 2001). Furthermore, some researchers are not protected by anonymity, and may receive harassment and death threats, unlike content moderators of big social media platforms (Martineu, 2019, para. 25). The literature suggests that risks associated with prolonged and consistent exposure to violent content, hateful rhetoric, and psychologically challenging content leads to desensitization, psychological distress, and/or potentially aggressive cognitions/behaviours.

Gibbons (2012) states that similar to military combatants and deployed military healthcare providers may be at increased risk of negative psychological side effects as a result of exposure to "life-threatening situations," including PTSD (p. 3). The study found that stress responses may differ depending on one's gender (p. 18-19). While Gibbons' study focuses on deployed military healthcare providers who were previously deployed, it does highlight the past neglect in literature of the mental and psychological effects on those who may not be directly involved in conflict. Both McCosker et al. (2001) and Martineu's (2019) work suggests that researchers may also be at risk of adverse side effects of dealing with sensitive research that involves violent, psychologically challenging, or hateful content (para. 4). McCosker et al. (2001) states that there are physical and psychological dangers to researchers in that researchers, those that they interview, and even their immediate families may be in danger (para. 9). Additionally, the psychological effect may come in the form of physical response, or "have a psychological impact" (para. 10). Martineu (2019) states that the suggested solutions to the increased spread of disinformation and online extremism rely too heavily on online platforms' abilities to self-police, or place too much responsibility on the user for their own media literacy (para. 11). Moreover, Martineu (2019) also suggests that some researchers of online extremism may face deterioration of mental health with similar symptoms to PTSD (para. 25).

Rogers' (2016) reports on the effect of exposure to a constant cycle of violent news on individuals of various ages and genders, suggesting it causes adverse psychological or emotional effects. According to psychologist Anita Gadhia Smith, an increase in the frequency of shooting or terror attacks builds up a sense of anxiety, vulnerability, and powerlessness and that it leads to heightened alarm and desensitization (Rogers, 2016, para. 6-8). Furthermore, Rogers (2016) explores the effects of exposure to violent images on social media, also stating that the effects can be traumatic, with similar symptoms to PTSD (para. 9). Additionally, the study found that participants that viewed violent images more often were more affected than participants who saw them less frequently (Rogers, 2016, para. 12). Finally, participants who self-described themselves as extroverts were at higher risk to be disturbed by the images (Rogers, 2016, para. 12).

Mrug et al.'s (2015) reports on the emotional and psychological desensitization of real-life violence versus TV or movie violence on college students (p. 2). The results from their study demonstrated that cognitive and emotional empathy increased from a low to medium level of exposure to real-life violence (Mrug et al., 2015, p. 20). However, cognitive and emotional empathy decreased when exposure to real-life violence reached high levels (Mrug et al., 2015, p. 20). Furthermore, exposure to televised violence was found to influence blood pressure but was unrelated to emotional functioning (Mrug et al., 2015, p. 2). Moreover, Mrug et al. (2016) studied the relationship between emotional desensitization and violent behaviour in adolescents. They concluded that emotional desensitization to constant exposure over multiple contexts of violence "in early adolescence contributes to serious violence in late adolescence" (Mrug et al., 2016, p. 1).

Dill & Brockmyer (2012) measured short term and long term psychophysiological and behavioural effects of exposure to media violence (p. 3-8). They suggest that the degree of psychological engagement with violent media may be more of a factor in determining the risk of negative effects on an individual (Dill & Brockmyer, p. 8). Furthermore, they conclude by recommending that newer imaging technologies be used by researchers to study the short and long term effects of desensitization on neural structures that are involved in empathetic responding (Dill & Brockmyer, p. 8).

Krahé's (2011) explores the effect of habitual exposure to violent media stimuli on desensitization, and aggressive cognition and behaviour in men and women through testing pleasant pr anxious arousal and skin conductance levels (p. 1-25). The study used violent, sad, and funny clips to compare and test results (Krahé, 2011, p. 1). It was found that increased habitual exposure to violent media content in a laboratory setting correlated with less physiological reactivity (Krahé, 2011, p. 21). In both women and men, there was a link between greater habitual exposure to media violence and greater pleasant arousal as opposed to anxious arousal (Krahé, 2011, p. 21). However, men only demonstrated a marginal significance when compared to women (Krahé, 2011, p. 21). Both men and women demonstrated a link between greater habitual exposure to media violence and more rapid accessibility to aggressive cognition (Krahé, 2011, p. 21). However, women only demonstrated a marginal significance compared to men (Krahé, 2011, p. 21).

Based off of current literature, the frequency of exposure, the degree of psychological engagement, the sex, and the nature of the content, seem to be key aspects in the mental and emotional health of consumers of violent, psychologically challenging, or hateful content. The findings broadly suggest that the more exposure to violence, the more psychologically and emotionally adverse side effects, and the more chances that exposure may result in aggressive cognition/behaviour. Additionally, the more violent, hateful, or psychologically challenging the content, the more adverse side effects in its consumption.

Research Challenges

Considering the effects of being exposed to this kind of language and rhetoric for extended periods of time, there are three research challenges when handling this type of data:

1. Finding qualified collaborators outside of the social sciences that are comfortable working with this type of data
2. The maintained well-being of the analyst during and after the project, which includes minimizing time spent immersed in the data
3. A scoring process to differentiate hate speech as defined in the criminal code, with soft violence that might have kinetic impacts

A requirement to further our research was to devise a method that incorporates filtering methods to reduce the time that researchers have to spend engaged with the psychologically challenging material, to the benefit of the analyst and the quality of the research output. To that end, a model has been developed by Simon Fraser University Computer Science Masters student, Bdour Alzeer, in collaboration with CASIS Vancouver in order to devise a model that was able to address the above research challenges. The model offers solutions in the following ways:

1. The data is presented in a way that gives the analyst a choice whether or not to view the raw data based on a summary of the entry.
2. Time spent immersed in the data is significantly reduced by the summary-system, hate-ranking system and topic modeling that allows researchers to draw conclusions from the data more quickly.

The Filtering Process

Alzeer’s model is unique because it ranks each individual entry, which in this case is in the form of online comments, from 0-5. A score of 0 indicates either a completely blank entry, an entry without enough data to gain points, or a photo. The higher the ranking, the more hate-speech indicators the entry has.

The entries from level 1 of the model display the score, thread title and username. Based on those three things the analyst can decide whether or not to go to the second level of the display. In level 2, the selected entry will show the predicted class (hate speech, offensive language, neither), the topic ID number, score, action words (violence-based terms such as kill, hit, etc.) and possible targets for the verbs (named entities, places and things). There is additionally an option to view entries that are similar to the one that is being viewed. Level 3 will display the original entry in its raw form without any blurred words or hidden content. This process greatly reduces any stress or potential harm to the analyst by not needing to look at the data in its raw form unless they choose to themselves.

Implications of the Scoring Process

The scoring process in Alzeer's model has additionally created a strong foundation for further research in how machine learning can be used to make the online space safer for users. The scoring process can be used to include social violence, including the posting of addresses and personal information of community members for the purpose of encouraging doxxing. It will also be used to include two lexicons the CCIBC has collated: an Incel lexicon, and a far-right lexicon that include terms unique to these groups in the online space that help them evade detection by bots on looking to exclude hateful rhetoric from their online platforms. Adding these lexicons into the scoring process will allow for a more accurate ranking process.

Currently the model uses Natural Language Processing tasks in order to filter the data. Because of this and as previously stated, the model cannot include images. While testing this model with a security problem and from the analysts perspective, it became evident that many entries ranked as 0 due to there being little to no actual text in the entry were actually photos. By finding other entries that are relevant to the image posts by using the option built into the model, we were able to see that several images ranked as 0 were associated with high scoring entries. This demonstrated the likelihood that these images contained hate-speech or offensive language with clear targets. The potential to integrate image recognition into this model means that in future we could decrease the number of miscored posts.

Another area for improvement arose when analysts found evidence of "bumping" during the test. Post bumping occurs when users post blank comments, one to two word comments, or punctuation in order to increase the popularity of the threads. The impact of this bumping creates the impression that a specific thread has more followers and activity than it actually has. The reality is the bumping is creating 'noise' inside the dataset and the end user is looking at a thread which may not have meaningful content. A future iteration of the algorithm can be improved to account for bumping and noise within the data set to allow for a more reliable analysis.

Concluding Thoughts

This project is unique in that it summarizes the data for the analyst. This allows for the well-being of the analyst to be at the forefront of any further analysis with hateful comments and language. As previously explored, the psychological effects that this kind of language has on the analyst is a continued problem that this project solves.by reducing the time an analyst has to deal with reading this type of information.

Based off of tests with CASIS Vancouver analysts, it allowed for them to identify the top posters and commenters that were recruiters for these forums. This demonstrates the ability of the filtering process to allow for analysts to focus more on the actual analysis rather than having to work through every data entry from the bottom up since unrelated or irrelevant comments are filtered out. Additionally, the scoring process could be elaborated on to include posts that have a higher number of replies than average as well as those with verbs associated with violence to help identify the top posters using this type of language.

Overall, the development of this model and its applications will be able to alleviate the stress of reading hateful comments and rhetoric and allow them to focus more on in depth analysis and identification of top comments, posters, and those who might be recruiting.

References

Dill, K., & Brockmyer, J. (2012, December 31). Media Violence, Desensitization, and Psychological Engagement. In The Oxford Handbook of Media Psychology. : Oxford University Press. Retrieved 18 Oct. 2019, from https://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780195398809.001.0001/oxfordhb-9780195398809-e-12.

Gibbons S.W., Hickling E.J. & Watts D.D. (2012). Combat stressors and post-traumatic stress in deployed military healthcare professionals: an integrative review. Journal of Advanced Nursing 68(1), 3–21.

Kelshall, C. M., & Meyers, S. (2019). PREPARED: A program to educate on the prevention and reduction of extremist discourse. Burnaby, BC: Simon Fraser University Library.

Koehler, D. (2014). The radical online: Individual radicalization processes and the role of the internet. Journal for Deradicalization, 1(2014), 116-34.

Krahé, B., Möller, I., Huesmann, L. R., Kirwil, L., Felber, J., & Berger, A. (2011). Desensitization to media violence: links with habitual media violence exposure, aggressive cognitions, and aggressive behavior. Journal of personality and social psychology, 100(4), 630–646. doi:10.1037/a0021711

Leuprecht, C., Hataley, T., Moskalenko, S., & McCauley, C. (2010). Containing the narrative: Strategy and tactics in countering the storyline of global jihad. Journal of Policing, Intelligence and Counter Terrorism, 5(1), 42-57.)

Martineu, Paris. (2019, February 5). The Existential Crisis Plaguing Online Extremism Researchers. Wired. Retrieved from https://www.wired.com/story/existential-crisis-plaguing-onlineextremism-researchers/

McCauley, Clark and Sophia Moskalenko. 2017. "Understanding Political Radicalization: The Two-Pyramids Model." American Psychologist 72 (April): 205-216. http://psycnet.apa.org/journals/amp/72/3/205/

McCosker, H., Barnard, A., & Gerber, R. (2001). Undertaking Sensitive Research: Issues and Strategies for Meeting the Safety Needs of All Participants. Forum Qualitative Sozialforschung / Forum: Qualitative Social Research, 2(1). doi:http://dx.doi.org/10.17169/fqs-2.1.983

Meyers, S. (2019). Is there a gap in Canada's hate crime laws? The identification of soft violence as a tool for current right-wing extremist social movements. Journal of Intelligence, Conflict and Warfare, 2(2).

Mrug, S., Madan, A., Cook, E. W., 3rd, & Wright, R. A. (2015). Emotional and physiological desensitization to real-life and movie violence. Journal of youth and adolescence, 44(5), 1092–1108. doi:10.1007/s10964-014-0202-z

Mrug, S., Madan, A., & Windle, M. (2016). Emotional Desensitization to Violence Contributes to Adolescents' Violent Behavior. Journal of abnormal child psychology, 44(1), 75–86. doi:10.1007/s10802-0159986-x

Neumann, P. (2013). Options and Strategies for Countering Online Radicalization in the United States. Studies in Conflict & Terrorism, 36(6), 431-459.

Perry, B., & Scrivens, R. (2019). Right-wing extremism In Canada / Barbara Perry, Ryan Scrivens. (Palgrave hate studies)

Pyrooz, D., Lafree, G., Decker, S., & James, P. (2018). Cut from the Same Cloth? A Comparative Study of Domestic Extremists and Gang Members in the United States. Justice Quarterly, 35(1), 1-32.

Rogers, K (2016, July 15) What is a constant cycle of violent news doing to us? The New York Times. Retrieved from: https://www.nytimes.com/2016/07/16/health/what-is-a-constantcycle-of-violent-news-doing-to-us.html?_r=0

Published by the Journal of Intelligence, Conflict, and Warfare and Simon Fraser University

Available from: https://jicw.org/