Identifying Misogynistic Rhetoric Within Advice-Seeking Communities

Tanvi Bhagwat
tbhagwat6@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Aiswarya Bhagavatula
abhagavatula8@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

Disha Das
ddas71@gatech.edu
Georgia Institute of Technology
Atlanta, Georgia

ABSTRACT

With easy access to the Web, it is easy for an individual to find an esoteric community of people that supports them in their interests that they wouldn't have found otherwise. Social websites such as Reddit lets the users be active under the cloak of anonymity, thus freeing them to discuss topics that they wouldn't have been able to in real life. Within Reddit, there exist communities that allow posting of personal problems for collaborative solutions. Unfortunately, the same cloak of anonymity that frees up a user to discuss topics without fear of exposure, allows another user to spread about vicious ideologies that are abhorred by the public. These advice-seeking communities are then forced to confront with ideologies shunned by the populace.

The presence of vile discourse might be jarring to those active within the community, and may be responsible for the vulnerable advice-seeker to go down avenues of thought that they wouldn't have considered during a better disposition. We seek to explore the possible rise in such ideas in these communities, their effect on the users who have been accosted with such values, and if there are communities that may be liable for the rise in such notions.

1 INTRODUCTION

Reddit is a social voting site that refers to itself as the "front page of the internet". [2]. It is now the 5th most visited website in the US, with an average of more than 430M+ monthly active users as of November 12, 2017 [7]. Users often ask for advice within the subreddits. Our research hopes to examine the increase in misogynistic ideas within the advice meted out to the user.

This is important as previous studies have shown that individuals with negative prior attitudes have a greater latitude of acceptance when they read negative comments about a subject. People with prior negative attitudes when confronted with refutational two-sided comments perceived the comments as more negative and acceptable than the positive comments. Thus, the presence of hateful comments might fuel notions for the already vulnerable advice-seeker [5].

Our research seeks to investigate the following questions:

(1) *Has the advice that has been doled out by the communities grown more misogynistic over the years?* We will conduct our research on this question by looking at the Reddit's datasets over the years for the selected communities, and applying Natural Language Processing to capture the presence of misogynistic ideas. The data will be compared to see if there is a rise in misogynistic rhetoric.
(2) *Which groups dole out such advice?* To answer this question, we will assess the communities that the users commenting with the negative rhetoric are active within. This will shed light if the feelings expressed by the user were a one-off or something they've felt for a while.
(3) *Check if the presence of negative community engagement drives down the user's engagement within these subreddits?* If a user is confronted with speech that is shunned by modern-day society, we are interested in seeing if this drives the user away from the community.

By answering these questions, we hope to contribute the following:

(1) *Help the users posting for advice understand the advice that comes their way.* For users who are posting in advice-seeking communities, they may be of a sensitive disposition. It might help them understand the advice being given to them better if they understand where it stems from.
(2) *Create safe spaces.* Often, the users posting in such communities may have undergone potentially traumatic experiences. Being exposed to inflammatory comments may lead to traumatic triggers in what the user thinks of as a safe space.
(3) *Check the spread of fringe ideas from becoming mainstream* Underprovision on monitoring might drive away users, thus letting fringe ideas become more acceptable within these online spaces leading these communities to disintegrate

2 PREVIOUS WORK

A study done by Farrell et al. [1], investigated the flow of language across seven online communities in Reddit. The communities they explored were contained to the communities that self-identified with the incel idealogy or men's rights activism. They gathered the comments using the pushshift API. To characterize the misogyny they built 9 lexicons for the different levels of hate speech. Then they used these constructed lexicons to calculate the amount of misogynistic posts per community, the amount of users posting such content per community, the top terms in each community, and the evolution over time of these communities for the different levels of misogyny. The study observed that hate-speech towards women has been on the rise in these communities, and that violent rhetoric usually co-occurs with such narratives.

However, this study was limited to the communities already infamous for their ideologies. As of 2/5/2020, two of the seven communities have already been quarantined. Users who don't identify with these values will not be active within these online spaces. However, it is imperative that spaces such as the advice-seeking communities be monitored for the presence of misogynistic ideas to prevent the spread of such values within the vulnerable.

Kumar et al. [4] showed that for communities that are targeted by other communities, the users within the targeted communities become less active. For users with a positive prior attitude, being

confronted by negative comments or "trolls" may perhaps lead to a similar drop in activity levels within these subreddits. If there is a significant increase in toxic ideas being utilized through discourse, it may lead to these communities degrading and suffering the fate of other subreddits that have disintegrated this way. Thus, we seek to explore implicit sentiments such as sarcasm, irony and misogynistic values to see if exposure to such values leads to the posters being less active.

Joshi et al.[3] presented a method of detecting sarcasm that is grounded on the theory of context incongruity. This work handled incongruity between the text and common world knowledge, leaving out sarcasm that might depend on the situation. The feature vector of a tweet consisted of four kinds of features: (a) Lexical, (b) Pragmatic, (c) Implicit congruity, and (d) Explicit incongruity features. An algorithm was implemented to detect sarcasm with an improvement of 5% over baseline. The same research could be applied to Reddit comments to give a nuanced understanding.

3 PLAN OF ACTION

3.1 What is new in your approach and why do you think it will be successful?

We plan to analyze data from a few subreddits using Natural Language Processing models like sentiment analysis. The motive is to analyze discourse to identify negative narratives. Pamungkas et al. [6] have implemented a similar model for identifying misogyny on Twitter data. Reddit as a platform is more prone to misogynistic comments and negativity due to the lack of user transparency and longer comments (when compared to Twitter's 280 character limit). The existing methods mainly concentrate on subreddits that already practice hateful speech as opposed to general advice seeking subreddits.[1] Our motive is to analyze if (and how) hate speech has increased over the years on such subreddits and make correlations between reactions and susceptibility towards negative comments.

We feel positive that this approach would work as a previous study has succeeded in analyzing misogyny for subreddits focused on "men's rights activism"[1]. Thus, applying basic models such as a Bag-Of-Words, followed by LSTM to further tune the gist of the comments, would help us succeed in identifying negative comments.

3.2 What are the risks and anticipated challenges that may be a roadblock for your project?

There always remains a risk of misdiagnosing sentiments due to the quality of the dataset and the Machine Learning model implemented. Initial perusal of the Web and blogs that have put together the datasets didn't sufficiently provide an answer if the datasets also contained comments deleted by the user or the moderator. These deleted comments usually contain inflammatory language, which has led to their removal. The presence of these within the dataset would give more fruitful results. If the datasets don't contain these, we would have to resort to web-mining the sites that keep track of removed Reddit comments.

3.3 Which dataset will you use? Which code repository will you start with, if any?

We will be using the reddit dataset dumps available to the public. We plan on using the pushshift API provided by Reddit to extract submissions, specifically concentrating on the following subreddits:

- *r/AskReddit:* the focus of this subreddit is to "ask and answer questions that elicit thought-provoking discussions". It currently consists of 26.3M members.
- *r/relationships:* This community made up of 2.8M users is for people who are seeking advice "about any aspect of their relationship".
- *r/relationship_advice:* This subreddit of $1.5M$ users is for people seeking advice for any kind of a relationship, ranging from romantic to basic human interaction.
- *r/AmITheAsshole:* This is a subreddit that helps people find out if they were in the wrong in an argument. Currently, it has 1.7M users
- *r/OffMyChest:* this safe space made of 1.8M users helps individuals unload emotional baggage, trauma or any positive feelings they have.

There also exists the Lexicon built by Farrell et al. [1] that we plan to use extensively. Our datasets don't require a ground truth as we are only identifying if there has been an increase in negative comments, which can be easily identified with the presence of keywords.

3.4 How much will it cost and how long will it take?

There is no cost involved in the project. We plan to use open source tools, datasets and technologies for this project. An approximate cost of 200 hours of labour work can be expected. The project is scheduled to be completed in 3 months. We will take around 2 weeks for data collection and preparation. The data will then be processed for 4 weeks with various Natural Language Processing techniques to identify and classify the content according to the lexicon. The we will further check if the original poster that received the negative comments has had a drop in their activity. We plan to use the remaining time for completion of documentation and analyzing the results.

3.5 What are the midterm and final exams to check for success?

For the midterm, we plan to:

- Complete data collection and processing.
- Count the comments utilizing the words within the lexicon.

For the final, the checkpoints will be:

- Fine tune the model for comments that may have been filtered.
- Checking user's activity within the subreddits.

REFERENCES

[1] Tracie Farrell, Miriam Fernandez, Jakub Novotny, and Harith Alani. 2019. Exploring Misogyny across the Manosphere in Reddit. In *Proceedings of the 10th ACM Conference on Web Science* (Boston, Massachusetts, USA) *(WebSci '19)*. Association for Computing Machinery, New York, NY, USA, 87–96. https://doi.org/10.1145/3292522.3326045

[2] Eric Gilbert. 2013. Widespread Underprovision on Reddit. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work* (San Antonio, Texas, USA) *(CSCW '13)*. Association for Computing Machinery, New York, NY, USA, 803–808. https://doi.org/10.1145/2441776.2441866

[3] Aditya Joshi, Vinita Sharma, and Pushpak Bhattacharyya. 2015. Harnessing context incongruity for sarcasm detection. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 757–762.

[4] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *Proceedings of the 2018 World Wide Web Conference* (Lyon, France) *(WWW '18)*. International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 933–943. https://doi.org/10.1145/3178876.3186141

[5] Moon J. Lee and Jung Won Chun. 2016. Reading others' comments and public opinion poll results on social media: Social judgment and spiral of empowerment. *Computers in Human Behavior* 65 (2016), 479 – 487. https://doi.org/10.1016/j.chb.2016.09.007

[6] Endang Wahyu Pamungkas, Alessandra Teresa Cignarella, Valerio Basile, Viviana Patti, et al. 2018. Automatic Identification of Misogyny in English and Italian Tweets at EVALITA 2018 with a Multilingual Hate Lexicon. In *Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018)*, Vol. 2263. CEUR-WS, 1–6.

[7] Reddit. 2020. Reddit. https://www.redditinc.com/