Identifying Misogynistic Rhetoric Within Advice-Seeking Communities

Course Project for CSE 6240: Web Search and Text Mining, Spring 2020

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

Reddit as a platform has made it easy for individuals to find communities that support their esoteric interests. 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. To better understand the prevalence of this problem, we analyzed the data from these spaces using Reddit's API. Our analysis did show an increase in the negative discourse taking place in these safe spaces. Not only were users posting about their problems attacked, but also advice givers were attacked with negative words. Further failure to curb the usage of such language could drive down the engagement of users.

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 [6]. 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. *What words are used in expressing foul feelings?* We wish to see what words are frequently used to express negative rhetoric and to see what communities the words attack.
3. *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.

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* Under provision on monitoring might drive away users, thus letting fringe ideas become more acceptable within these online spaces leading these communities to disintegrate

2 LITERATURE SURVEY

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 ideology 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 today, 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 DATASET DESCRIPTION AND ANALYSIS

3.1 Data preparation

Selection of subreddits: We chose the following subreddits that are advice seeking communities for our dataset:

- *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.

Source: From the two subreddits we have chosen to analyze, we mined the posts and comments for our data using Pushshift.io [1], which is a Reddit Search Application. After extracting the posts for each subreddit, to mine the comments made under each post we used Reddit's API called The Python Reddit Wrapper(PRAW)[2]. Since Reddit only allows 30 requests per minute, PRAW helps break it into multiple API calls with two second delays. Along with the text contained in the posts and the comments, we also have access to other details such as the authors, the time the object was created and the score of the object.

Data preprocessing: Raw reddit data is available in the form of a CSV file. The following preprocessing steps were implemented on the dataset:

- Undesired columns which did not contribute to the analysis were dropped.
- Data in the columns 'author_created_utc' (date-time of creation of account) and 'created_utc' (date-time of creation of post) were converted from Unix Epoch Time to datetime object to make analysis easier.
- The dataset was sorted in order of the earliest date of posting to the latest, in accordance with the column 'created_utc'.
- The column 'comments' consists of a list of strings, each string being a top-level comment to the post. These strings are processed by removing non-alphabet letters, converting to lower case and splitting the comments down to individual words.

This dataset suits the requirements for our analysis because it provides us with date-time data of the posting, the frequency of the user's posts for a given period of time as well as the content of the comments.

3.2 Raw Data Statistics

Available information: The following information can be extracted from the dataset- *Author, post creation time, number of comments, over 18 or not, number of upvotes, title of post* and *comments on post.*

Lexicon: We used the Lexicon built by Farrell et al [1]. It consists of 1,300 terms that belong to one of the 9 categories of misogyny, as mentioned in table 1. This lexicon will help identify if the words used within the submissions are toxic or not, and if they are precisely what category of negative rhetoric they belong to.

Raw data statistics: The data statistics are given in table 2.

3.3 Data Analysis

Since this is an analysis project, the results of data analysis has been presented in the Experiments section.

4 EXPERIMENT SETTINGS AND BASELINES

4.1 Experiment Settings

The experiment was run on Google Colab and Kaggle. Kaggle's four cores were used for multiprocessing to speed up the server requests.

Runtime Type: Python3

Hardware Accelerator: GPU

RAM: 12 GB

4.2 Baselines

The approach used by Farrell et al. in their paper [1] was used as the primary baseline for this project. They chose six Reddit communities that have identified themselves as "incels". They gathered all the posts and comments of each community from their creation all the way through to 2019. They used the Pushshift API to crawl for all the posts and the comment submissions. Then these comments were checked for the presence of words from the hate lexicon and for misogynistic content. This paper was chosen as a baseline as it archived the growing trend of using negative discourse within Incel-identifying communities. We plan to showcase similar trends but for more general advice-seeking communities. The paper guided us in the right direction on choosing the correct lexicons, and to highlight the correct results. The original paper calculated the frequencies of the misogynistic content across categories such as misogyny, belittling, flipping narrative, homophobia, hostility, patriarchy, violence, racism and stoicism. It also looked at the users using such words across the similar categories. It then mapped the usage of these words over time for all the communities.

We tried to emulate similar results. Additionally, we also try to analyse specific user behavior on these subreddits. We identified the frequent negative rhetoric using posters and looked at what communities they usually associate with.

[1] https://pushshift.io/

[2] https://praw.readthedocs.io/

Table 1: Lexicon of Misogyny

Category of Misogyny	Num terms	Examples
Belittling	58	femoid, titties, stupid cow
Flipping the narrative	7	beta, normie, men's rights
Homophobia	126	dyke, fistfucker, faggot
Hostility	303	bitch, cunt, whore
Patriarchy	8	alpha male, subjugate, suppress
Physical Violence	73	hit, punch, choke
Racism	670	nigger, raghead, pikey
Sexual Violence	22	rape, sodomise, gangbang
Stoicism	33	blackpill, cuck, hypergamy

Table 2: Raw data statistics

Community	Num Posts	minDate	maxDate	Avg num of comments	Max num of comments	Min num of comments	Comment Vocab size	Avg num of words per comments section	Avg score	Max score	Min score
r/r/AmITheAsshole	4670	2015-06-02	2020-04-17	573.97	11231	0	196127	11878.95	3296.55	71339	1
OffMyChest	10158	2010-02-25	2020-04-25	75.4	2932	0	163711	3275.89	692.94	17443	0

Table 3: Percentage of Misogynous posts across categories for Subreddits OffMyChest and *r/r/AmITheAsshole*

Community	OMC	AITA
Belittling	32	54
FlippingNarr	2	8
Homophobia	3	13
Hostility	66	99
Patriarchy	2	28
P. Violence	63	99
Racism	26	98
S. Violence	11	34
Stoicism	12	54

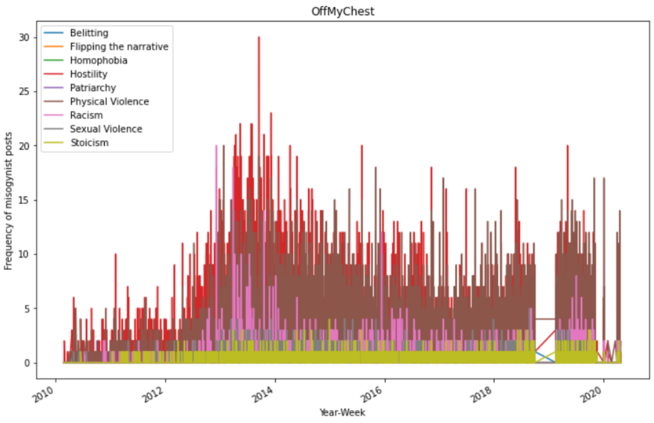
5 PROPOSED METHOD

Our method seeks to utilize the data from the advice-seeking subreddits to identify trends in usage of hate speech, if any. These steps can be applied to an assortment of Reddit spaces. As for the hate words corpus, we chose the ones that helped identify misogynistic rhetoric, which is the same as that used by Farrell et al. [1]. But the reader could utilize anything that would help identify specific problems. Then we analyzed the trends across subreddits to see if any patterns could be identified.

The existing methods mainly concentrated on subreddits that already practice hate speech as opposed to the general-advice seeking subreddits. Unlike [1], we present analysis on a variety of fronts. For one, the *Redditors* who frequently use hateful terms may frequent toxic communities. This has been verified in this report. Moreover, we also present analysis on the exposure of posts suitable for under 18 years of age to the hateful comments on posts meant for over 18 years of age. Lastly, we take into account the score of the post and find the amount of negativity associated with it. This analysis is meant to find a relation between the down-votes received on a post and the content of the comments section.

After pre-processing the data, the comments were processed to extract individual word tokens. The common words found among the comments and lexicon under each category were considered and counted. This data was then used to generate plots of frequency trends and other observations as shown in the Experiments section.

Figure 1: Frequency of misogynous words over time for *r/r/OffMyChest*



6 EXPERIMENTS

Thus, drawing inspiration from Farrell et. all [1], we decided to analyze three subreddits that are known within the Reddit community as places that dole out relationship advice. We used the same evaluation metric that Farrell et. all used. A lexicon made of seven

Table 4: Top 5 Misogynous words used across categories in *r/r/AmITheAsshole*

Belittling		FlippingNarr		Homophobia		Hostility		Patriarchy		P. Violence		Racism		S. Violence		Stoicism	
Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq
Dumb	1985	Mra	100	Homo	464	Ho	4641	Overwhelm	724	Er	4639	Ike	4511	Rape	885	Rope	2348
Female	1216	Prevail	97	Queer	144	Asshole	4622	Oblige	435	Hit	3799	Abo	4418	Pound	762	Cope	748
Failure	727	beta	90	Fag	95	Hos	3474	Compel	365	Ram	2584	Nig	2592	Molest	165	Fuel	362
Dumbass	653	Mgtow	79	Faggot	57	Hate	3174	Suppress	181	Cut	2584	Amo	2529	Spank	108	Incel	242
Puss	416	Normie	22	Lez	25	Hurt	2763	Rope	12	Force	2330	Leb	2096	Incest	90	Cuck	126

Table 5: Top 5 Misogynous words used across categories in OffMyChest

Belittling		FlippingNarr		Homophobia		Hostility		Patriarchy		P. Violence		Racism		S. Violence		Stoicism	
Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq	Word	Freq
Female	1529	Beta	87	Queer	103	Hate	3989	Suppress	105	Hit	2775	Black	1488	Rape	792	Cope	833
Dumb	1417	Mra	60	Faggot	93	Hurt	3311	Overwhelm	59	Cut	2166	Bigger	1345	Pound	223	Fuel	233
Failure	720	Prevail	40	Homo	74	Asshole	2054	Oblige	28	Kill	1911	Hun	217	Conquer	101	Rope	114
Boobs	253	Overthrow	25	Fag	51	Beat	1543	Compel	26	Force	1544	Arab	64	Incest	94	Redpill	33
Dumbass	225	Normie	10	Dyke	22	Harm	1150	Omega	15	Kick	1246	Cracker	43	Molestation	79	Incel	28

Figure 2: Frequency of misogynous words over time for *r/r/AmITheAsshole*

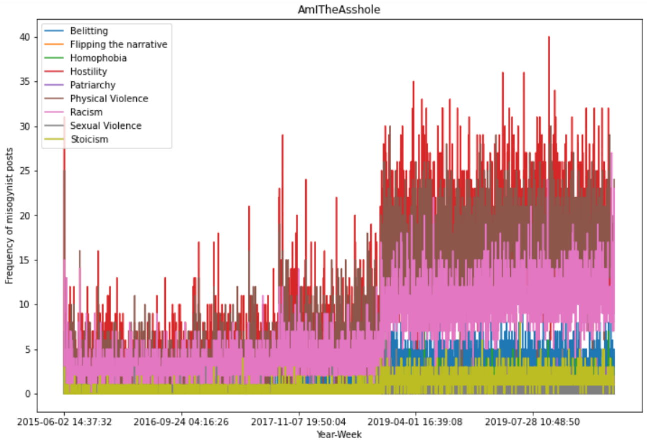
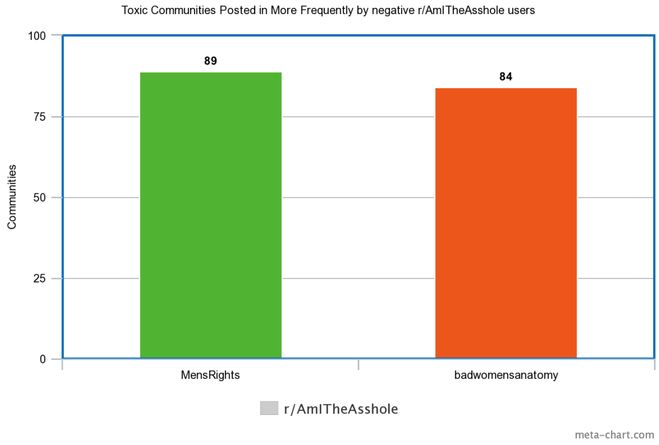


Figure 3: Frequency of comments in toxic subreddits by users of *r/r/AmITheAsshole*



already existing lexicons of hate speech and specific misogynistic rhetoric was used. Each comment under a submission was assigned a "hate score" if the presence of hate words from within the corpus was identified. The offending words were added to a list to keep track of the most frequent words. Side-by-side comparisons of the baseline and our project showed similarly increasing trends albeit with different rates of increase. The original study was done within groups already know for their misogynistic rhetoric. The negative discourse used in those communities is often and specific. While the size of data sets could be the reason, but the offending words for the general communities were more "common place" compared to the ones used in the Incel subreddits.

As observed from Table 3, *r/OffMyChest* is not very different from the AskReddit community analyzed in the mid-term report. The most frequent theme of the comments section seem to be Physical Violence, Hostility and Belittling. From Table 5, it's observed that the theme of Hostility dominates over all the other categories for *r/OffMyChest* subreddit. The most popular words were *Hate, Hurt, Asshole, Hit*,etc. Rhetoric that is exclusive to certain toxic communities finds its way into this subreddit. Some of these words include *Beta, Mra, Normie, Omega, Redpill*, etc.

As for the subreddit *r/AmITheAsshole*, which frequently sees posts that cause much polarity among the original posters and the advisors, we observe a dominance of the themes of Hostility, Physical Violence and Racism. It is also observed that mean posts are made far more frequently as compared to other subreddits. This could very well be due to the polarizing nature of the posts. From Table 4, words that are blatantly derogatory have a clear dominance. For example, we see a high frequency of words such as *Ho, Asshole, Dumb*, etc. Certain rude remarks are expressed such as *Er*, which stands for 'Eye-roll', is the second most frequent word used. It could point to a hostile or generally disagreeable banter. The theme of Racism, which has been in the backdrop for the other communities,

Figure 4: Frequency of comments in toxic subreddits by users of *r/r/OffMyChest*

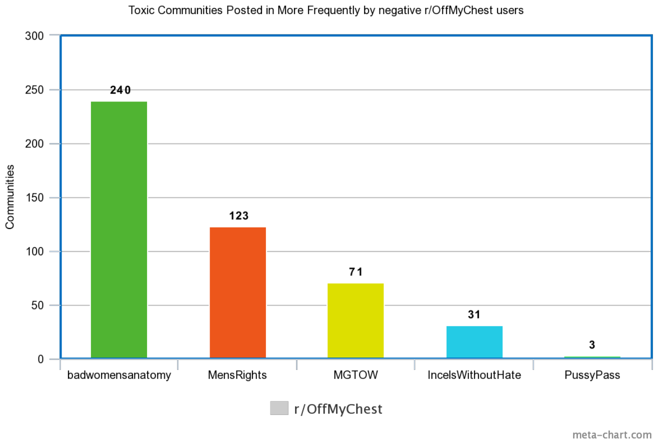


Figure 5: Age group vs Frequency of negative words for *r/r/AmITheAsshole*

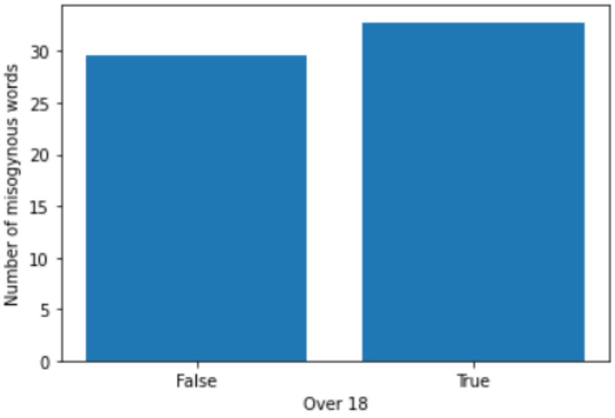
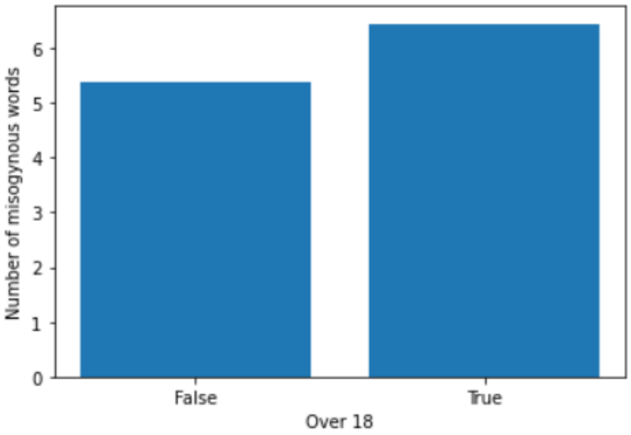


Figure 6: Age group vs Frequency of negative words for *r/r/OffMyChest*



sees a rise. Racism is unabashedly expressed through words such as *Ike*, which refers to beating or abusing a woman, *Abo*, which is an Australian disparaging term for Aboriginals, and *Nig*, which is an American offensive slang for African-Americans.

From Figure 1, a spike in abusive terms in comments is observed around the year 2014 for the subreddit *OffMyChest*. It is difficult to account any explanations to this peak. The frequency seems to taper off. This could be due to a change in community rules and its rise in popularity. From Figure 2, a more interesting trend is observed for the *r/AmITheAsshole* community. It's clear that themes like Racism, Hostility and Physical Violence have not only been dominant, but they have been on the rise over the years. It's interesting to note that the hatred seems to shoot up towards the beginning of 2019. It could be accounted for by the subreddit loosening their rules on posts, and the overall popularity of the subreddit, not only on the front page, but also popular media.

The amount of negativity for posts meant under the age of 18 in *r/OffMyChest* is in no comparison to that of *r/AmITheAsshole*. The amount of negativity is almost equal for both the age-oriented posts in the two communities, but the frequency is far greater for *r/AmITheAsshole* than for *r/OffMyChest*. The plots can be observed from the figures 5 and 6.

The ratio of negativity in the posts having a low score seem to be far higher than those with high scores. This trend is similar for both the communities. This goes to show that the posts receiving most of the down-votes are commented upon negatively. The plots can be observed from the figures 7 and 8.

The users that made the most negative comments were found to be involved in other communities that upheld certain negative ideologies. The plot for this data is given in figure 3. The users of AITA were found to be most involved in AITA itself, which itself is found to be a toxic community from the data we presented. In case of OMC, from figure 4 we can see that the users seem to be posting less frequently, but they're still involved in communities flagged for hate-watch by the Southern Poverty Law Center.

7 CONCLUSION

Our work has shown results that align with the objective - we noticed an increase in negative discourse in posts on advice seeking communities and we could relate the frequenters being active in posting on various other negative communities. The analysis has been limited to 2 subreddits. The paper by Farrell et al. [1] has scraped a wider dataset than the one we have. This, coupled with the already toxic communities they are looking at, the frequencies of the negative words utilized is higher. The results also show that the vocabulary used in a general community is less specific than the vocabulary used by the Incel communities.

Therefore, this work can be extended to other subreddits in advice

Figure 7: Ratio of negative words used in comments with respect to the score for *r/r/AmITheAsshole*

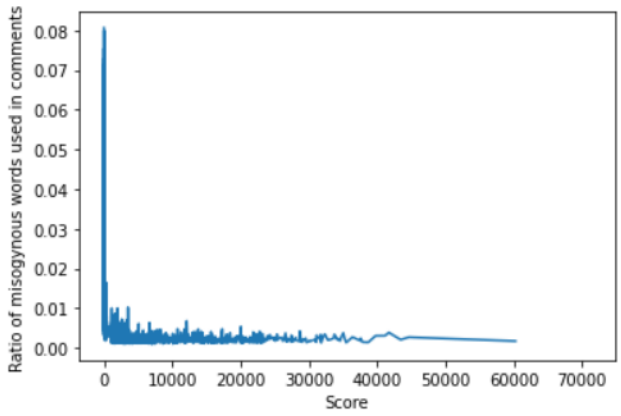
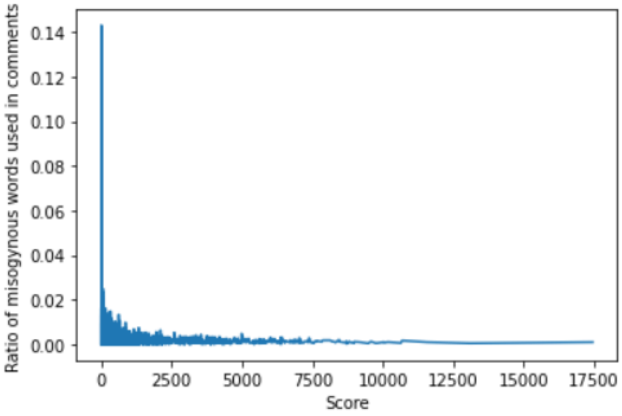


Figure 8: Ratio of negative words used in comments with respect to the score for *r/r/OffMyChest*



seeking communities to gain a larger dataset and analyse the trends across other communities as well. Additionally, more in-depth analysis can be made to identify a trend in the content projected by a given user to spot possibilities of severe problems. Algorithms using the concepts of Machine Learning and Deep Learning can be incorporated to train a model on the dataset to alert moderators quicker, or indicate to the user how seriously to take advice espoused by fellow Redditors.

8 CONTRIBUTION

All the team members have contributed a similar amount of effort.

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