Processing what makes a good reddit comment

Andrew McLeman, Joe McDonald, Jordan Youngman,
Murray Lyne, Scott Thomson

A report submitted as part of the requirements for the degree
of MSc in Cyber Security
at the School of Computing
Robert Gordon University
Aberdeen, Scotland

April 2019

Supervisor Eyad Elyan

Abstract

The Abstract of the report should be written here, it should provide a short summary of the work encompassing no more than 300 words.

Declaration

I confirm that the work contained in this MSc project report has been composed solely by myself and has not been accepted in any previous application for a degree. All sources of information have been specifically acknowledged and all verbatim extracts are distinguished by quotation marks.

Signed .... Date ....

Andrew McLeman, Joe McDonald, Jordan Youngman, Murray Lyne, Scott Thomson

Contents

List of Tables

List of Figures

Chapter 1

Introduction

The goal of this paper was to determine what makes a good comment on the social media platform known as Reddit. This research was done to not only help people realise how to form an adequate post that can prove to be popular but also help train chatbots to make them seem more lifelike and to give the illusion of human speech. The data was parsed using a parser to filter down the comments of each subreddit. This made it easier to define a context for analysis. This research can prove to be useful for a number of reasons such as the previously mentioned training of AI. It could also be useful for the likes of marketing companies who need to know how to best target their online content to a wider audience. It may also be of great interest to the likes of politicians to see which side the general public stand on political events. The hypothesis of this paper would be that comments don't stick to a predetermined formula. Their success is instead determined by varying factors such as time of post, captivating or "clickbait" titles and how relevant it is with today's current events.

1.1 Background

Reddit is an online Score aggregator Website. It is used for a wide variety of discussions and by all kinds of communities.

1.2 About this Thesis

This is the thesis of *Insert Full Name Here*, submitted as part of the requirements for the degree of MSc Computing: Software Technology at the School of Computing, Robert Gordon University, Scotland.

A number of paragraphs detailing the main expectations of this body of work.

1.3 Conclusion

A short conclusion summarising the chapter.

Chapter 2

Background Research

This chapter provides some background research on the project and examines some previous work.

2.1 Political Motivations behind Subreddits

Reddit is home to several political subreddits, many of which have high subscriber counts and activity per day. As stated in the paper titled Automating power political actors have used bots in social platforms to influence public opinion. They are used to not only raise awareness in certain political campaigns but also to follow politicians on platforms such as twitter and Facebook to give the illusion of popularity. They can even do an attack on news outlets where they flood their wall with misinformation as to avoid and sidetrack public attention. [10] the above research strongly shows why in our paper it is imperative that we distinguish AI from human users to get accurate and truthful data to work with. Also stated by the paper titled Melanization of Politics it describes and researches in the rise of political discussion due to the rise of media in the late nineteenth century early two-thousands it discusses how politics is now spoken on such a large scale. [11]

2.2 Psychological Traits in Users

Psychological traits play an important part in social media posts.

This paper [?] suggested that Users with Stronger personality traits tend to have stronger engagements in online discussions and therefore have increased interactions on Reddit. These users tend to use social media more because they have a more extroverted personality, and are much more open on social media platforms. This also means due to their extrovert traits, that these users tend to have better leadership qualities. The results, however, had issues with statistical significance.

Users tend to display “Cyborg-like” behaviour when trying to attract attention to their posts. This doesn’t always work and a significant number of posts studied [1] failed to garner the attention they wanted.

2.2.1 User Behaviour Patterns

It was found that women tend to post far less than men, and age determines how common a user will post on the platform. Users who had higher news engagement tend to comment and vote more on news related posts. As well as that, Older users were much more likely to post on Reddit. Voting patterns have no significant predictors, with the only exception being that news engagement can predict voting. It was suggested that the reason women post less is due to the fear of suffering from online harassment, as Reddit has a history of harbouring subreddits that have “existed to only annoy other redditors”. Egregious examples include /r/fatpeoplehate and /r/rapingwomen. Of which, were banned by Reddit in 2015. [2]

When it comes to positive events occurring on Social Media, it was found that there is an increase in negative emotions in social media on Twitter. It was found that on Reddit, highly upvoted stories tend to gather a proportionate amount of downvotes, therefore indicating that the same effect can happen on Reddit as well.[3]

Most posts on Reddit “die” after one day in terms of activity, which is a pattern that has been observed on many other social media platforms. For posts that only had 1 posted comment, it was found that 72.78 Per cent of these post did not last more 600 seconds in terms of new activity [1]. Users who comment frequently on other user posts tend to have more highly scored posts and therefore if you want to have good interactions with other posts, you need to be reciprocative. 67 Per cent of authors had more effective comments than posts compared to 22 Per cent who has less effective comments than posts.

Most users are only active on a handful of subreddits. It was found that during a 1-year data collection on 309 Reddit users, only 109 unique users(104 subscribed and 44 unsubscribed) had subscription events fired (This is when a user subscribes to a subreddit). Users tend to have varying attention spans as 73 Per cent of posts are rated without viewing the content of the post first. It was found that most users were “headline browsers” who only look at the headline of a post before voting on it. It was also found that most users also only look at the top headline posts on the front page, which is known as position bias. This means that users tend to gravitate towards posts that are on top of the front page of a subreddit, compared to posts further down the page. Users often vote on posts before viewing the comments on the post. Over 50 Per cent of users vote before actually checking the comments replying to the post. [4]

It has been noted that the variety of subreddits browsed by users is lacking. A severe lack of browsing and voting variety was noted which indicates that users tend to create "echo chambers", in which they only view content that they agree with. [4] This is a very commonly observed phenomenon across many social media platforms, and Reddit is no exception. [4]

Moderation doesn't always affect user behaviour on Reddit. Subreddits themed around real-world discussion tend to value analytical and objective based comments compared to less serious subreddits. Even when filters are enabled preventing certain types of content being submitted, this makes little difference in the behaviour patterns of users.

2.3 Significant Events

Governments use bot farms on social media to manipulate public perception on social media, and Reddit is no exception. Many Governments, militaries and significant state actors have been outed as having used social-bots to appear either more relevant and influential or more controversially to directly manipulate legitimate users opinions on political events and military regimes. [5]

2.4 Machine Learning

Reddit is host to multiple AI experiments as well as many bots to reply to incorrect post syntax. Reddit is commonly used due to having various different subreddits from politics down to humour subreddits. This makes for an excellent place for AI to learn through varying teaching methods such as reinforcement learning to learn from peers in the subreddit or to strive for higher upvotes then previous posts. [6]

The popularity of Reddit posts can be hard to classify without analysing the content of the post first. It can also be hard to classify posts if the subreddit in question covers a broad subject. Post popularity is determined on the content of the post, when the post was posted and the particular subreddit it was posted on, which, can be regarded as the context for discussion. An example of this would be the subreddit /r/pics, which had a poor performance in error rate due to the broadness of the subreddit's context. Whereas other subreddits follow more specialized trends which are easier to classify. This paper ran classification algorithms and found that it ran better on simpler datasets. It also found that features like a Reddit post's title lacked "indicative" power on whenever it made it a popular post or not. It suggested that more features from the Reddit API would improve prediction ability. [7]

With that said, this paper [8] suggests that it is possible to predict the score of Reddit comments, even if the communities are loosely defined (in terms of context) and disorganized. It found of note that user flairs (which are often given by moderators of a subreddit) can be an excellent predictor of highly popular comments. The model could also retrieve, with good precision, the highest rated comments within a particular post. It found that lowly rated comments tend to contain outdated jokes or information, which by that point had been cycled by the community to the point of exhaustion. Machine learning on Reddit needs to take into account community-related factors into what makes a comment popular. Posts on expertise related subreddit communities tend to use more technical and analytical terms. Likewise, comments on news related subreddits tend to be less emotional and value analytical insight over emotional responses. Some subreddits can vary in language use, the /r/worldnews subreddit is more diverse lexically, whereas /r/worldpolitics tends to use more "netspeak" compared to other subreddits. Emotional comment preference can vary between subreddits, with some subreddits heavily disliking having emotion in their comments. Time can play an important factor in determining how popular a comment can get. Comments on communities such as /r/news tend to be much more popular compared if posted later in time compared to comments made earlier, whereas other communities show the opposite happening. Moderation doesn't always affect user behaviour on Reddit. Subreddits themed around real-world discussion tend to value analytical and objective based comments compared to less serious subreddits. Even when filters are enabled preventing certain types of content being submitted, this makes little difference in the behaviour patterns of users. This means that moderation doesn't necessarily need to be taken into consideration when ranking comments, as the user behaviour is always affected by a significant amount.

2.5 Conclusions

The main conclusions for this chapter.

Chapter 3

Method

3.1 The dataset

The dataset was the entire collection of every Reddit comment (including deleted comments) posted on December 2018. Due to the sheer size of the dataset being very hard to parse as one large dataset, it was decided to parse down comments to specific subreddits. (Table to be added in R) The data was stored in the json format.

Each subreddit was placed its own json file and kept in cloud storage, allowing to be accessed by any of our team members. Each dataset has a varying level of bot created comments, which depends on how heavy the automated moderation is on each subreddit. To deal with this, we filtered out any users with a "-bot" suffix and any users that had a history of posting the the same type of comment repeatedly. An example of this would be AutoModerator, a popular bot used to automate moderation of subreddits.

3.2 Data preparation

After clearing out the bot users, the data is further filtered out to only the required fields. The main ones used were Score, User, the content of the message, the date it was posted and the length of the message.

For text mining and processing. It was required to prepare the text of the comments for meaningful analysis. The first step was to remove common words and filter out "stop words" used in the English dictionary. Afterwards, the text was stemmed to reduce words to their core meaning. An example of this would be reducing "enhancing" just to "enhance".

Document term and term document matrices were created after the stemming process was completed, at this point it was possible to perform exploratory text analysis and to gain a better understanding of the vocabulary used by each subreddit community. This was done in order to prove that each subreddit provides a context which radically changes the vocabulary that each user uses.

3.3 Basic Analysis

3.3.1 Relationship between original post time vs time since posted.

One of the vectors investigated was the relationship time has on the popularity of posts. Analysis shows that posts that are posted earlier are much more likely to get a higher score. The mean score goes down further on as time progresses before flat-lining at 100 5 minute intervals. Therefore, we can conclude that the mean score settles after 8.3 hours after the original post was created if you divide 500 minutes by 60.

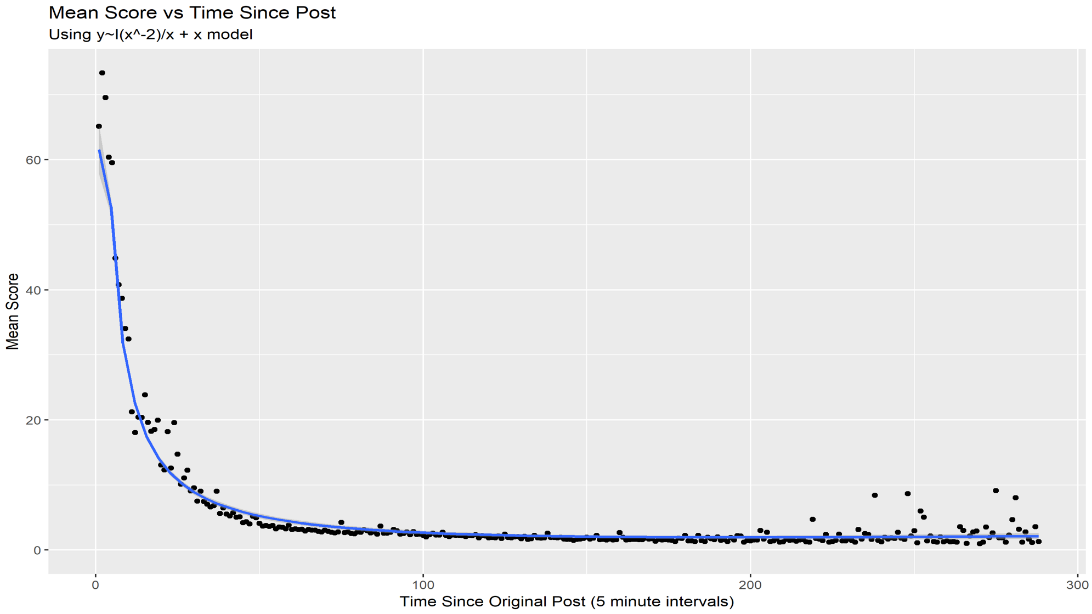

Figure 3.1: *The time since the Original post was sent versus the mean score of every post within that 5 minute interval. Posts closer to post time tend to garner a much higher score. With this said, there is some outliers that begin to form near the tail end of the line. that beat the average.*

3.3.2 Relationship between length of comment and post score

This was tested to see if there was any correlation between the length of the comment measured in number of characters, vs the posts score which is an integer value. What we can see in the graph below is that there is no correlation between a posts score and the length of the comment. From this we can assume that a comments score is determined by the content of said comment, which is likely to vary from each subreddit. For example a posting a meme in /r/politics may not get the same reception that it would in /r/DankMemes as the subculture of politics orients itself to be for more serious conversations compared to /r/DankMemes which is all about humour.

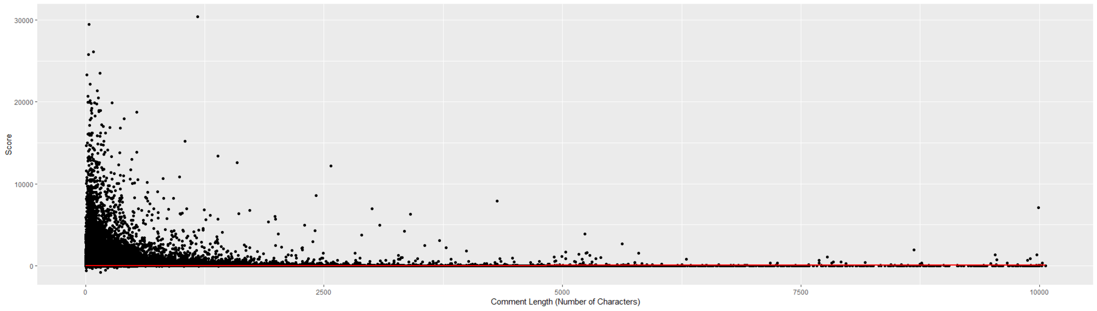

Figure 3.2: *The length of a comment versus the mean score of all comments of that particular length*

3.4 Exploratory Text Analysis

3.5 Comment features

3.6 Machine learning experiment

3.7 Conclusions

Chapter 4

Results

4.1 Conclusions

Chapter 5

Discussion

5.1 Conclusions

Chapter 6

Conclusion

This chapter summarises the main outcomes and conclusions resulting from this body of work.

6.1 Conclusions

The main conclusions that may be drawn from the body of work.

6.2 Future Work

Further development that could be carried out in the future.

Bibliography

[1] Thukral S, Meisheri H, Kataria T, Agarwal A, Verma I, Chatterjee A, et al. Analyzing Behavioral Trends in Community Driven Discussion Platforms Like Reddit. In: 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 2018. p. 662–669.

[2] Robertson A. Reddit bans "Fat People Hate" and other subreddits under new harassment rules. [online]. 2015;Available from: https://www.theverge.com/2015/6/10/8761763/reddit-harassment-ban-fat-people-hate-subreddit.

[3] Van Mieghem P. Human Psychology of Common Appraisal: The Reddit Score. IEEE Transactions on Multimedia. 2011 Dec;13(6):1404–1406.

[4] Glenski M, Pennycuff C, Weninger T. Consumers and Curators: Browsing and Voting Patterns on Reddit. IEEE Transactions on Computational Social Systems. 2017 Dec;4(4):196–206.

[5] Woolley S. Automating power: Social bot interference in global politics. First Monday. 2016;21(4). Available from: https://uncommonculture.org/ojs/index.php/fm/article/view/6161.

[6] Kilgo DK, Ng YMM, Riedl MJ, Lacasa-Mas I. Reddit's Veil of Anonymity: Predictors of engagement and participation in media environments with hostile reputations. Social Media + Society. 2018;4(4):2056305118810216. Available from: https://doi.org/10.1177/2056305118810216.

[7] Segall J, Zamoshchin A. Predicting Reddit post popularity. nd): n pag Stanford University. 2012;.

[8] Horne BD, Adali S, Sikdar S. Identifying the Social Signals That Drive Online Discussions: A Case Study of Reddit Communities. In: 2017 26th International Conference on Computer Communication and Networks (ICCCN); 2017. p. 1–9.