Characteristics of a Successful Reddit Post

Stephen Smart, Emily Southern, Shruthi Sukumar, and Adam Winchell*

University of Colorado Boulder – INFO 4602/5602 – Spring 2018

Abstract

The goal of this project was to determine the characteristics of a successful Reddit post. We analyzed data from eight subreddits consisting of text-only posts. We created three visualizations to assist Reddit users interested in generating the most karma. The first visualization shows the mean karma score for different topics across all of our subreddits. The second visualization shows the number of posts within each subreddit, as well as the number of those posts above and below a given karma threshold. The third visualization shows the mean karma score (normalized by the number of active users on that subreddit) over time. In addition to supporting subreddit analysis, these visualizations also tell interesting stories about what happens within these subreddits.

Keywords: subreddit, topic model, D3.js

1 Introduction

Reddit is a popular social news and media aggregation website. Registered users submit content, such as text posts, links, and images, to user-created communities called 'subreddits'. These communities cover a wide range of topics including, but not limited to, news, science, movies, video games, music, books, and food. Some subreddits allow for text-only submissions, which we focus on here. Specifically, we chose to analyze a collection of subreddits dedicated to posts asking focused questions. These are communities like AskHistorians, AskScience, and AskCulinary, where members of the subreddit pose a question specific to the given community and experts within the community collectively provide their perspectives on the answer.

Each user who views a post can choose to either up-vote (i.e. like the post) or down-vote the post. The sum of the up-votes and down-votes, treated as +1 and -1, respectively, is defined as the karma of the post, which serves as a proxy for the success of a post. We aim to determine the characteristics of a successful text-only post in various Ask-based subreddits. These findings can be used to assist Reddit users interested in generating the most karma.

2 Related Works

Our research began with trying to understand Reddit's algorithm in displaying posts [7]. Reddit's core object is to prioritize new posts as to keep content fresh for users; as a result, newer posts are prioritized over older ones.

Due to Reddit's relative obscurity, we looked for publications about a different social media platform, Twitter, as the two are comparable for our intents and purposes. To that end, we found papers [3, 5, 8] that have had modest success in predicting the popularity of tweets on Twitter. These methodologies incorporated the use of the content of the messages, temporal information, and meta data of the tweets for use in prediction. In [8] the authors use early retweeting data in order to anticipate the popularity of a tweet; the Reddit equivalent would be the comments on a post. Unfortunately, such a strategy would not necessarily work with the subreddits analyzed due to the strict moderation of which comments are considered relevant to the post at hand.

D3 is a JavaScript library that allows developers to map objects to the DOM [2], and is a powerful tool in visualizing data online. We wanted to create an experience where users could explore what it means to create a successful Reddit post. Thus, D3 was the obvious candidate for allowing the types of interactions we desired. Since our visualizations include animations, we researched whether this could possibly detract from the user's ability to analyze trends and determine what comprises a successful Reddit post. The work done by [6] showed that while animations can detract from visualizations' effectiveness, it comes with the boon of being more fun and engaging. Thus, one must strike a balance in the use of animations, as to not detract too much from the effectiveness of the visualization.

Data storytelling is more effective when visualizations are memorable and engaging [4]. Thus, we included animations at the potential cost of effectiveness and chose not to modify some of the less sensible topics recovered in the topic modeling. In regards to [1], Latent Dirichlet Allocation (LDA) has been shown to be a profoundly useful tool in the analysis of large corpuses of text. LDA is an unsupervised machine learning method that takes in a corpus of documents and returns a set of topics that describe the corpus. Further, each document may then be assigned a distribution of topics that describes the content of the document.

3 Description of Project

We created three visualizations to help Reddit users understand the characteristics of a successful subreddit post. Visualization 1 shows the topics of successful posts in a given subreddit. It also provides information about which subreddit it is easiest to be successful in. Visualization 2 helps users decide which time of day to post in order to generate the most karma. Visualization 3 allows the user to define the karma threshold that distinguishes between a "good" and a "bad" post, and then displays the number of posts in each subjective category for each subreddit.

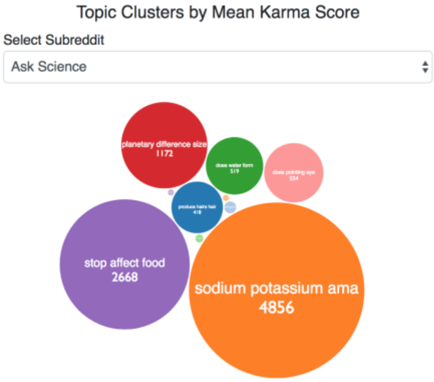
We collected data from eight Ask-based subreddits using Reddit's provided Python API. We collected data over course of

* adam.winchell@colorado.edu

one week from the following subreddits: AskHistorians, AskAcademia, AskAnthropology, AskCulinary, AskMen, AskScience, AskSocialScience, AskWomen. The process consisted of checking each subreddit every hour for new posts and recording the time the post was created (UTC), the number of users on the subreddit at the time of the post, and the title of the post. We recorded the karma of the post 24 hours after the post was made.

3.1 Visualization 1

Our first visualization is a bubble chart showing the different topics of posts in a given subreddit, generated using topic modeling.


Topic Clusters by Mean Karma Score

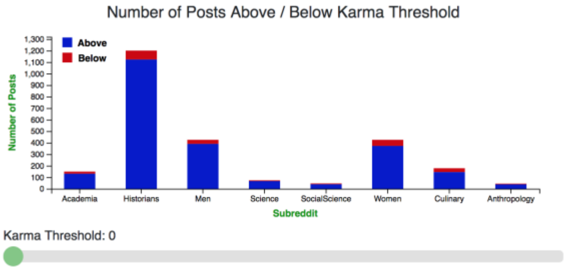
Select Subreddit: Ask Science

Figure 1. Screenshot of visualization 1: topic clusters by mean karma score

The user can choose which subreddit to visualize by selecting it from the dropdown menu above the bubble chart. Each different colored bubble represents a different topic, which is displayed within the bubble along with the average karma score for posts that contain one of the words from the topic. The size of the bubble corresponds to the average karma of the posts included in that topic. Each topic contained ten words total, and we elected to display the top three words within the each bubble. We also used a redundant encoding (position and color) to distinguish topics. The size of the topic labels and mean karma values are scaled according to the size of the bubble they are contained in.

3.2 Visualization 2

Our second visualization is a stacked bar chart displaying the number of posts above and below a given karma threshold for each of the subreddits. The user can change the threshold that differentiates between a “good” and “bad” post using the slider tool below the visualization. Every time the slider changes, the visualization animates by adjusting the number of posts that are greater than or equal to the specified karma threshold. The colors chosen were a medium blue (#00D HEX color code) to represent the number of posts greater than or equal to the karma threshold, and a medium red (#D00 HEX color code) to represent the number of posts less than the karma threshold. Red was chosen to represent posts will low karma since red is often associated with negative values or failure, although we do recognize that this is not true across all cultures. Blue was chosen to prevent color vision deficiencies from causing any problems.

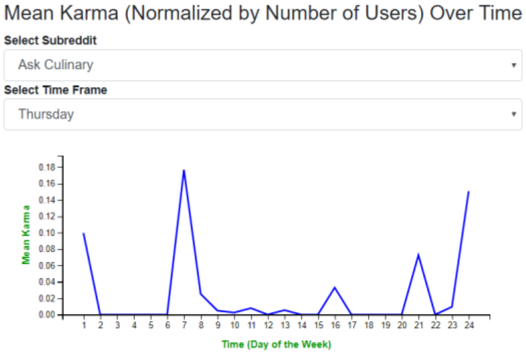

Number of Posts Above / Below Karma Threshold

Karma Threshold: 0

Figure 2. Screenshot of visualization 2: number of posts above / below karma threshold

3.3 Visualization 3

Our third visualization is a line graph showing the popularity of posts in a given subreddit over time. The horizontal axis represents time. The user can choose to visualize the entire week of data (Monday-Sunday) or all 24 hours of one specific day of the week using the dropdown menu. The vertical axis is the mean karma over all posts made within the time frame normalized by the number of users on the selected subreddit. When the user selects to “drill down” in order to view one of the days instead of the whole week, the line graph animates from the week view to the day view and vice-versa.


Mean Karma (Normalized by Number of Users) Over Time

Select Subreddit: Ask Culinary

Select Time Frame: Thursday

Figure 3. Screenshot of visualization 3: mean karma (normalized by number of users) over time

4 Discussion

Our system of visualizations was designed to determine the characteristics of a successful Reddit post.

From our first visualization, we see that posts in the top ten topics of some domains have fairly similar average karma scores (equal-sized bubbles) while posts in the top ten topics of other domains have a greater spread of average karma scores (a mix of very large and small bubbles). Posting to a subreddit with topics that have similar average karma scores is a ‘low risk, low reward’ situation. For example, a post in any one of the topics for AskHistorians would be likely to receive a decent karma score

(48-73). Alternatively, posting to a subreddit with topics that have a large range of average karma scores is a ‘high risk, high reward’ situation. For example, a post in the AskScience subreddit could be included in a topic with a very low average karma score (less than 10) or a very high karma score (greater than 1000).

From our second visualization, we are able to see the distribution of posts across the domains that we analyzed. Additionally, as we increase the karma threshold, the number of posts above the threshold (shown in blue) decreases and the number of posts below the threshold (shown in red) increases. It is interesting to note that the chart changes much more drastically at lower karma thresholds than at higher karma thresholds; when moving the slider from high to low, graphical changes are nearly imperceptible until reaching a karma threshold of about 20. A Reddit user may wish to post in a subreddit with very few posts, such as AskSocialScience or AskAnthropology, because there is less competition for karma. Alternatively, it is possible that posting in a subreddit with many posts, such as AskHistorians, may have more users willing to give karma to new posts.

From our third visualization, we see a consistent trend of more successful posts later in the week (Thursday through Sunday) across all domains. For example, when looking at the entire week of posts in the AskCulinary subreddit, we see that there is a peak on Thursday. If we drill down and just focus on Thursday, we see peaks in the morning and late at night. It is possible that posts made late at night or very early in the morning are more successful than posts made during more popular Reddit user hours because they stand out due to less competition at those times.

5 References

[1] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." Journal of Machine Learning Research 3:993-1022, 2003.

[2] Michael Bostock, Vadim Ogievetsky, and Jeffrey Heer. "D³ data-driven documents." IEEE Transactions on Visualization and Computer Graphics 17(12):2301-2309, 2001.

[3] Liangjie Hong, Ovidiu Dan, and Brian D. Davison. "Predicting popular messages in twitter." Proceedings of the 20th International Conference Companion on World Wide Web, pages 57-58, 2011.

[4] Robert Kosara. "Presentation-oriented visualization techniques." IEEE computer graphics and applications 36(1):80-85, 2016.

[5] Jey Han Lau, Nigel Collier, and Timothy Baldwin. "On-line trend analysis with topic models: #twitter trends detection topic model online." Proceedings of COLING 2012: Technical Papers, pages 1519-1534, 2012.

[6] George Robertson, Roland Fernandez, Danyel Fisher, Bongshin Lee, and John Stasko. "Effectiveness of animation in trend visualization." IEEE Transactions on Visualization and Computer Graphics 14(6):1325-1332, 2008.

[7] Amir Salihefendic. "How Reddit ranking algorithms work." Hacking and Gonzo, 2015.

[8] Tauhid Zaman, Emily B. Fox, and Eric T. Bradlow. "A Bayesian approach for predicting the popularity of tweets." The Annals of Applied Statistics 8(3):1583-1611, 2014.