

Predicting Comment Popularity within Reddit Communities through Text and Metadata based
Multiclass Classification Models

(Technical Paper)

Exploring the Influence of Reddit Mechanics on Content Popularity Factors and the
Counterinfluence of User Activity on Content Popularity Factors and Reddit Mechanics

(STS Paper)

A Thesis Prospectus Submitted to the

Faculty of the School of Engineering and Applied Science
University of Virginia • Charlottesville, Virginia

In Partial Fulfillment of the Requirements of the Degree
Bachelor of Science, School of Engineering


Siddharth Nanda
Spring, 2020

Technical Project Team Members
Cory Kim

On my honor as a University Student, I have neither given nor received
unauthorized aid on this assignment as defined by the Honor Guidelines
for Thesis-Related Assignments

Signature  Date April 22, 2020
Siddharth Nanda

Approved _____ Date _____
Rich Nguyen, Department of Computer Science

Approved  Date 10/15/2020
Toluwalogo B. Odumosu, Department of Engineering and Society

Introduction

The last two decades have seen an explosive increase in popularity for social media websites: websites that enable online communication of information, ideas, and other content through virtual communities (Merriam-Webster, n.d.) (Alexa Internet, Inc., n.d.). Social media websites have a few key traits including: some structure for organizing communities, some structure for sharing content with other users, and some metric for indicating the community reception to shared content. Analysis of these websites from a technical and sociological perspective could help provide insight into the evolution of the mechanics of interaction between persons. The STS research proposed analyzes the results of a technical research project that investigates the characteristics of popular content on social media website Reddit, the 6th most popular website in the US (Alexa Internet, Inc., n.d.), through the use of Machine Learning methods. Prior to the articulation of the specific of the STS research question and the methodology used to investigate it: it is important to understand the specifics and basis for the technical project.

Research Context

Reddit is a social media website that has somewhat of an unconventional structure compared to its peers. Founded in 2005 by UVA students Steve Huffman and Alexis Ohanian, Reddit attempts to focus user interactions around user submitted content. Content such as links, text, or video can be shared through two means: the creation of an individual discussion space for the content (“thread”), or a reply to some existing content (“comment”). Virtual communities are organized through self-selecting groups (“subreddits”), which a user can join or leave at will, and groups serve as a collection of threads of a related topic. The typical Reddit user can choose to interact with others simply by observing content they’re interested in (browsing a subreddit, its threads, and the comments contained within those threads), by actively interacting within a community (joining a subreddit, posting threads to it, and commenting on threads), or through a

combination of the described observation/interaction mechanics. Reddit users can indicate their reception to posted content simply by clicking an up-arrow (indicating a positive response) or down-arrow (negative response) available on each thread and comment. Reddit displays the cumulative value of all upvotes and downvotes as a score on each thread and comment. For content that provokes a strong positive reaction, Reddit offers the option for users to purchase “awards” for content which are then displayed prominently. The pricing of these awards varies: and the more expensive the award given, the more benefits the user who posted the content that received the award enjoys. Reddit is also unique in that users are partially anonymous: aside from a selected username, users are not required to provide personal data that could be used to identify them offline although some choose to. Reddit’s communities are moderated by users themselves and each subreddit can have unique rules and restrictions as determined by its moderators. For example, some subreddits do not allow new users to post, requiring them to acquire a certain amount of karma (a user metric depending on the overall popularity of a user’s content) before posting in that community. Karma mechanics have allowed “influencers”, users that generate popular content and are closely followed by other users, to rise on Reddit as well. Therefore, despite the somewhat basic mechanics of the website (subreddits, threads, comments, upvotes, awards, and karma): this social media website has a reputation of being difficult to interact with due to its diverse subcultures and user generated exclusivity.

Technical Project Description

In the technical project, comment data was gathered from a certain number of threads on selected subreddits using a researcher created tool. The data collected included: comment text, the time a comment was created, the time difference in when the comment was created and when the comment or thread it was replying to was created, the depth level of the comment (how far down a comment was in a chain of comments), the score of the comment, and the number of awards the comment had received. Certain characteristics of content were derived from the

collected data including: the comment length in words, the comment length in characters, and the sentiment score of the comment (how positive or negative of a tone the comment took). Machine Learning methodologies (essentially automated statistical methodologies) were used to create various classification models from the characteristics of the data collected that could be used to classify data into score groups (stratified by popularity). Models created solely from either the text data itself or solely from metadata (such as length, time difference, depth, awards, etc.) both demonstrated acceptably high, as well as comparable, accuracy in the classification tasks (predicting the score group the input data belonged to).

STS Project Description

My STS research question proposed attempts to supplement the results of the technical project, asking: "How do Reddit's mechanics cause certain factors to be more important than others in creating popular content on reddit and how does activity on Reddit influence those same factors and mechanics?". In order to understand the scope of the STS research: it is important to identify the frameworks that will be used to answer those questions, the various actors within the context of the question, the importance of the question itself, and the timeline for the research project. (Latour, 1992)

The two frameworks that will be used answer the STS research question are the notion of "Configuration and Script" (N.E.J. Oudshoorn, 2003) and "Actor-Network Theory" (Latour, 1992). The former approach focuses on the process of defining the users, setting constraints on their actions, and how the design of objects conceptualizes a method of their use. This framework can be used to understand the usage of Reddit, how its creators intended to have users interact on the website, the intended usage of site mechanics such as upvotes/downvotes and subreddits, and how users ultimately interpret the design to enable their unique forms of usage. The latter framework, Actor-Network Theory, attempts to form an understanding of users and website mechanics as "actors" within a "network" where actors (both living and non-living) can

exert influence on each other. This framework can be used to understand elements of Reddit culture and how Reddit users affect each other and the development of the site itself. The understanding derived from analysis through the configuration and script framework could help explain why certain features depending on site mechanics themselves such as awards or depth level have a certain weight in determining overall popularity while the second framework could help explain how certain features depending solely on the user decisions (comment length, word choice, etc.) have an effect on overall popularity.

The actors within the context of the question can be divided into two groups: living and non-living. The first consists of many sets of individuals including designers of the website, moderators, influencers, users who post, and users who interact with content (upvote/downvote) but do not post. All of these subgroups have unique characteristics that affect how they interpret Reddit and each other. The second group consists of the content posted to Reddit as well as the existing mechanics of the website.

STS Project Timeline

Finally, it is important to consider a timeline for the completion of this research project. First, two to three weeks will be used to define the context of the problem as well as detail various actors and actants. Next, three to four weeks will be used to analyze the question under the Configuration and Script framework. Similarly, three to four weeks will be used to analyze the question with the Actor-Network Theory. After the 10-week mark, a week or so will be spent reevaluating the technical project and forming deeper connections between the results of the technical project and the STS investigation. Two weeks will be spent completing the research paper detailing the STS project and any remaining time will be used to revise.

References

Alexa Internet, Inc. (n.d.). *Top Sites in United States*. Retrieved from Alexa:

<https://www.alexa.com/topsites/countries/US>

Latour, B. (1992). Where are the missing masses? The sociology of a few mundane artifacts. In W. E. Law, *Shaping Technology/Building Society: Studies in Sociotechnical Change* (pp. 225-258). Cambridge, Massachusetts : MIT Press.

Merriam-Webster. (n.d.). *social media*. Retrieved from Merriam-Webster: <https://www.merriam-webster.com/dictionary/social%20media>

N.E.J. Oudshoorn, T. P. (2003). Introduction: How users and non-users matter. In T. P. N.E.J. Oudshoorn, *How users matter. The co-construction of users and technology* (pp. 1-25). Cambridge, Massachusetts: MIT Press.