Thesis Project Portfolio

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

(Technical Report)

Exploring the Influence of Reddit Mechanics on Content Popularity Factors and the Counterinfluence of User Activity on Content Popularity Factors and Reddit Mechanics
(STS Research Paper)

An Undergraduate Thesis

Presented to the Faculty of the School of Engineering and Applied Science

University of Virginia • Charlottesville, Virginia

In Fulfillment of the Requirements for the Degree

Bachelor of Science, School of Engineering

Name: Siddharth Nanda

Signature  Date: 12/02/2020

Fall 2020

Department of Engineering and Society

Table of Contents

The ability to individuals to communicate with one another and the methods by which they do so have experienced great change with the rise of social media websites. In recent years, online “hive minds” have become increasingly concerning phenomena on social media and present concerns about disinformation and online radicalization. This set of studies can be viewed as an attempt to understand the development of factors that influence online popularity via sociotechnical analysis of Reddit: a partially anonymized social-network with a unique set of popularity mechanics including direct voting.

The technical research, entitled “Predicting Comment Popularity within Reddit Communities through Text and Metadata based Multiclass Classification Models”, focused on using machine learning classifiers (automated statistical methods) to predict content popularity and identify content popularity factors on Reddit. Data on and relating to comments was collected and derived using an original, extensible data-collection tool built around an existing API wrapper for Reddit including: 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 (time difference from parent comment), how far down a comment was in a chain of comments (the depth level of the comment), the score of the comment, the number of awards the comment had received, the comment length in words, the comment length in characters, and the AFINN sentiment score of the comment (how positive or negative of a tone the comment took). Score categories were derived by scaling and stratifying raw comment scores and various machine learning classifiers were trained on these features and used to predict the score: achieving a high degree of accuracy and identifying the time difference in when the comment was created and when the comment or thread it was replying to was created, comment length (both character and word), and sentiment score of the comment (how positive or negative of a

tone the comment took) as features that served as strong content popularity predictors. The STS research, entitled “Exploring the Influence of Reddit Mechanics on Content Popularity Factors and the Counterinfluence of User Activity on Content Popularity Factors and Reddit Mechanics”, focused on analysis of the results of the technical research project through the use of two STS frameworks: Configuration and Script; and Actor-Network Theory. Configuration and Script 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; meanwhile Actor-Network Theory forms 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. The former framework was used to contextualize the technical research as a study in how the mechanics (commenting, subreddits) influence user behavior (voting, commenting), and ultimately asserted that repeated usage of popular “scripts” defines a culture (essentially that common usages of the commenting mechanic have a strong influence on subcultures). The latter framework helped provide commentary on the counterinfluence of users and other actors on the social network on mechanics and ultimately asserted that when networks share a goal, actors influence each other to assert a culture (popularity on a social network is self-driving in identifying and building popularity factors in that popularity is the goal of a social network).

These two studies were somewhat unique in that the STS research question was directly coupled with the technical project: as the technical research was conceived and completed prior to the identification of the STS research question. The direct influence of the technical research on the STS research allowed for the STS research to serve as an important complementary piece to the technical research and significantly enrich the overall value of the pair of studies. The work can serve as a basis for other studies focused on social media popularity.