Analysis of popularity of News Links in Reddit Communities

Bharadwaj Janarthanan

I. INTRODUCTION

Humans are fundamental elements of the society. Humans form groups that are macro units of information sharing network. Groups differ in their preferences and beliefs. What might seem popular among one group, might not be for another. Groups can even have conflicting behaviours.

With rapid adoption of online social discussion forums, more data about interaction in human networks have been made available. Reddit dubbed as the front page of the internet is a popular information sharing platform, with information sharing predominantly being within communities(subreddit). Subreddits are dedicated to specific areas and its members discuss and exchange information relevant to their common interests. Each subreddit, is governed by moderators and have policies and rules defined for content posting. Members upvote or downvote posts and comments for several reasons such as content credibility, controversiality, specificity and relevance to name a few. In this project, we would like to broadly explore factors associated with news content popularity across communities and how communities under same domain with different levels of moderation vary in their definitions of popularity.

Since, reddit can be viewed as a collection of groups with members cross linked across communities in membership, it forms a good approximation to complex communication chains existing across human groups in real life. For this project we would like to use observational data on news website submissions from subreddits to analyse popularity in context of inter and intra news sharing communities.

We wish **to explore the trends in popularity among different news sharing communities on Reddit.** Further, we want to compare popularity of news from same source among different communities. More specifically, we wish to observe **if different communities perceive news from same source different based on the source popularity within the community**. Additionally, we wish to study how moderation of communities/ governance influence news popularity.

II. RELEVANT WORK

In this section we briefly discuss, and review work relevant to our proposal. Our main goal is to find commonalities and distinctions in news popularity trends across communities.

[1] presents detailed analysis to predict popularity of a submission within a subreddit based on the number of comments it would have. Extensive work has been done in data collection and feature extraction. Popularity was shown to be a temporal phenomenon and features used were broadly user, community, content and title centric. Popularity was predicted in two stages- initial popularity versus long-term popularity and it was shown that initial submissions get more attention over subsequent submissions due to novelty in content. However, the author doesnt explore commonality in popularity patterns across communities.

[2] presents extensive work in predicting which comments receive high attention within subreddit communities, based on score of comments. A comparison study of feature importance across different subreddit models then reveal, what are the common and distinctive factors that are characteristic of popularity in different communities. It is shown that, popularity is different even among similar communities. Sentiment and relevance of content is identified as key characteristics, while extent of temporal effects differ across communities. Moderation doesnt always alter popular behaviour in communities.

Our proposed exploration is different from other relevant work ([1], [2] and [3]), surveyed in that, we focus specifically on news source popularity in communities and we wish to study how factors influencing news popularity varies across different communities with different levels of moderation. Also, we would analyse submissions and comments together and not independently, like done in most of the literature we surveyed.

III. PROPOSED METHODOLOGY

In this section, we discuss our analysis plan. We use an observational study approach to compare trends and patterns in news popularity across subreddits sharing news links of top 5 news websites based in US, as of May 2018. [6]

A. Data Collection and Feature Extraction

All Reddit data is available as dumps on a monthly basis from [7]. For the purpose of our analysis, we only use November 2016 dump for our initial phase. Our intuition behind choosing November 2016 for our initial analysis is to study news source popularity during U.S presidential elections. We believe a lot of news links from different sources were shared across communities around that time. We filter subreddit submissions that have one of the top 5 news links [6] mentioned in the submission and their relevant comments. We would look only for normalized URL and query parameters if any would be ignored for the purpose of our analysis.

Our initial feature set consideration includes, content sentiment, content topic, number of comments, relevance of comments to post topic, relevance of posts to subreddit domain and the time of submission/ post.

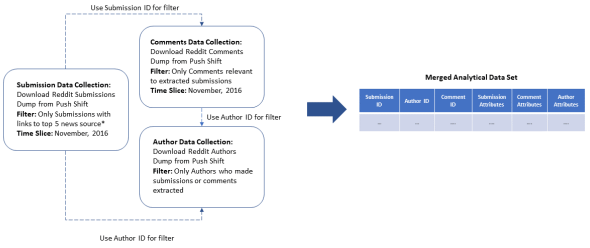

Fig. 1. Data Collection Process

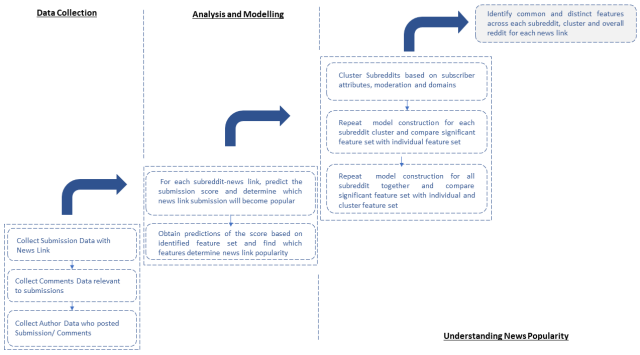

Fig. 2. Flowchart depicting our proposed methodology

B. Analysis and Modelling

We would need to assess sentiments of submissions and comment threads to assess the sentiment of submission and distribution of sentiments in comments. A Nave Bayes Classifier would be used for this purpose to be able to tag each of the submission and comment to a sentiment. Following this, we will use an LDA model to identify the theme of each submission or comment from an identified set of topics. These developed features along with cosine-based distance metric of comment to submission and submission to subreddit measuring content similarity and content flair, tags provided by authors and moderators for content is used as features into a regularized regression model to predict the score for a submission.

We then introduce temporal effects into the model and observe, if popularity depends on the timing of a submission. Subreddits are clustered into 10 different groups based on number of subscribers, karma of participants of subreddit and level of moderation, which is later used for group comparisons. Each subreddit, news source pair has its own model and each subreddit group has its own model and we construct an overall reddit model.

Our entire dataset would be split into 60% training, 20% validation set for model tuning and remaining 20% hold or test set for results validations. Do note that since we incorporate day of week and hour of day effects into predicting popularity, we will use a sliding window k-fold validation technique.

C. Comparing News Link Popularity

Now that we have decoded popularity definition across subreddit communities, groups of subreddit and overall reddit, we then look at how similar are these definitions (feature sets) through a heatmap. This way we would be able to tell, what features are intrinsic to communities and what features of popularity are an effect of moderation, number of subscribers and their activity levels and what features of popularity are universally valid across all communities.

D. Project Proposed Timeline

This section provides estimated time to be spent on each task for this project.

Week	Task
1 - 02/24	Data Summary, Data Cleaning and Wrangling
2 - 03/03	Submissions and Comments Sentiment Analysis
3 - 03/10	LDA model for identifying relevance
4 - 03/17	Feature extraction and regularized regression
5 - 03/24	Profile based clustering of communities and comparing popularity
6 - 03/27	Project Report and Presentation

E. Progress Update

Of the total submissions in Nov 2016, a total of 42,093 submissions were linked to one of the top 5 news source links in the study and came from a total of 1,787 different subreddits. Of, the total subreddits with a submission with a link to the news source, only 4% of them had more than 30 submissions. We restrict our analysis to this 4% of data and only look at 37,081 submissions made during Nov 2016. We then download, comments relevant to the submissions in the consideration set and the authors information relevant to the submissions only.

References

[1] Katyaini Himabindu Lakkaraju, Demystifying content popularity on Reddit,

[2] Benjamin D. Horne, Sibel Adal, and Sujoy Sikdar; Rensselaer Polytechnic Institute, Identifying the social signals that drive online discussions: A case study of Reddit communities, 2017 26th International Conference on Computer Communication and Networks (ICCCN)

[3] Mohamed Ahmed, Stella Spagna, Felipe Huici, Saverio Niccolini, A Peek into the Future: Predicting the Evolution of Popularity in User Generated Content

[4] Eshwar Chandrasekharan, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein, and Eric Gilbert. 2018. The Internets Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. Proc. ACM Hum. -Comput. Interact. 2, CSCW, Article 32 (November 2018), 25 pages. https://doi.org/10.1145/3274301

[5] M. J. Salganik, P. S. Dodds, and D. J. Watts, Experimental study of inequality and unpredictability in an artificial cultural market, Science, vol. 311, no. 5762, pp. 854856, 2006.

[6] Most popular news websites in the United States as of May 2018, ranked by unique monthly visitors, https://www.statista.com/statistics/381569/leading-news-and-media-sites-usa-by-share-of-visits/

[7] PushShift Data Dumps, https://files.pushshift.io/reddit/