Analysing Distinctions in Popularity of Different Contents using Reddit

Bharadwaj Janarthanan
Department of Computer Science, University of Torronto
bharadwaj@cs.toronto.edu

Abstract— **Creating and delivering content effectively is a challenging problem to solve. To enable effective content creation, it is important to understand what factors are associated with the popularity of different types of content. Popularity of content depends on presentation style, content relevance, author popularity, size of target audience and time of delivery. In this study, we used observational data from Reddit posts during November 2016 to analyse popularity of submissions across domains in news, social media, blogs and user-generated independent content. Specifically, we address the research question, are different content- self generated vs blog links vs news links, popular due to similar influences of user, community, content and time aspects. We proposed a modelling framework to predict popularity of submissions based on a diverse set of features such as content type, similarity to other submissions, length of submission, time of submission, user popularity, community activity and level of moderation in community. While user and community influences are similar across different domains, content-based features have varying level of influences across domains. News submissions are more popular when their content structure is like other submissions in the news domains, while user-generated content is more popular when they are more unique in their content.**

I. Introduction

With the rapid adoption of content sharing websites people share contents of their choice and vote or comment on contents shared by other or those that are of theirs. Reddit is one such content sharing platforms in which users can post content, upvote or downvote content and comment on posts made by others. Users post their content on sub-groups/ social communities called subreddits which are dedicated to share information of a specific type as described on the subreddit page. Subreddits are moderated and reviewed for content through policies enforced by Reddit and subreddit moderators. Reddit posts/ submissions are of two types mostly, they are either links to third party content or self-generated content by the user.

Now that we have such a rich source of content sharing open source data, it would be interesting to study and observe what aspects of a content causes it to be virally discussed in societies. To be more specific, do different content types have different aspects that influence its virality, how should different types of content be structured and placed on social communities such as subreddits such that they are more popularly discussed within the community. Does moderation influence the extent to which contents of different type can be discussed on communities? Does subreddit member community perception about content influence popularity?

In this project, we conducted a preliminary research to examine factors such as content similarity, content length, time of submission and other such features in relation with the number of comments a submission receives. This study provides us with a better understanding of associations between content structure, member influence, community moderation and content popularity.

Our main research question in this study was, **Are different content- self generated vs blog links vs news links, popular due to similar influences of user, community, content and time aspects?** We developed models to predict content popularity based on the number of comments it received. We also compared across different content types as to what features influence popularity.

II. Relevant Work

In this section we briefly discuss, and review work relevant to our research. Our main goal in this study was to find commonalities and distinctions in popularity trends across news links, blog posts, social media posts and self generated content in Reddit. Based on our survey of existing literature on reddit analysis researchers have studied submission popularity among communities and compared them across select subreddits/ communities. However, to our best of knowledge there hasnt been any research on how different content types structural aspects influence its popularity on reddit in terms of the number of comments it receives.

In [1] Lakkaraju performed detailed analysis to predict popularity of a submission in a subreddit, based on the number of comments it received. 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 and distinctions in popularity patterns across different domains from which content originated. A further formal extension of this study has been performed by the same author in [7].

In [2] Benjamin et. al. present extensive work in predicting which comments and submission receive high attention within subreddit communities, based on the number of votes each comment receives. A comparison study of feature importance across select subreddit models showed which factors are common and distinctive in determining the popularity among 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. Moder-

ation doesnt always alter popular behaviour in communities. Our proposed exploration is unique from other work on content popularity analysis in that, we study and compare popularity of different content types and identify factors influencing popularity of different styles of content such as self generated vs social media vs blogs vs news websites.

III. Analysis Methodology

We considered Reddit.com for analysing aspects of different types of content that are associated with the content going popular among communities. Users make submissions on subreddit, that are either self-generated or are links to third party content creator websites. Larger the user community of a subreddit, more the visibility and more popular the submission would be. Peoples sentiment about the submission are captured in the form of votes- upvotes or downvotes which in turn determines if a submission should be displayed on the top of the page, which further encourages more visibility and discussion on the submission as a result. Content submitted to communities can be images, video link, article links, social media posts, news articles, self-generated content by user, GIFs. Every subreddit has policies for content submission and moderators remove posts that dont comply.

A. Dataset Description

For the purpose of this analysis we restrict ourselves to Reddit submissions data made during November 2016. Pushshift.io has made available monthly reddit submissions data [7] for download and we would use the same for this research. Our choice of November 2016 period was based on our interest in studying what cahnnels of news content were popular during U.S Presidential elections. We only considered subreddits with more than 1000 submissions and 100 users for our analysis to ensure that we studied content popularity associations only in communities that had a reasonable size and activity during November 2016.

We only study the top domains with most submissions during November 2016. These were majorly News sharing websites along with blogging sites and social media websites. We also included self-generated content in our analysis, to see what aspects of the content from different origins/ domains are associated with its popularity in communities on Reddit.

Further, we excluded submissions that had images, video links, GIFs or any other non-textual data in them. Also, submissions from deleted authors and submissions that were removed or deleted were excluded from the analytical dataset.

In total we have 2,177,368 submissions made on 1,010 different subreddit communities by 858,586 Users. The content was identified to be from one of the 12 domains (ref: Table 1 shows distribution of submissions across domains),

News Websites: miamiherald.com, washingtonpost.com, theguardian.com, nytimes.com, cnn.com, huffingtonpost.com

Blog Sites: blogspot.com, wordpress.com, wikipedia.com

Social Media Websites: twitter.com, facebook.com

Self-Generated: reddit.com As can be seen from Figure 1, most submissions across different domains receive fewer

TABLE I: Total Submissions across Domains on Reddit for November 2016

News Links	Total Submissions
theguardian.com	6712
nytimes.com	6634
washingtonpost.com	5370
cnn.com	5004
huffingtonpost.com	4275
miamiherald.com	218

Blog Sites	Total Submissions
blogspot.com	12664
wordpress.com	10721
wikipedia.org	6088

Social Media Websites	Total Submissions
twitter.com	55428
facebook.com	8307

Self-Generated	Total Submissions
reddit.com	2742460

comments with blog sites mostly having no comments on their submissions except for Wikipedia. On an average across submissions in different communities that originate from any of the domain, a submission by a User receives about 10 comments. Our analysis will be focussed on studying submissions that receive one or more comments and what aspect of the submission content is associated with getting more users to comment on the submission.

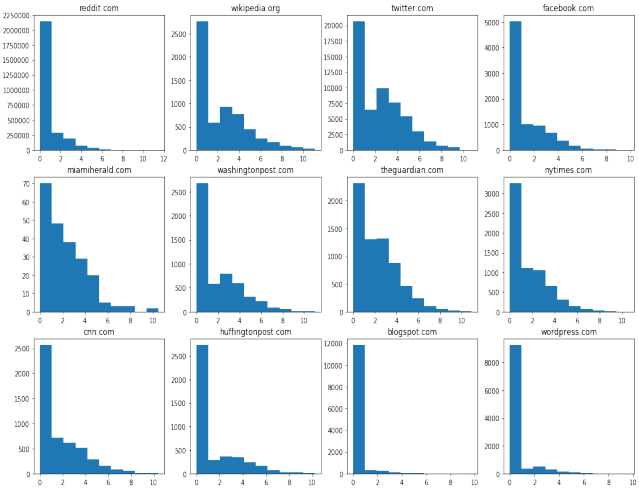


Fig. 1: Histogram of number of comments a submission receives across different domains

B. Feature Extraction

In order to understand factors that influence submission popularity we developed features based on content, community, user and time of a submission. In this sub-section, we describe in detail features used in analysis and how they relate to submission popularity.

Note: All features were tested for significance of their relationship with number of comments and found to be significant in data used.

1) Content based

Content based features are developed to identify how the quality and relevance of a submissions content to a subreddit community influences its popularity. In reddit, submission content that receive the most upvotes are featured on top of the page, thereby providing more visibility to submission and facilitating more discussion on submission, which in turn increases the number of comments.

*i. Content Type:*In this feature set, we examine if the submission has content that is restricted to only users over 18

years of age to see how the target user restriction influences the number of comments a submission receives. We also study how submissions with flair tags vary in number of comments they received from the submissions that dont have any tag on them.

ii. Content Length: The length of the title and description (only for self-generated submissions) could suggest how specific and descriptive is the submission about the subject it is talking about. We examined how the submission comments were influences by the length of its title and self-text (only in case of reddit user generated self-posts).

iii. Sentiment and Subjectivity: As is evident from our discussions so far, the score or votes a submission receives from the user community strongly influences the number of comments/ popularities of it, as reddit.com would feature most upvoted submissions on top of the page. In our study we approximated the score of a submission that is the difference between the upvotes and downvotes it receives as the sentiment of the submission. Additionally, we also performed a bag of words approach to determine the sentiment and subjectivity score of submission title based on the proportion of words that are from different sentiment tags and subjective/ objective tags.

iv. Content Similarity: We hypothesize that the more similar

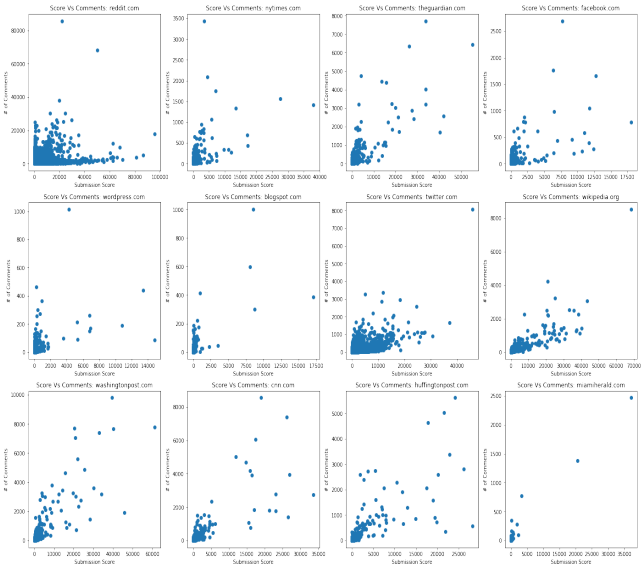


Fig. 2: Higher the number of upvotes that a submission receives, the more the number of comments for the post across domains and subreddits

submission is to other submissions from the domain in its semantic structure, the more popular it is likely to be, due to users familiarity with the content style and likewise to submissions in subreddit. In order to determine the semantic similarity of a submission to other submissions within the subreddit we constructed feature vectors from submission title using Doc2Vec models. Doc2Vec models are extension of Word2Vec models which translate words into vectors denoting its similarity to other words in the corpus based on semantic features. [7] provides details of working Word2Vec and Doc2Vec model for further reading. Now that, we have a feature vector for each submission we need to determine how similar is a submission title vector to other title vectors within a subreddit and within a domain. We computed cosine-based similarity between a titles vector and the average title vector for a subreddit and for a domain separately to determine how similar is the submission to the average submission style of a subreddit and of a domain.

Fig. 3: Submission content that are more similar to domain content posting style receive more number of comments on an average than the less similar ones

2) Community based

Submission in community with more users and in communities with more active users are more likely to receive more comments than submissions in small subreddit communities that are less active. To account for this, we developed features that determine subreddit popularity across reddit based on the total number of Users with at least one submission in the community, the average number of comments submissions in a subreddit receive and the immediate activity of the subreddit is determined based on total number of comments the previous submission received and the time gap since the last submission in that subreddit.

3) User based

Submissions made by users who are more active and who are associated with more subreddits are more likely to receive more comments for their posts than others. We also examined content diversity in user posts in terms of the number of different domains in which user posts and how does that relate to the number of comments their submissions receive. Further, we explored how flair tags on user influence the number of comments their submission receives.

4) Time of Submission

Popularity of submission also depends on the time during the day and the day of the week during which the submission was made. We account for this temporal effect, by using the

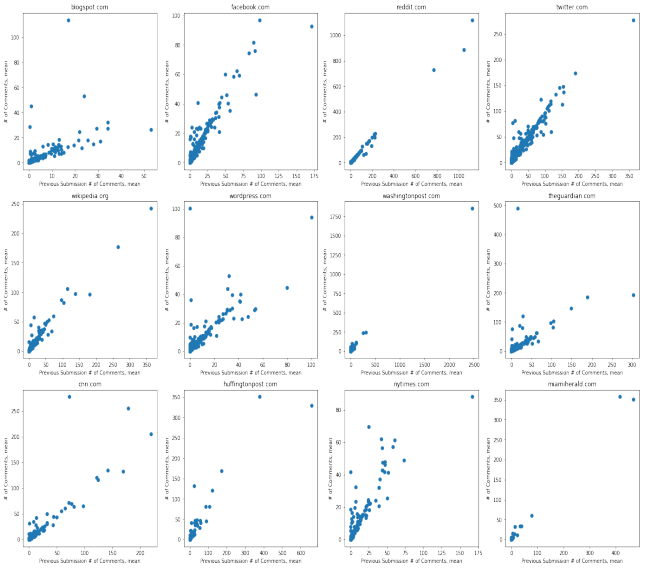


Fig. 4: The avg. # of comments the previous post receives shows a strong correlation with the number of comments the immediate next submission would receive

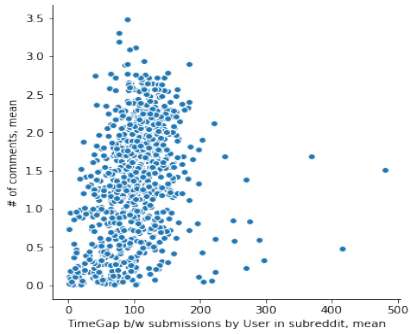


Fig. 5: With increasing gap between submissions made by User in a subreddit community the chances of the submission being more discussed becomes higher

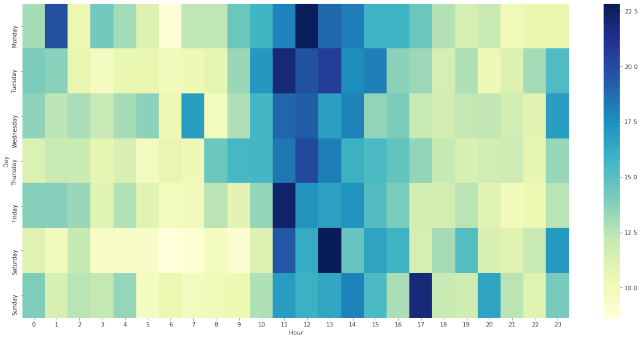


Fig. 6: Most submissions posted around 10 AM to 5 PM are more likely to become popular, with posts on weekdays being more popular than the submissions made over weekends

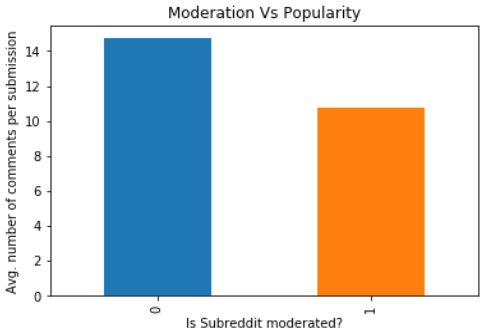


Fig. 7: Moderated community posts are less popular than that of unmoderated community posts

average comments per submission across different hours and days of a week.

5) Community Moderation

Popularity of a submission is constrained by the extent to which the subreddit in which the submission is made is moderated. This hypothesis is studied by examining relationship between avg. number of comments a submission in a subreddit receives against the total number of removed posts in the community.

C. Content Popularity Model

In this sub-section we propose a modelling framework to predict popularity of submissions based on features designed in previous sub-section. Models developed were at a domain level with one model for each of the 12 identified domains in this study. Since, subreddits different in the volume of submissions made by their user communities, we reweighted training data to balance for the different submission volumes across subreddits for each domain model.For all of our modelling experiments we trained on 70% of the data and reserved 30% of the data for validation.

We used a gradient boosting ensemble modelling technique to predict the number of comments a submission would receive based on the features used in the model. In order to examine the explanatory capabilities of the develop features groups, we developed a baseline model with number of comments as response variable and the score of submission as the covariate. We then incrementally add feature sets pertaining to content, user, community and time to the baseline model and observe the changes in R-Squares of the model. This would tell us the incremental explanatory capabilities added to the model by the included feature set. As illustrated in Figure 10, we observed that the proposed modelling framework had improved the explanatory capabilities of the baseline model which used only the number of votes a submission received as the feature. A detailed summary of the increase in R-Squares of model by including feature groups is provided in appendix C. Our choice of baseline model to use only score as the covariate is based on our understanding of reddits content display structure wherein submissions with higher upvotes are featured at top of the page and thereby receive more visibility. This could further cause the submission to be discussed more often irrespective of the content style, author or subreddit popularity nor time

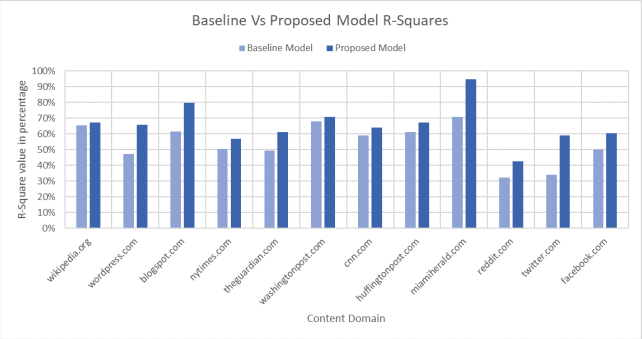


Fig. 8: Proposed model performs better than the baseline model which uses the submission score (votes) as the predictor variable

of submission.

IV. Results and Discussion

In this section we summarize the results from our modelling and analysis in predicting the popularity of content submissions from different domain.

Indeed, features along the dimensions of community, user, time and content style can provide increased explanatory capabilities to the popularity prediction model as can be observed from figure 11 and Table II. We validated our models against 30% of the data across each domain and the results are as shown in Table II. Further, a feature importance rank matrix shows features that strongly influence popularity of content across different domains. The age of a submission in terms of how long it has been since the submission has been made tells us the duration of visibility for a submission, the longer the duration the more the number of comments and we control for the same in our prediction model. As can be seen in figure 11, the submission score along with submission age, and size of the subreddit user community to determine the number of comments a submission sourced from a domain in a specific subreddit would receive.

Further, theres a strong time aspect to popularity of submissions depending on what time the submission was made to the community. Other time component that is associated with submission popularity is the time gap between submissions, the longer it has been since the last post in the community by the user, the more popular the post would be.

From the features developed to capture content style we observed that more the number of words in title for news link and blog submissions, more the number of comments they receive. Self-Generated content with more words in self-text are more popular than the ones with short descriptive text and less like other submissions in subreddit. Content similarity to domain style is more important for news link submissions than for content from other domains. Also, adding flair tags to submissions and authors improves its visibility to relevant discussion communities, increasing submission's popularity. It is also seen that, submissions made by users that are more spread across reddit in terms of multiple subreddit membership and posts on content from multiple domains are likely to have their submissions be more popular. However,

Domain	Train R-Squares	Test R-Squares
reddit.com	37%	37%
nytimes.com	59%	51%
theguardian.com	55%	51%
twitter.com	57%	55%
facebook.com	58%	46%
wikipedia.org	69%	60%
washingtonpost.com	69%	59%
wordpress.com	61%	37%
blogspot.com	76%	57%
cnn.com	60%	52%
huffingtonpost.com	64%	60%
miamiherald.com	92%	25%

TABLE II: Domain Model Validation Results. Overall, the developed models have an average train R-Square of 63% and test R-Square of 49%

our analysis is limited in that it uses only 1 month of reddit submissions data and studies only specific domains identified in the study. Domains not in the study might have other factors influencing its popularity, like Fox News links not included in the study might have different features influencing its popularity unlike the news links studied here. Also, users might have typos and short informal acronyms in their submissions which denote the same word, our analysis doesnt normalize submissions for such differences and would treat them as different words.

Since, this is an observational study any significant relationship between user, community or content style features with submission popularity is only associations observed in the data used in our analysis and can't be used to make any causal inferences. However, by validating on unseen data we have tried to evaluate how likely are we to see such associations in new submissions.

Another limitation of our study is that we dont have a methodology to analyse domains and subreddits that are relatively new with no or few submissions. Our approach currently requires historical submissions to be made to be able to analyse what factors are associated with popularity of submissions from a specific domain across communities in Reddit.

Further, there are potential confounds of external events that might have caused significant interest in certain topics that in turn caused submission in such topics to become more popular, in this phase of our analysis we dont explore this direction of analysis. It could be a potential future direction to our work, in which we determine popular topics being discussed on Reddit over a time frame and how similar are the identified topics.

V. Conclusion and Future Work

In conclusion, in this study we developed a framework to predict popularity of different domain-based submissions on Reddit and identified features that are common and

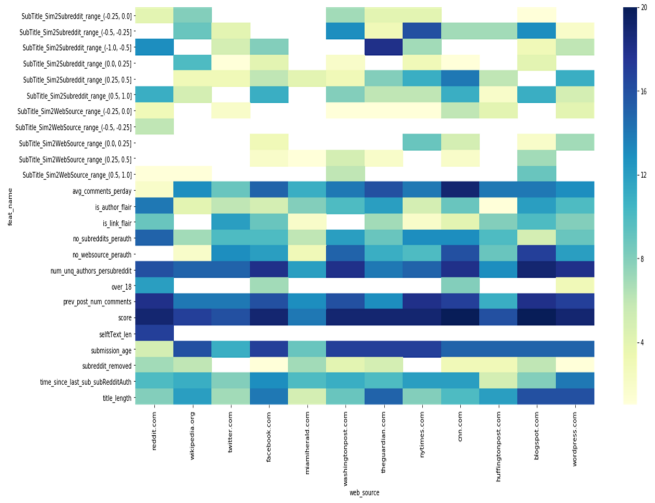


Fig. 9: Feature Rank Matrix- Each column represents a domain model and each row denotes a feature, a darker colour indicates, higher rank and a more important feature for the domain model

distinctive across domains in determining submission popularity.

[1] Our choice of domains for this study was based on the top domains in Nov 2016, that had most submissions on reddit.com. Also, we only analysed subreddits with more than 1000 submissions and more than 100 users subscribed to them. Submission popularity within small communities and relatively less active or new subreddits might have very different features impacting them.

However, from our analysis it is evident that different domains such as blogs, social media, news links and user generated content have different associations with user popularity, community activity, time of submission, similarity of content, length of content and extent of moderation in community. This helps us understand what aspects of a community, author, content or time, influence submissions of different formats such as tweets vs news articles.

An analysis of this kind provides us insight into how content should be structured to get the attention of a community, which would help in targeted advertising. It also helps understand how information of different type such as news vs blog posts vary in their popularity. News articles that are similar to other submissions in news domain are more popular than blog posts in which submission similarity to other submissions in blog domain isnt predictive of its popularity.

However, this study only presents a preliminary framework to predict popularity and evaluated its potential in being able to predict submission popularity and capturing differences in significance of relationships between user, community, content features and content popularity across news, blogs, social media and user generated content. Theres potential for more detailed analysis and repeating the same analysis for larger than one-month period to observe if the associations are consistent across different time periods. Further, we could also explore how comments on submission causes more discussion by studying nested comments in submissions.

Overall, our modelling framework helps identify when, where and what format should a submission on a domain content be made, for it to be popular and we showed inherent differences in information delivery style of different types of content such as news, blogs that make it more popular.

[1] All codes can be accessed at github.com/bharat92/CSC2552_ContentPopularityOnReddit

Acknowledgement:

This project was done as part of the coursework for CSC2552- Topics in Computational Social Science during Winter 2019 at University of Torronto. I would like to thank Professor Ashton Anderson for helping me scope the research question for the purpose of this study and for providing me with feedback that helped me improve upon my work.

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] PushShift Submission November 2016 Data Dump, https://files.pushshift.io/reddit/

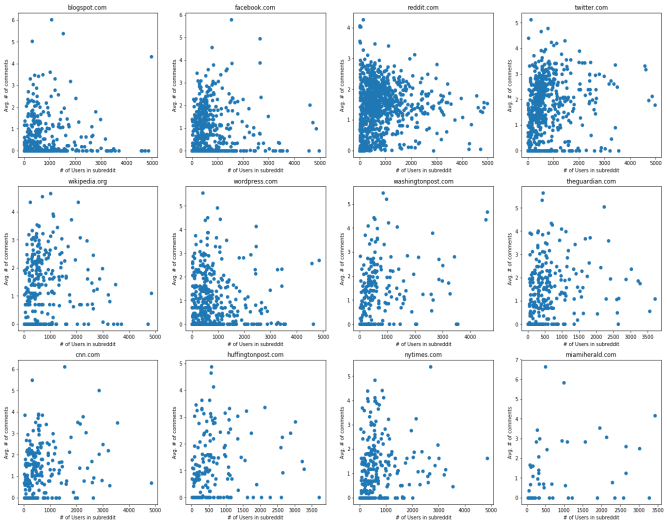
[7] Quoc Le, Tomas Mikolov, Distributed Representations of Sentences and Document

[8] Himabindu Lakkaraju, Julian McAuley, Jure Leskovec, Whats in a Name? Understanding the Interplay between Titles, Content, and Communities in Social Media

APPENDIX

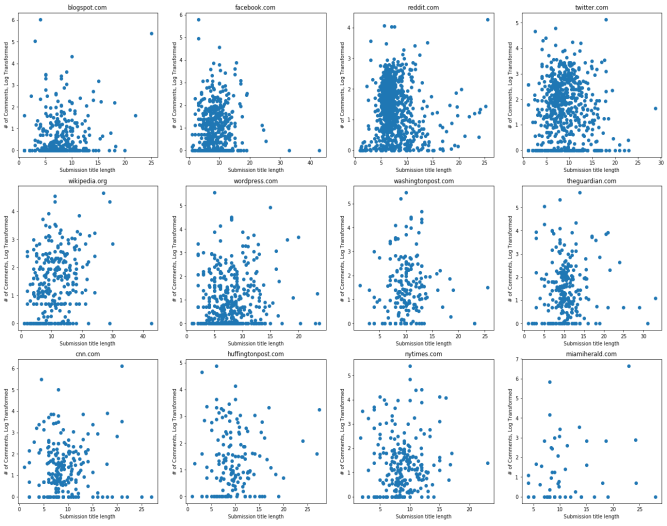
A. Additional Feature Explorations:

1) Size of subreddit user community:



Subreddits with a larger user community receive more comments per submission on an average. However, for subreddits with more than 1000 users the above plot across domains show a decreasing trend.

2) Submission Title Length:

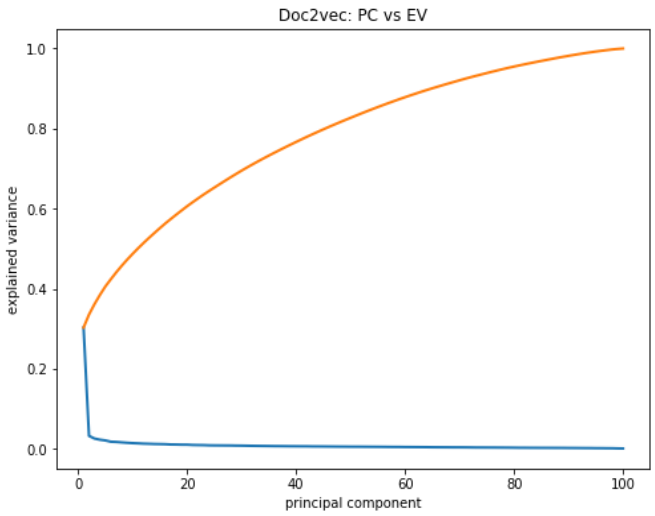


Submission with more words in title have more people making comments and discussing about them across domains.

3) Submission Similarity to other submissions in community:

News titles that are more like subreddit community content style receive more comments as compared to dissimilar news title styles whereas for self-generated content the most novel or dissimilar content receives more comments in a community.

B. Dimensionality Reduction: PCA on Doc2Vec Feature Vectors:



As can be see the top 2 principal components explain about 25% of the variance in Doc2Vec feature vectors for submissions while the remain components explain a small amount of variance. We could reduce to top few components and look at submission similarity.

C. Analysis of additional explained variance by developed feature sets, incrementally

Content Domain	No. of Submission, Train	No. of Submission, Test	Avg. no. of Comments, Train	Avg. no. of Comments, Test	Baseline Model	Baseline + Community	Baseline + Community + Moderation	Baseline+ Community + Moderation + User	Baseline + Community + Moderation + User + Time of Submission	Baseline + Community + Moderation + User + Time of Submission+ Content
wikipedia.org	2,775	224	11	10	65.41%	71.62%	71.62%	71.17%	70.18%	69.38%
wordpress.com	1,087	320	7	7	47.22%	58.89%	58.89%	59.30%	59.59%	60.74%
blogspot.com	682	221	6	5	61.60%	74.52%	74.66%	73.32%	73.70%	76.47%
nytimes.com	2,763	199	6	6	50.24%	58.49%	58.49%	58.09%	58.12%	58.87%
theguardian.com	3,458	202	12	12	49.44%	54.92%	54.92%	54.21%	54.19%	54.56%
washingtonpost.com	2,527	163	11	12	67.97%	69.82%	69.82%	69.02%	68.95%	69.45%
cnn.com	2,206	167	11	11	58.93%	61.45%	61.45%	60.25%	60.83%	60.13%
huffingtonpost.com	1,363	124	14	13	61.20%	63.25%	63.25%	64.52%	63.89%	64.20%
miamiherald.com	106	32	9	13	70.74%	88.71%	88.71%	92.08%	91.74%	92.30%
reddit.com	12,80,113	964	7	7	32.15%	37.06%	37.61%	37.62%	37.61%	36.52%
twitter.com	23,653	494	14	13	33.88%	57.49%	57.49%	58.44%	57.14%	57.49%
facebook.com	2,389	360	9	8	49.89%	58.47%	58.47%	57.90%	58.86%	58.23%