Demystifying content popularity on Reddit

Katyaini Himabindu Lakkaraju
Project Group Number: 25
himalv@stanford.edu

December 10, 2012

Abstract

World Wide Web is now teeming with user generated content as more and more sites allow users to become an integral part of content generation and dissemination. Several sites such as Youtube, Digg, Twitter, Facebook, Reddit allow users to post content of their choice, comment on and rate the content posted by other users. This user generated content has given rise to a whole bunch of interesting problems such as prediction of content popularity, user popularity and many more. Solving these kind of problems involves understanding the underlying sociological processes and coming up with appropriate algorithmic solutions. The goal of this report is two fold. Firstly, I present my findings about the influence of various factors such as community aspects, user attributes, content quality, multiple content exposures on the popularity of the content on Reddit. Secondly, I propose models that can effectively predict the content popularity in terms of the number of comments a particular post would get at any given point in time.

Introduction

With the inception of sites such as Twitter, Facebook, Reddit, Digg etc., the web is now a power house of user generated content. It is often seen that users share content of their choice, post comments, write reviews, upload data, acknowledge content posted by other users on these sites. In some sense, users are leaving their digital footprints almost everywhere on the web. This kind of data when harnessed appropriately can be used for various purposes ranging from understanding customer segments for better product marketing to understanding sociological aspects of human behavior. The surge in this kind of data has reinforced the need for computational algorithms and techniques which effectively capture human behavior, both at an individual and aggregate level.

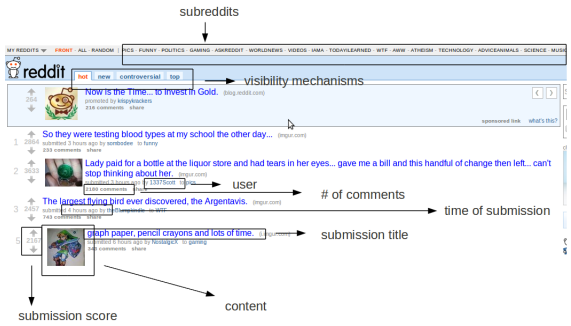


Figure 1: Snapshot of reddit front page

This paper primarily focuses on analyzing the content popularity and the associated dynamics on a social content sharing site called Reddit. Reddit has a very interesting approach to organizing content and determining content visibility. The principal entities of reddit are users and subreddits[1]. A user can follow (this is 'friends' in reddit terminology) other users and subscribe to various subreddits. At any given point in time, user's homepage provides seamless access to various content views. A user can browse through the activity of people he follows. He can see the most recent posts and highly popular posts submitted across various subreddits. In addition to all these, the content from those subreddits which the user subscribed to also shows up on his homepage. The user can browse through each of these content views by clicking on various tabs on his home page. Further, users can share interesting links and textual content in the subreddits of their choice and other users vote and comment on these

[1]I will be using the terms communities and subreddits interchangeably through out this writeup

submissions. The interactions between all these entities would render the task of predicting content popularity more challenging and interesting.

The concept of *content popularity* is subjective. In this work, we associate popularity of a post with the number of comments it receives. The questions that we aim to answer in this work are -

- How much impact does the quality of the content have on its popularity ?
- What role does the subreddit play in the content popularity ?
- What is the impact of the user (who posts the content) on the content popularity ?
- Does the title of the submission play a role in its popularity ? If so, what would constitute a good submission title ?
- How would the submission popularity vary with multiple subsequent submissions of the same content ?
- Can we design a framework which accounts for the submission popularity and accomodates the effects of the various factors highlighted above ?
- How can we account for the popularity deviations of the submissions with time ?

Surprisingly, there is not much published work on analyzing the dynamics of Reddit. This work is one of the initial attempts towards the same.

Related Work

In this section, I will briefly discuss the literature that is most relevant to this piece of work. This work cuts across two aspects - content popularity prediction on social media and analysis of voting / commenting patterns on social sharing sites. Hence, this section deals with the discussion of work along these two lines.

Content Popularity on Social Media

There has been a surge in the interest in problems of this kind in the recent past. [2] deals with the problem of determining if a given post on twitter would elicit any response or not. No distinction is made between various kinds of responses such as reply, retweet, favoriting etc. This work presents a detailed analysis of content features based on natural language heuristics such as number of stop words, positive and negative sentiment words and identifying lexical items with high correlation with previous tweets which elicited responses. Though this work gives some detailed insights into the various kinds of content features, the problem that this work solves is a limiting case of what we are proposing. [1] attempts to solve the problem of predicting the popularity of an article on a social networking site (twitter). The measure of popularity here is determined by how many times the article url is shared across twitter. The authors employ classification and regression based approaches with various features such as categories / topics the article belongs to, subjectivity of the article, named entities mentioned and the source of the article However, this work also does not deal with the aspect of predicting the popularity of an item at various points in its lifetime. [3] deals with the problem of predicting if the popularity of some content exceeds certain threshold. The approach employed has its roots in survival analysis framework. The main idea behind this work is to employ initial popularity measures to detect the long term popularity of a piece of content. There is also a huge body of work on related lines about predicting the popularity of posts on social media and other controlled forums on social media [4] [5].

Analysis of Voting and Commenting Patterns on Social Media

Though there is a lot of literature on analyzing user behavior on social media, this subsection focuses on analyzing user behavioral activity with respect to social sharing sites such as Digg. [6], [7], [8] present an extensive analysis of Digg. These papers present simple models capturing the user behavior and content visibility to capture the dynamics of digg. These papers highlight user influence and interestingness of the content as the two primary factors behind the popularity of a piece of content. Further, the authors use this model to predict the long term content popularity. However, the dynamics of the data that we are dealing with is a lot more complex since there is additional component of topic (content) based communities and also the content that we are dealing with comprises of a lot of personal content (pictures, queries etc.) in addition to other knowledge based postings (as opposed to mostly news based urls on digg).

Our work is different from the state-of-the-art in two aspects. Firstly, the process of content generation, organization and visibility on Reddit has not been explored so far. Secondly, we attempt to predict the popularity of the content at various phases of its lifetime rather than just predicting the initial popularity or the long term popularity.

Dataset and Features

This section describes the data and the feature space in detail.

Dataset

Data collection process spans three phases.

- First phase comprised of crawling those posts which are duplicate submissions of the same content. As discussed in the introduction, one of the objectives of this work is to understand how popularity varies with multiple submissions of the same content. In order to achieve this, the first phase of the crawl primarily constituted of crawling such resubmissions. Note that Reddit alerts a user who is trying to submit duplicate urls. However, there is currently not any mechanism which alerts the user who is submitting duplicate images. Hence, the content of the submissions obtained during this crawl primarily comprises of images. This dataset comprises of about 102K posts spanning 928 communities. The number of unique images in this set is 9,022 which means that on an average there are roughly about 10 - 12 resubmissions of the same image.

- Since the data that we had crawled in the first phase was highly skewed towards duplicate content, a second phase of crawling was initiated which crawled data from those communities which were found in the data crawled in the first phase. This was also an attempt to obtain more data from communities such as politics, technology which were under represented in the data from the first phase. From this second crawl which was carried out bi-hourly for about 4 days added another 20K posts to the dataset from the first phase.

- The final phase of the crawl was carried out primarily for monitoring the front page of reddit and to figure out the thresholds of content promotion on to the front page. This was necessary to model the popularity changes with time. The data that I obtained in this third phase comprised of 1248 posts which hit the front page and this data is only used to estimate the score threshold which qualifies a submission to appear on the front page. This data was not used in any other analysis.

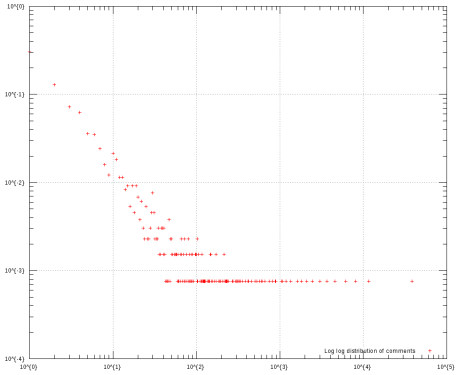


Figure 2: Log log plot of the comment distribution (Power law fit: xmin = 789, alpha = 3.5)

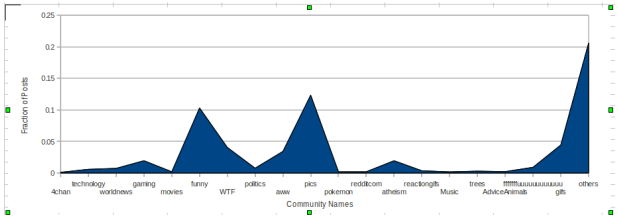


Figure 3: Community vs. Fraction of Posts

The dataset on which the entire analysis in this report has been carried out comprises of about 120K posts from 226 communities. Note that the number of posts and the communities is smaller than the numbers specified above. This is because of the fact that there were several communties with very low activity (less than 50 posts). So, I had to prune down such communities and all the associated posts. The basic statistics of this dataset and correlations between the number of upvotes, downvotes and comments are shown in the Table 1. It can be seen that there is a very high correlation between the number of upvotes, downvotes and comments. Another unsurprising aspect empirically verified

Table 1: Basic statistics of the dataset

# of Posts	120423
# of users	58032
# of communities	1012
# of active communities	226
# of up votes	89M
# of down votes	68M
# of comments	3.8M
Average # of posts per user	1.83
Average # of posts per community	114.1
Average # of up votes per post	845
Average # of down votes per post	642
Average # of comments per post	36

Post Level Correlations	
correlation between # of upvotes and downvotes	0.99
correlation between # of upvotes and comments	0.73
correlation between # of downvotes and comments	0.72
Community Level Correlations	
correlation between # of posts and comments	0.97
correlation between # of posts and upvotes	0.96
correlation between # of posts and downvotes	0.97
correlation between # of upvotes and downvotes	0.99
correlation between # of upvotes and comments	0.96
correlation between # of downvotes and comments	0.95

by the results in Table 1 is the strong correlation between the number of posts, upvotes, downvotes and comments at the community level.

Figure 2 shows the distribution of the comments in the dataset. It can be seen that the density function follows a power law. Also, Figure 3 shows the fraction of posts in the data which belongs to the respective communities. It can be seen that pics, funny, aww, gaming, atheism are amongst the well represented communities in the dataset.

Features

This subsection discusses in detail all the features that I will be employing in the approaches described in the subsequent sections. The features in the dataset that we are dealing with can be broadly categorized into user features, post title features, community features, submission features and content features. I present here a detailed analysis of each of these features and also describe how the feature set has been pruned to come up with an appropriate feature list for the prediction tasks.

- User Features - As discussed in the introduction, one of the goals of this work is to study how user's influence effects the popularity of a submission. In sites like Twitter, this factor would be very crucial for the popularity of a post. On the other hand, Reddit's content organization mechanism allows for content to take precedence over the user influence in accounting for the submission popularity. However, it was important to understand the role of a user in such an environment. Some of the features which would allow us to capture this are average number of comments, upvotes and downvotes received by the user's previous submissions and the user's subscriber list. Unfortunately, user's subscriber lists are mostly private on Reddit and hence it is not possible to access them. Also, our dataset did not provide enough representation of each user to confidently capture metrics such as average number of comments, upvotes and downvotes. To overcome these issues, I extracted each user's link karma from reddit and used it as a proxy for the user influence. Infact, doing so actually yielded some interesting correlation with the number of upvotes, downvotes and comments.

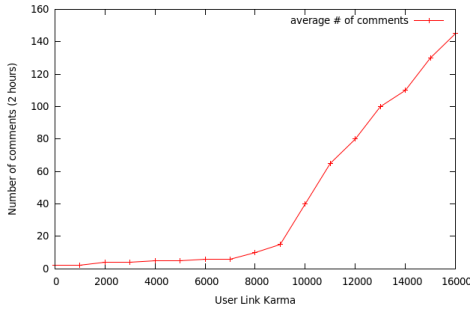


Figure 4: Link Karma vs. Average Number of Comments (in the first 2 hours of post submission)

Figure 4 shows the plot of user link karma vs. average number of comments received by their submissions. It can be seen that this curve exhibits an exponential distribution and users with link karma greater than 10K have a relatively higher average number of comments. This is interesting because it indicates that users with high link karma are expert users who have mastered the art of submitting posts on Reddit in such a way that they receive high attention. Also, users with higher link karma typically have more visibility because of lots of other users subscribing to them and the high attention received by their submissions can also be attributed to this reason.

- Post Title Features - Post title features comprise

of the attributes of the title of the post (or the descriptive text written by the user). There are multiple features relevant to the text of the title that we can consider. I explored a bulk of features in this space such as KL divergence of the words used in the title with respect to the community's language, presence of named entities and positive and negative sentiments, subjectivity of the title, number of words in the title, fraction of content specific words in the title, fraction of community specific words in the title. Amongst all these features, fraction of content specific words, fraction of community specific words, subjectivity of the title proved to be the most effective. KL divergence was also another feature which showed some promising insights, however, this had a high correlation with fraction of community specific words and metrics such as information gain suggested that usage of fraction of community specific words was more beneficial as compared to KL divergence.

In order to determine the community specific words and content specific words, I used the following metrics. Let us denote subreddit specificity score of a word with respect to subreddit s as w_s and content specific score as $w_{c,s}$.

$$w_s = \frac{\#\ of\ times\ w\ appears\ in\ titles\ on\ subreddit\ s}{\#\ of\ times\ w\ appears\ in\ the\ entire\ dataset}$$

$$w_{c,s} = \frac{\#\ of\ times\ w\ appears\ in\ titles\ of\ content\ c}{\#\ of\ times\ w\ appears\ in\ the\ subreddit\ s}$$

These two scores determine the the specificity of a word with respect to its community and the associated content respectively. I designated thresholds for these scores and any word with a score higher than the designated threshold was tagged as a content specific or a community specific word and fraction of such words in each post is computed. Formally the subreddit specific (sp_p^{subr}) and content specific (sp_p^{cont}) scores of a particular post p are :

$$sp_p^{subr} = \frac{\#\ of\ subreddit\ specific\ words\ in\ title\ of\ post\ p}{\#\ of\ words\ in\ title\ of\ post\ p}$$

$$sp_p^{cont} = \frac{\#\ of\ content\ specific\ words\ in\ title\ of\ post\ p}{\#\ of\ words\ in\ title\ of\ post\ p}$$

Subjectivity of the title is a much more involved concept. Initially, I started with the naive approach of using lexicons for positive and negative words. I found that another approach which involved subjectivity detection at a phrase level[2] was much more effective. I employed this approach which assigns a score to each title in the range of 0 to 1 indicating the extent of its subjectivity. Note that we are not distinguishing between positive and negative sentiments here, but just extracting the subjectivity of each title. This aspect exhibited interesting correlations with the number of comments at a community level.

[2] www.lingpipe.com

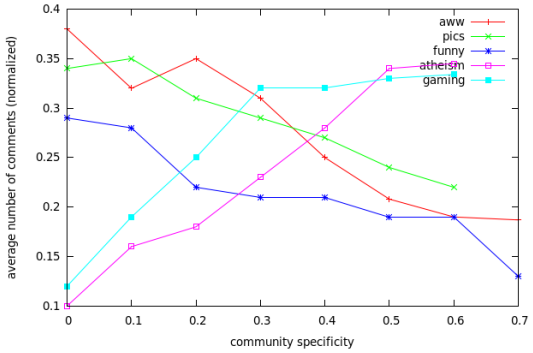


Figure 5: Community specificity score of a post vs. Average # of comments (normalized)

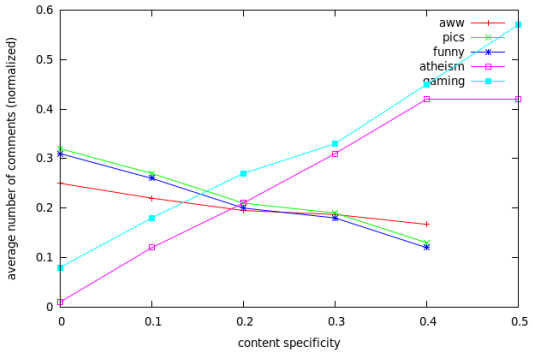


Figure 6: Content specificity score of a post vs. Average # of comments (normalized)

Figures 5, 6 and 7 show the effect of aforementioned three factors on the average number of comments received by the submissions exhibiting the corresponding feature values.

- Community Features - The major feature in this category was 'activity' in the community. Table 1 shows a positive correlation between the number of posts in a community and the number of comments obtained by posts in that community. So, the number of posts in the community could serve as a good indicator of the activity within the community. Further, I tried to analyze the peak activity periods for each of the

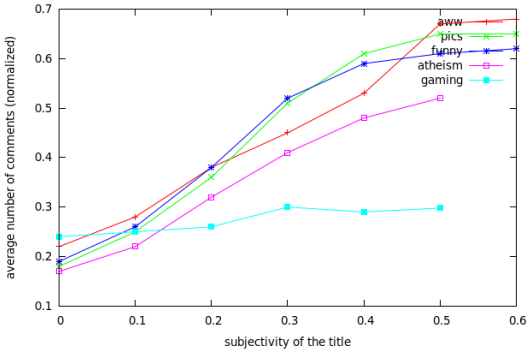


Figure 7: Subjectivity score of a post vs. Average # of comments (normalized)

communities. Figure 8 shows the hour of the day (in UTC) plotted against the average number of comments in that hour for top 5 communities in the dataset. This shows that the hour of the day is one of the factors that influences the popularity of a particular post. Number of posts within the community during any particular hour of the day also exhibits similar sinusoidal behavior. In order to quantify the 'activity' in a community, the feature of average number of posts in the community at that particular hour turned out to be the most effective feature offering highest information gain.

- Resubmission Features - One of the key aspects of our study is to understand how content popularity varies with multiple submissions of the same content, albeit with different community, submission and user attributes.

  Figures 9 a and b show submission number (at the corpus level) plotted against number of comments and submission number (at community level) plotted against number of comments respectively. It can be seen that the average number of comments decrease with the increase in submission number both at corpus and community level. This is primarily because the first submission usually is novel and gets more attention from the subscribers and once the content becomes stale, the number of people who would be interested in commenting on it again decreases. However, this decrease is non-monotonic and a few peaks can be seen on and off in the curve.

  Our analysis showed that this non-monotonicity can be explained by a couple of interesting factors. Firstly, if the difference in time between two submissions is high (of the order of several days), then the resubmission's popularity remains unaffected by the previous submission. Another interesting aspect that we observed during our analysis was that the number of comments a submission would get depends on the number of comments obtained by the previous submissions. Figure 10 demonstrates both of the aforementioned aspects. If the previous submissions failed to get adequate visibility or generate appropriate interest due to being posted in an inappropriate community or with an uncatchy title, then the probability that the current submission gets more visibility and hence more comments increases. On the other hand, if the previous submissions received the attention they are due, then the current submission is likely to receive lesser comments because users have already seen the previous submissions of the same content. This dependency on the previous submissions can only be broken by the difference in time between two adjacent submissions. The higher this difference, the lower the dependence of the popularity of the current submission on the previous ones.

  In order to capture this behavior, let us define a metric called potential of the post p (both across communities and within the same community) $\phi_{c,p}$ and $\phi^s_{c,p}$ where p denotes a submission of some content c and s denotes a specific subreddit. Let A_n and $V_{c,n}$ denote the average number of comments a post with submission number n would get (averaged across all communities) and the actual number of comments obtained by that submission respectively (Adding a subscript s to these quantities corresponds to their subreddit specific values). The value of k is set based on empirical performance and this will be discussed in detail in experiments section.

$$\phi_{c,p} = \frac{1}{k} \sum_{i=n-1}^{n-k} \frac{A_i - V_{c,i}}{\Delta t_{c,p,i} + 1}$$

$$\phi^s_{c,p} = \frac{1}{k} \sum_{i=n-1}^{n-k} \frac{A_{i,s} - V_{c,i,s}}{\Delta^s t_{c,p,i} + 1}$$

where $\Delta t_{c,p,i}$ denotes the time difference in days between the submissions p and i of content c. Similarly, $\Delta^s t_{c,p,i}$ denotes the time difference in days between submissions p and i of content c

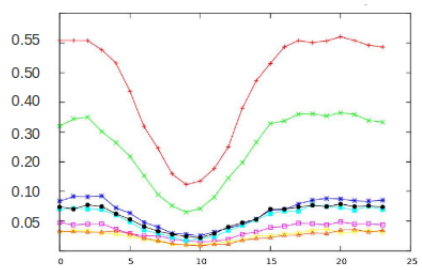


Figure 8: Hour of the day vs. Average Number of Comments (normalized against maximum number of comments received by a post)

(both p and i are posted in subreddit s). So, we pick the following resubmission features -

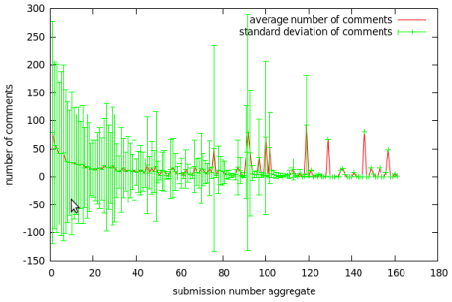
- submission number
- potential function of post p evaluated across previous k submissions

Note that each of the aforementioned features are computed both across the communities and within community.

- Content Features - This corresponds to identifying the interestingness of the image itself. This is a latent attribute and needs to be inferred from the multiple submissions of the same content. I will be parameterizing this as a numeric quantity and the higher this value the better the content. Note that we can also quantify interestingness of the content with respect to various communities but our dataset does not carry enough submissions of the same content in multiple communities. Hence, I will be sticking with the notion of interestingness which holds across all the communities.

Our Approach

This section discusses in detail the framework that I employ in order to predict the popularity of any given submission at various points in time. This framework has two parts to it - First part involves proposing a model for predicting initial popularity as a function of all the features discussed in the previous section. Second part involves proposing a model which takes as input initial predictions from the first part and then predicts the popularity of the submission at subsequent points in time.



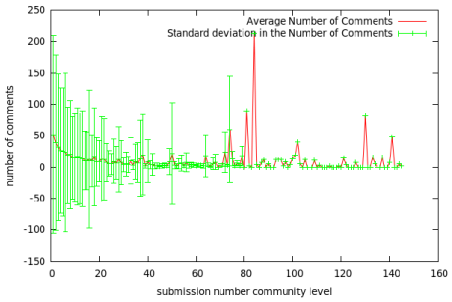


Figure 9: a. Submission Number (Aggregate) vs. Number of Comments b. Submission Number (Community level) vs. Average Number of Comments (normalized against maximum number of comments received by a post)

Predicting initial popularity

The typical approaches to predicting the initial popularity given a set of features involves using off-the-shelf classification or regression based approaches. However, in our case, as seen in the previous section, most of the features have different effects on the popularity depending on the subreddit we are talking about. For instance, certain subreddits such as politics favor submissions with titles which are specific to the community and the

Table 2: List of all the features and Notations

Notation	Feature Description
c	Index denoting content
s	Index denoting subreddit
u	Index denoting a user
p	Index denoting a post
n	submission number of the post across all the subreddits
n_s	submission number within the subreddit s
lk_u	link karma of user u
sp_p^{subr}	subreddit specific score of post p
sp_p^{cont}	content specific score of post p
sp_p^{subj}	subjectivity score of post p
$avgc_{s,h}$	average # of comments in subreddit s during the hour of the day
ϕ_p	potential function of the post p carrying content c (across all subreddits)
ϕ_p^s	potential function of the post p carrying content c with respect to the subreddit s
Δt	time in days since previous submission (across all subreddits)
Δt^s	time in days since previous submission within the subreddit s
r_c	interestingness of content c (latent variable)

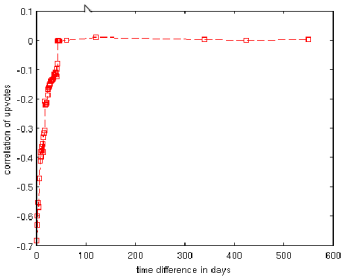


Figure 10: Time difference vs. Correlation (explained in the text of the section)

content. On the other hand, subreddits such as pics, funny do not emphasize this aspect. Further, the interestingness of content is an important aspect that needs to be modeled as a latent variable and should be integrated as an explanatory factor into the model. Putting together all these details, it becomes essential to come up with our own model that can incorporate these aspects.

I propose a regression based formulation encapsulating a latent variable for modeling the content interestingness in order to solve the problem at hand. Let $\hat{V}_p$ denote the estimated initial popularity of a post p. It can be modeled as :

$$\hat{V}_p = \beta_0 + \beta_1 \ exp(\alpha_1 \cdot lk_{u_p}) + \beta_2^s \ sp_p^{subr} + \beta_3^s \ sp_p^{cont}$$
$$+\beta_4^s \ sp_p^{subj} + \beta_5 \ avgc_{s,h} + \beta_6 \ exp(\alpha_2 \cdot \phi_p)$$
$$+\beta_7 \ exp(\alpha_3 \cdot \phi_p^s) + \beta_8 \ exp(\alpha_4 \cdot n_p) + \beta_9 \ exp(\alpha_5 \cdot n_{s,p}) + \beta_{10}^s \ r_c \quad (1)$$

The unknown variables in the equation above are all the α, β quantities and r_c. Note that the parameters with superscript s correspond to subreddit level parameters and r_c is a content level parameter. The objective function to be minimized can be formulated as :

$$min \ \frac{1}{2} \sum_p (V_p - \hat{V}_p)^2$$

I employed gradient descent algorithm with simultaneous update in order to solve the above optimization problem. The algorithm is executed till the difference in the squared error between subsequent iterations is < 0.0001

Predicting content popularity at next time instant

The next step is to generalize this framework to predict the popularity of a particular submission at various time instants. The popularity of a given post p at the first time instant is given by equation 1.

Autoregressive models (AR) constitute simple and effective approaches to time series modeling. AR models posit the value of a variable at time 't' as a function of the variable's values at p previous time instants. The AR(m) model is defined as

$$V_t = c + \sum_{i=1}^{m} \psi_i V_{t-i} + \epsilon_t \quad (2)$$

c is a constant, ϵ_t is the white noise and ψ_i correspond to the coeefficients which need to be estimated. A major limitation with AR is that it does not capture bursts effectively. In the application that we are looking at, submission popularity increases in bursts when the submissions are promoted to the front page. So, I propose an Augmented Autoregressive model (AAR) which can capture the bursts in the submission popularity.

AAR is designed to capture the sudden bursts in the commenting activity due to promotion to front page. Let t_{cur} denote the difference in hours since the submission was posted till the current time instant. Let t_{pr} denote the time the upvotes on the submission exceeded the threshold and the submission was promoted to the front page. t_{de} is the time to decay which is the average number of hours after the submission's promotion to the front page at which the submission is removed from the front page. The augmented autoregressive model can be formulated as :

$$\hat{V}_{p,t} = c \ u(t_{cur} - t_{pr}) \ u(t_{pr} - t_{cur} + 2)$$
$$+ \sum_{i=1}^{m} \psi_i V_{t-i} - d \ u(t_{cur} - t_{pr} - t_{de}) \ u(t_{pr} + t_{de} - t_{cur} + 2) \quad (3)$$

c and d are coefficients which need to be estimated and u() is a step function. u(x) = 1 if x > 0 and u(x) = 0 otherwise. The above equation precisely induces the increase in the rate of comments from previous time instants due to promotion of a post on to the front page. The first term in the summation in equation 3 models the sudden increase in the rate of comments due to promotion of a submission to the front page. u(t_{cur} - t_{pr}) = 1 when the content has been promoted to the front page. u($t_{pr} - t_{cur} + 2$) = 1 for the first 2 time instants after the submission's promotion to the front page. This term in the summation will be effective only when both the step

functions evaluate to 1. This implies that this term is effective during the first two time instants after the submission's promotion to the front page. After this interval, the step functions evaluate to 0 and then the onus is on the second term in the summation to model the comments as a function of the prior popularity. In an analogous manner, the last term in the summation is employed to account for the decrease in the rate at which comments pour in after the submission has been pulled off the front page.
In the formulation above, the parameters that need to be estimated are c, d, t_{pr}, t_{de} and ψ_i. Note that t_cur corresponds to the number of hours elapsed since the submission of the post. The parameters t_{pr} and t_{de} are set from the empiricial observations of the data. As discussed in the section on Dataset, I obtained 1248 posts which were promoted to the front page. Analysis of this set of posts revealed that submissions promoted to the front page need to satisfy a specific threshold on the number of upvotes. This threshold needs to be achieved within about 10 hours of the submission. The minimum number of upvotes needed for a post to be featured on front page is not published by reddit, however from the data collected, the minimum number of upvotes of any submission which was featured on the front page is 1012. Typically, the rate of accumulation of comments increases steeply from the instant the content shows up on the front page and the submission shows up on the front page typically for an average of 5 hours (1 - 9 hours is the typical time that was observed in the dataset) since the time it shows up first on the front page.
From these observations, $t_{de} = 5$ (average number of hours after which submission is removed from the front page). t_{pr} is defined as follows :

$$t_{pr} = \begin{cases} t & \text{if } V_{p,t}^{upvotes} \geq 1012 \text{ and } t \leq 10 \\ 0 & \text{otherwise} \end{cases}$$

The above equation reads that t_{pr} is the time instant t when the number of upvotes of post p exceed 1012 and if this time instant t is within 10 hours of the post's inception, it is 0 otherwise.
In order to estimate c, d and ψ_i, the formulation posed in equation 3 is solved using minimization of least squares and employing gradient descent to obtain estimates as discussed in the previous subsection.

Experimentation

This section discusses in detail the experimentation carried out to demonstrate the effectiveness of the proposed framework and analysis of the effect of various features on the popularity of the submission. This section is divided into three parts. The first subsection discusses the evaluation of the initial predictor. The second subsection discusses the performance of AAR model on the task of predicting the number of comments obtained by a submission. The last subsection discusses the relative importance of features in explaining the popularity of the content.

Setup All the experiments have been performed using 75% of the data as training set and the remaining 25% of the data as the test set. All the feature values have been normalized to have a mean of 0 and unit variance in order to facilitate faster convergence of the gradient descent algorithm and also easier interpretation of the weights assigned to the respective features.

Baselines Since there has not been much work on popularity prediction for reddit, I compared our model with its ablations. In addition, since the data has lots of submissions which are duplicates, we want to examine the explanatory capability of the resubmissions. In order to do this, I used two more baselines - First one is a model of exponential decay of the number of comments depending on the submission number. Second one is a model of exponential decay in the number of comments depending on the submission number with the subreddit.
Baseline 1:

$$\hat{V}_p = \beta_0^c + V_{c,1} - \beta_1 e^{\alpha n_p}$$

$V_{c,1}$ denotes the number of comments of the first submission of content c (this will be a known quantity) and n_p indicates the submission number of post p. Other parameters are again estimated using least squares minimization process.
Baseline 2:

$$\hat{V}_p = \beta_0^c + V_{c,s,1} - \beta_1 e^{\alpha n_{p,s}}$$

$V_{c,s,1}$ denotes the number of comments of the first submission of content c in subreddit s (this will be a known quantity) and $n_{p,s}$ indicates the submission number of post p within subreddit s. Note that β_0^c is a content level parameter.
Baseline 3: This baseline just corresponds to the linear regression formulation without the latent variable r_c and with all the coefficients taken at a global level.

Table 3: Initial Prediction Results

Approach	R^2
Our Model	0.672
Our Model - User	0.603
Our Model - Title	0.588
Our Model - Resubmission	0.267
Our Model - Content	0.512
Baseline 1	0.287
Baseline 2	0.376
Baseline 3	0.412

Evaluating Initial Predictions

Table 3 shows the results of predicting the # of comments in the first one hour. The predictions of the models are evaluated using the R^2 metric. R^2 is the coefficient of determination and relates to the mean squared error and variance as :

$$R^2 = 1 - \frac{MSE}{Var}$$

The results of other ablations of the model are also highlighted in the Table. Our Model - feature indicates that the ablation model does not capture the effects of that particular feature. For instance, Our Model - User model does not contain the user features.

Discussion It can be seen that the model we propose is the most effective for explaining the submission popularity. Table 3 also indicates that removing resubmission features decreases the accuracy of the model considerably. On the other hand, user and title features impact the overall accuracy much lesser. It is interesting to see that ignoring content modeling also results in significant dip in the accuracy. Further, the baselines especially the second baseline accounts for a decent portion of the popularity too indicating that initial submissions in the subreddit typically get a lot of attention as opposed to the subsequent submissions of the same content.

Evaluating Predictions at Subsequent Time Instants

Figure 11 highlights the predictive performance of AAR and AR models for time series prediction task at hand. In addition, it also shows the performance of using the initial prediction model [3]. It is important to discuss the details of the training phase here. AAR and AR models are fit to the data by observing the # of comments in the previous 2 time instants. On the other hand, initial prediction model was also trained separately for each time instant t to see how effective it would be in making the predictions at various time instants. Firstly, note that to make the initial prediction model work in a set up like this, I had to train the model separately for each slice in time causing additional overhead. From this, it can be seen that it is not ideal to train the model to obtain the coefficients for each feature separately at each time instant. ARR and AR models on the other hand have relatively parameters to keep track of and are performing efficiently.

Discussion From Figure 11, it can be seen that AAR model, despite having very simple augmented additive terms, performs better at almost all time instants. AR model exhibits comparable performance at time instants $>$ 16 hours. This is the time period when the flow of comments would have been stabilized. The initial prediction model despite of being trained separately at each time instant performs slightly poorly compared to the AAR model. It was observed that the features could explain the popularity of the post better during the intial phases of the post's life time. However, when it comes to modeling the bursts and sudden decay in the rate of comments, conditioning the popularity only on the features performs poorly.

Evaluation on Bursty Data The main objective of using an augmented model as opposed to AR model directly was to account for the bursts and steep decays in the rate of comment flows. In order to study this more closely, I picked up those posts from the test set for which the rate of increase of comments suddenly increases and decreases beyond a particular threshold. The entire dataset comprises of about 28.32% of such posts. The test set in particular has 8.92% of these posts (percentage computed across entire dataset) which exhibit this kind of properties.
Figure 12 shows the plots for comparing the performance of AAR and AR models. The differences in performance are now much more clearly visible. It can be seen that AR model performs poorly in the beginning, then adapts itself to the changes in the comment rates and then the performance degrades again when there is a decline in the comment rate

[3]Note that all the other ablations and baselines were also tested on the same lines as that of the initial prediction model for this task, however, they exhibited poorer performance than the initial prediction model and hence have been left out of the Figure to avoid clutter.

Table 4: Communities with low and high coefficients for various features

Community Specificity		Content Specificity		Subjectivity		Content r_c	
High	Low	High	Low	High	Low	High	Low
atheism	aww	politics	funny	pics	politics	pics	politics
gaming	pics	business	pics	funny	business	food	business
technology	funny	atheism	gifs	gifs	food	apple	technology
politics	gifs	gaming	reactiongifs	aww	worldnews	funny	games
worldnews	reactiongifs	movies	aww	food	technology	gifs	movies

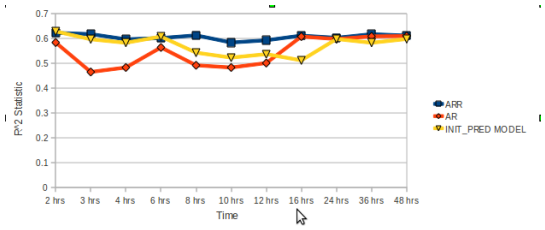


Figure 11: Time since posting of submission vs. R^2 prediction

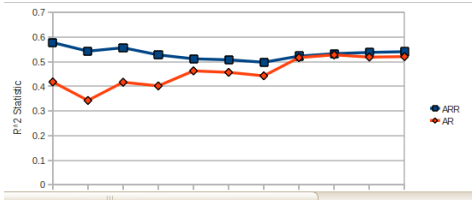


Figure 12: Time since posting of submission vs. R^2 prediction (of bursty data)

and then becomes stable towards the end of the curve. As can be seen in the figure, AAR can model these abrupt changes much more effectively.

Feature Analysis

In this subsection, we present answers to the questions such as which feature is the most important in explaining the content popularity. This is done by comparing the coefficients obtained from the initial prediction model. Note that not all coefficients are at the global level, there are several subreddit level coefficients. We begin by analyzing all the coefficients at the global level. β_1, β_6, β_7, β_8 and β_9 have multiplicate exponential terms. Comparing these coefficients, it was found that potential function within the subreddit is the most influential factor. This is followed by submission number in the subreddit. Then, we have the potential function across the entire corpus followed by the submission number of the entire corpus. Last ranked is the user influence.

Moving on to the subreddit specific coefficients, it was interesting to see that different subreddits emphasized different aspects. As discussed in the Features section of this write up, few reddits favor usage of community specific words and others dont. Table 4 highlights some of the top communities which place higher and lower weights on various factors that we are considering. It is interesting to see that communities such as politics, business, worldnews, technology which are focussed towards certain content place higher emphasis on content specificity and community specificity and discourage subjectivity in the title of the post. On the other hand, communities such as pics, gifs, funny etc. tend to exhibit opposite behavior. Another interesting insight is that the coefficients placed on the content attribute are high for communities such as pics, gifs, funny etc. and low for communities such as politics, technology, games etc. This is rather surprising. When I delved deeper, I realized that the reason for this kind of behavior is that the titles in the communities such as pics are not enough descriptive to discriminate the content or the image, hence the actual image or the content becomes important. On the other hand, in communities such as politics, technology etc. titles tend to be enough descriptive to indicate what the content is. Hence, the importance given to the topic is split between the title and the actual content.

Conclusion

In this report, I have presented in detail my work on analyzing popularity of a given post. I have outlined methods which are suitable for the prediction of initial popularity and also popularity at subsequent time instants. This work is a preliminary attempt at understanding the richness of the content on Reddit. As is evident from this writeup, there are several factors which are contributing to the popularity of the content on reddit. I have made an effort to analyze their interplay and propose a framework with all these aspects tied together. There is definitely much

left to be done in terms of exploring the factors a lot more rigorously and coming up with more scalable models for handling the data on Reddit.

References

[1] Roja Bandari, Sitaram Asur, Bernardo A Huberman: *The Pulse of News in Social Media: Forecasting Popularity*. In ICWSM, 2012.

[2] Yoav Artzi, Patrick Pantel, Michael Gamon: *Predicting Responses to Microblog Posts*. In HLT-NAACL 2012: 602-606.

[3] Jong Gun Lee, Sue Moon, Kave Salamatian. *An Approach to Model and Predict the Popularity of Online Contents with Explanatory Factors*. In 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2010.

[4] Himabindu Lakkaraju, Jitendra Ajmera. *Attention Prediction on Social Media Brand Pages*. In CIKM 2011.

[5] Sitaram Asur, Bernardo A Huberman: *Predicting the Future With Social Media*. In WI-AIT, 2010.

[6] Kristina Lerman, Aram Galstyan: *Analysis of Social Voting Patterns on Digg*. In WOSN, 2008.

[7] Tad Hogg, Kristina Lerman: *Social Dynamics of Digg*. In ICWSM, 2010.

[8] Kristina Lerman, Tad Hogg: *Using a Model of Social Dynamics to Predict Popularity of News*. In WWW, 2010.