What's in a name? Understanding the Interplay between Titles, Content, and Communities in Social Media

Himabindu Lakkaraju, Julian McAuley, Jure Leskovec
{himalv, jmcauley, jure}@cs.stanford.edu
Stanford University

Abstract

Creating, placing, and presenting social media content is a difficult problem. In addition to the quality of the content itself, several factors such as the way the content is presented (the title), the community it is posted to, whether it has been seen before, and the time it is posted determine its success. There are also interesting *interactions* between these factors. For example, the language of the title should be targeted to the community where the content is submitted, yet it should also highlight the distinctive nature of the content. In this paper, we examine how these factors interact to determine the popularity of social media content. We do so by studying *resubmissions*, i.e., content that has been submitted multiple times, with multiple titles, to multiple different communities. Such data allows us to 'tease apart' the extent to which each factor *influences* the success of that content. The models we develop help us understand how to better *target* social media content: by using the right title, for the right community, at the right time.

Introduction

When creating and presenting social media content, we are faced with many challenges: How should we phrase the title of our submission? Should the title be long or short, specific or generic, unique or derivative, flowery or plain? Which community should we submit our content to? At what time? How should the title change when targeting a different community? As content creators, our end-goal in asking such questions is achieving high popularity for the content we submit. Therefore, answering these questions involves analyzing the effect of each of these aspects on the popularity of the submissions. Insights stemming from such analysis are extremely valuable as it is much more appealing to a content creator to increase the chances of the content's success by changing the title, posting time, and community, rather than changing the content itself.

Naïvely, given *enough* content and *enough* submissions we might train classifiers that predict the success of social media content directly from its title, community and posting time. While this is a reasonable strategy for predicting the overall success of the content, it does not effectively capture the confounding effects arising from the complex interplay between these factors. For example, we might learn that occurrences of certain words in the title increase the chances of a submission's success (Brank and Leskovec 2003). While this kind of insight is interesting, it might not accurately tease apart the effect of the content and the title on the submission's popularity. To illustrate, consider the case of kitten images. Such images are accompanied by titles such as 'cute kittens', and while such titles are correlated with high popularity, it would not be appropriate to conclude that words such as 'kittens' contribute to a 'good' title in general. Rather, such submissions are popular because of the images themselves, and not because there is anything particularly effective about the title.

It is exactly this interplay between the content, title, community, and posting time that makes it difficult to study factors that determine the success of social media content (Artzi, Pantel, and Gamon 2012; Yang and Leskovec 2011; Yano and Smith 2010). Our goal in this paper is to directly study this interplay by observing how well the same content performs when posted with multiple different titles to multiple communities at different times. To do this, we consider submissions to the website *reddit.com*, each of which is an image, uploaded with a particular title to a particular community at a particular time, and rated by the community. The success of a Reddit submission depends on many factors other than the quality of the content. When the same content is submitted to Reddit multiple times, the popularity of each submission is not independent: an image is far less likely to be popular the twentieth time it is submitted regardless of how well the other factors are accounted for.

A particularly unique and important property of our dataset is that every image we consider has been submitted *multiple times*, with *multiple titles* to *multiple communities*. This means that our dataset represents a natural large-scale experiment, which allows us to tease apart the quality inherent in the content itself, and directly assess the extent to which factors such as the title, community and posting time *influence* the popularity of the content.

We develop a statistical model which accounts for the effect of four factors all of which play an important role in influencing the popularity of online social media content: (1) the content of the submission, (2) the submission title,

(3) the community where the submission is posted, and (4) the time when it is posted. Our approach consists of two main components. First is a *community model* that accounts for factors such as the number of times an image has been submitted previously, the time of day it is submitted, and the choice of communities that it has been submitted to. The second is the *language model* that accounts for the quality of the title. A good title is then considered to be one that further improves the success of that submission.

Our community model accounts for all the factors that influence a submission's success *other* than the title itself. The language model then uncovers properties of good titles. For instance, we discover that the vocabulary used in the titles should account for the preferences of the targeted community, yet the titles should be novel compared to the previous submissions of the same content within the community. Such findings have applications when targeting existing content (e.g. movies, books, cars) to new markets, though our findings are sufficiently general that they can be applied even when brand new content is submitted.

A Motivating Example. An example of the type of data we collect is shown in Figure 1. Here, the same image (of a bear riding Abraham Lincoln) is submitted to Reddit 25 times, with several different titles, in several communities ('subreddits'). The top plot shows the popularity of each resubmission (the number of upvotes minus the number of downvotes), together with our community model's prediction about how well each submission ought to do *without* knowing what title was used.

From this picture we begin to see the complex interplay between content, community, and title. The choice of title and community has a significant impact on whether a submission becomes popular. The title 'Merica!' (or some variation) is used four times, but is successful only the *second* time it is used. Thus its success must be due to factors such as the time of day the image was submitted or the community it was submitted to.

The purpose of our community model is to account for such factors. We make use of features such as the number of times an image has been previously submitted (resubmissions of the same content are naturally less likely to be popular than original submissions); the community (subreddit) the image was submitted to (submissions to more active subreddits have the potential to become more popular, but also face more competition); and the time of day of the submission (submissions posted during certain hours of the day are more popular).

In this example (Fig. 1), the submission is first successful the fifth time it submitted. However, our community model predicts a similar level of success (due to factors such as the subreddit and time of day etc.), thus this initial success was *not* due to the title. The submission first achieves major success the tenth time it is submitted. Indeed, our community model predicted that the tenth submission would be popular, though it was even more popular than expected. Later, the submission achieves renewed popularity when a novel title is proposed ('God bless whoever makes these'). It is precisely these effects (Fig. 1, bottom) that we use to determine

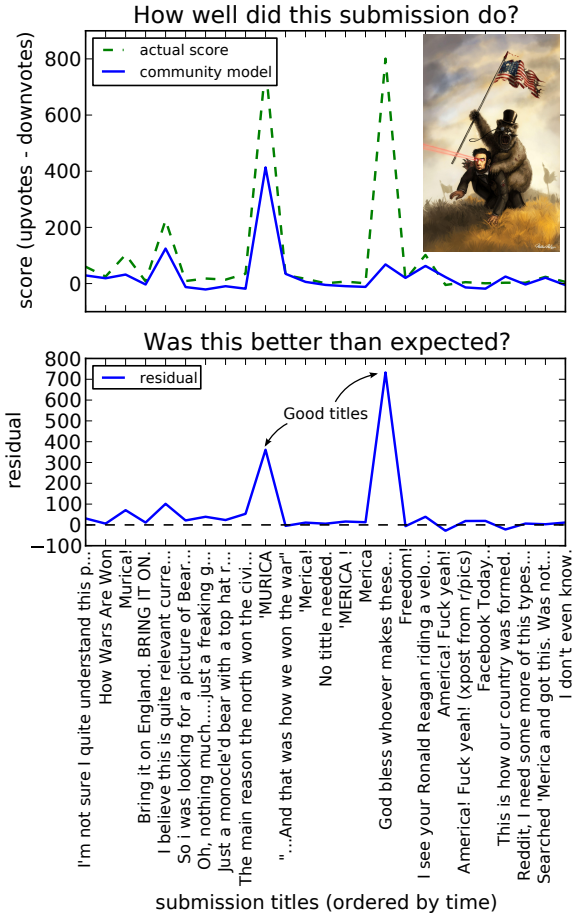

Figure 1: Submissions of the same image (a bear riding Abraham Lincoln, top right) to Reddit, with different titles. The top plot shows how successful each submission is, along with our community model's prediction of its popularity *without* knowing the title. Our community model accounts for factors such as the number of previous submissions of the same content, the community (i.e., 'subreddit') the image is submitted to, and the time of day of the submission.

which titles and communities are good, having factored out the inherent popularity of the content itself.

Contribution and Findings

Our main contribution is to study the effect that titles, submission times, and choices of community have on the success of social media content. To do so, we introduce a novel dataset of ∼132K Reddit submissions with a unique property: every piece of content has been submitted *multiple* times. This number is made up of ∼16.7K *original* submissions, each of which is resubmitted ∼7 times on average.

We use this data to disentangle how much of a submission's popularity is due to its inherent quality, and how much is due to the choice of commutniy, submission time, and title. The former part is handled by our *community model*, while the latter part is handled by our *language model*.

These two models are the main contribution of our paper.

The models we develop help us to uncover several novel results. Indeed, we confirm our intuition that good content 'speaks for itself', and can achieve popularity regardless of what title is used. Choosing a good title has a secondary—though still important—effect. We find that features such as length, descriptiveness, and even sentence structure can be predictive of whether a title will be successful. We find that the choice of community also plays a major role.

Related Work

Predicting the popularity of social media content is an important problem, and has been approached from many angles. One approach is to use measurements of an item's early popularity, such as view counts on *youtube* and *digg.com* (Szabo and Huberman 2010) to predict its future success (Lee, Moon, and Salamatian 2010; Tatar et al. 2011), though others have tried to forecast popularity *before* an item is submitted (Tsagkias, Weerkamp, and de Rijke 2009; Bandari, Asur, and Huberman 2012). Others have even used such media to predict the outcomes of external events, such as using tweets to predict box-office revenues (Asur and Huberman 2010).

Works that predict the future success of social media content include (Bandari, Asur, and Huberman 2012; Tsagkias, Weerkamp, and de Rijke 2009; Yano and Smith 2010). Typically, the goal of such works is to predict the popularity, or the number of comments that will be generated by an article based on its content. Although this is a similar goal to our own, such approaches differ from ours in that they typically rely on the *content* of the article rather than its title. For instance, in (Yano and Smith 2010) the authors learn that words like 'Obama' are predictive of comment volume: while a valuable insight, this is precisely what we do *not* want to learn—such content is successful because it discusses Obama—*not* because it has a clever title or was posted to a particularly appropriate community.

Many authors have used Twitter to study the relationship between language and social engagement (Boyd, Golder, and Lotan 2010; Danescu-Niculescu-Mizil, Gamon, and Dumais 2011; Hong, Dan, and Davison 2011; Petrovic, Osborne, and Lavrenko 2011; Suh et al. 2010). For example, in (Artzi, Pantel, and Gamon 2012), the authors use language models (in addition to social network data) to predict which tweets will be successful. This approach is related to our own, in that the authors consider a similar predictive task using lexical features, however the data is critically different—in the case of short messages such as those on Twitter, there is no meaningful distinction between title and content.

Another related study is that of (Danescu-Niculescu-Mizil et al. 2012), which considers how phrasing effects the memorability of a quotation. Among other findings they discover that a memorable phrase should use less-common word choices, but should be syntactically familiar. Although their problem setting is quite different, this is similar to our own finding that a successful title is one that employs novel words, yet at the same time conforms to the linguistic norms of the community to which it is submitted.

To our knowledge, few studies use *Reddit* as a source of data (Wang, Ye, and Huberman 2012; Gilbert 2013). However, other works have considered similar 'social news' sites (i.e., sites where users rate each other's submissions), such as *digg.com* (Hogg and Lerman 2010; Lerman and Galstyan 2008; Lerman and Hogg 2010). These works study features such as the relationship between 'visibility' and 'interestingness' (which is somewhat analogous to our study of the relationship between title and content), the effect of 'friendships' on social voting, and aspects of the website design to predict the popularity of social news content (Hogg and Lerman 2010; Lerman and Galstyan 2008; Lerman and Hogg 2010).

Proposed Method

We consider submissions to *reddit.com*, a community news site where users create, comment on, and evaluate content, which essentially consists of a title and a url. Feedback comes in the form of positive and negative ratings ('upvotes' and 'downvotes'). These ratings are then used to promote content, so that highly rated submissions are more visible (closer to the top of the page). Content can be posted to one of hundreds of communities, or 'subreddits'; posting to a large community means that the submission faces more competition, but also that it will be more visible *if* it becomes successful.

Dataset Description

Rather than studying all submissions, we focus on submissions of *images*, which consist of a title and an image url. For each submission we obtain metadata including the time of the submission, the user who submitted the image, the community they submitted it to, the number of upvotes and downvotes, and the comment thread associated with the image.

Naturally, we could use other types of web content other than images, though using images has several advantages over other types of web data. For example, webpage data may not be static, meaning that multiple submissions of the same url may not actually refer to the same content, while for an image submission we can more reliably detect duplicates. We note that images are the dominant form of content in many of the most active communities on Reddit.

To identify resubmissions of the same image, we use a reverse image search tool specifically designed for Reddit.[1] This allows us to discover resubmissions of the same image even if the submission urls are not the same. This tool maintains a directory of resubmitted content, from which we collect resubmissions of content dating back to 2008.

In total, we collect 132,307 images, 16,736 of which are unique. In other words, each of the images we obtain has been submitted 7.9 times on average. This data consists of roughly 5 million comments, 250 million ratings (56% upvotes, 44% downvotes), from 63 thousand users to 867 communities ('subreddits'). Our dataset is made available for public use.[2]

[1] `karmadecay.com`

[2] http://snap.stanford.edu/data/

$$\hat{A}_{h,n} = \overbrace{\beta_h + \phi_h}^{\text{inherent content popularity and resubmission decay coefficient}} \underbrace{\exp}_{\text{resubmission popularity decays exponentially (Fig. 2, center)}} \left\{ -\sum_{i=1}^{n-1} \overbrace{\frac{1}{\Delta_{i,n}^h}}^{\text{resubmission penalty disappears given enough time (Fig. 2, right)}} \underbrace{\left(\delta(c_{h,i} \neq c_{h,n})\lambda_{c_{h,i}} + \overbrace{\delta(c_{h,i} = c_{h,n})\lambda'_{c_{h,i}}}^{\text{penalty for submitting to the same community twice (Fig. 3, diagonal)}}\right)}_{\text{penalty from communities of previous submissions (Fig. 3, rows)}} \underbrace{A_{h,i}}_{\text{penalty from the success of previous submissions}} \right\} \quad (1)$$

Symbol	Description
h	an image
p	an image title
c	a community (or 'subreddit')
$V_{h,n}$	the rating (upvotes - downvotes) the image receives the n^{th} time it is submitted
$p_{h,n}$	the title used for that submission
$c_{h,n}$	the community used for that submission
$avg_{c,t}$	average popularity of submissions to community c at time t
$A_{h,n}$	the rating normalized by the overall popularity of content in that community at that time
$\hat{A}_{h,n}$	our community model's prediction of $A_{h,n}$
$y_{h,n}$	the residual $A_{h,n} - \hat{A}_{h,n}$
$\hat{y}_{h,n}$	our language model's prediction of $y_{h,n}$
$\Delta_{i,n}^h$	time between the i^{th} and the n^{th} submission of the image h

Table 1: Notation.

Problem Setup and Evaluation

Our dataset consists of images, each submitted with a certain title to a certain community. Since we are studying images that are submitted multiple times, we use h to donate a specific image, and $V_{h,n}$ to denote the success of that image the n^{th} time it was submitted. Initially we will assume that this quantity refers to the *rating* (upvotes - downvotes) of the image, but later we will show that it can straightforwardly be replaced by other measures, such as the number of comments a submission receives. The notation we use throughout this paper is briefly summarized in Table 1.

We first develop the *community model* by considering those aspects of a submission's success that are *not* related to its title. A significant factor is the choice of community ('subreddit') and the time of day of the submission. Submissions can potentially receive more attention if they are submitted to more active communities, and at busier times of day (though they will also face more competition). Figure 2 (left) shows the average popularity of submissions to some of the most active communities (and one less active one) at different times of day. From this we observe that there are vast differences between the most active communities compared to smaller communities (such as GifSound). There is also an apparent periodicity, though interestingly the peak does *not* occur when the website is most active (around 8pm UTC).

We compute the average popularity of submissions to a community c at time t (in one-hour intervals), $avg_{c,t}$. This is the quantity depicted in Figure 2 (left). Niche communities, where fewer than 50 submissions are available in any one-hour interval, are combined to use a single background parameter.

We use this quantity to normalize our output variable $V_{h,n}$. Instead of predicting $V_{h,n}$ directly, we predict

$$A_{h,n} = \frac{V_{h,n}}{avg_{c,t}}, \quad (2)$$

i.e., how much *more* popular was this submission compared to others in the same community, at the same time. Initially we performed training to predict $V_{h,n}$ directly, but found that this skewed our evaluation to favor submissions to popular communities. By normalizing the submission in this way, we attempt to treat each submission equally, regardless of whether it was submitted to a popular or a niche community.

Our goal shall be to propose a model whose predictions $\hat{A}_{h,n}$ are similar to the observed values $A_{h,n}$. We shall measure performance using the coefficient of determination (the R^2 statistic) between A and $\hat{A}$, using appropriate train/test splits, though we defer further explanation until our experimental section.

Community Model

Having accounted for the overall activity of each community at each time of day through normalization by $avg_{c,t}$, we move on to model more subtle community effects that influence a submission's popularity.

Our final community model, whose parameters we shall now explain, is shown in Equation 1. Firstly, β_h is intended to capture the inherent popularity of the image h, while all other parameters capture how the popularity changes as it is submitted multiple times in different communities.

In Figure 2 (center), we observe that on average, content becomes less popular each time it is submitted. We model this using an exponential decay function with importance ϕ_h, and decay parameters λ and λ' to be described shortly. We find that the exponential decay effect diminishes with the increase in the time interval (in days) between successive submissions, as shown in Figure 2 (right). To account for this, we use $\Delta_{i,n}^h$ to denote the time difference (in days) between the i^{th} and the n^{th} submission of the image h. The presence of $1/\Delta_{i,n}^h$ in the exponent means that the resubmission penalty gradually disappears if there is enough time between submissions.

Finally, we want to model the interaction effects between communities, i.e., the cost of resubmitting an image to the *same* community versus the cost of submitting it to a *different* community. For instance, if a submission first becomes

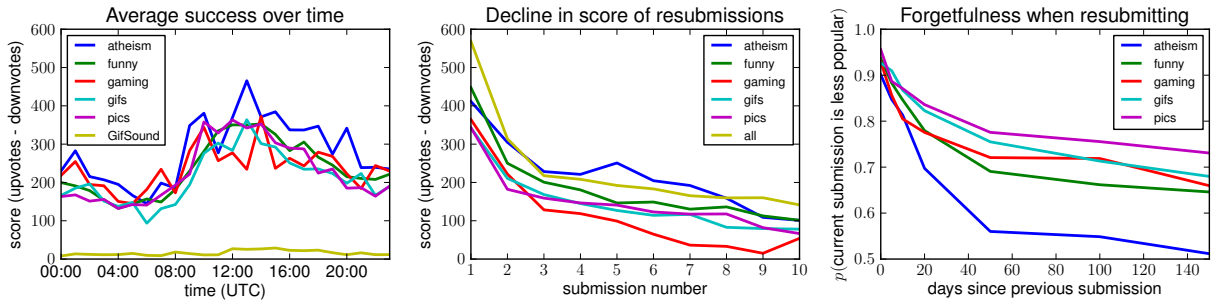

Figure 2: Three factors that affect the popularity of submissions. Submissions posted during certain hours of the day are more popular, certain communities are much more popular than others (left). Content is less likely to become popular each time it is resubmitted (center). Resubmitted content may become popular again, if there is enough time between resubmissions (right).

successful in a high-visibility community (such as the subreddit 'pics'), then it is unlikely to be successful if it is later posted to a smaller community (since users have already seen it). However, this is not symmetric, since a successful submission from a low-visibility community still has a potential audience in a larger community.

Figure 3 shows the probability that the n^{th} submission of an image is less popular ($A_{h,n} \leq 1$) given that the previous submission of the same image was successful in the community it was posted ($A_{h,n-1} > 1$). There are two main observations: Firstly, a submission following a previously successful submission of the same content within the same community is unlikely to be popular, as evidenced by high values along the main diagonal. Secondly, a submission is unlikely to be popular if it has previously been successful in a high-visibility community, as evidenced by high values in rows corresponding to popular communities. In short, content submitted to a popular community will not be successful again, whereas content submitted to a niche community may yet become popular in a *different* community.

Critically, since this interaction matrix is essentially a low-rank matrix plus a diagonal term, it is sufficient to model it using two first-order parameters. The decrease in popularity due to have seen the same content previously in a *different* community c is modeled by λ_c, and the penalty for having seen it in the *same* community is modeled by λ'_c.

We fit all of the above parameters by minimizing the least-squares criterion $\|A - \hat{A}\|_2^2$ using L-BFGS, a standard quasi-Newton procedure to optimize smooth functions of many variables (Nocedal 1980).

Language Model

Having accounted for the effects coming from the choice of the community and submisison time, we shall now move on to model *linguistic* aspects of a submission's title. At this point, our goal is to model the *impact* that a title has on a submission's success. After factoring out the effects captured by the community model, we attribute the residual popularity to the title of the submission:

$$y_{h,n} = A_{h,n} - \hat{A}_{h,n}. \quad (3)$$

This is depicted in our introduction: Figure 1 (top) depicts our community model's prediction of $\hat{A}_{h,n}$, while Figure 1 (bottom) depicts the residual $y_{h,n}$ of this prediction, i.e., the difference between the actual and the predicted value.

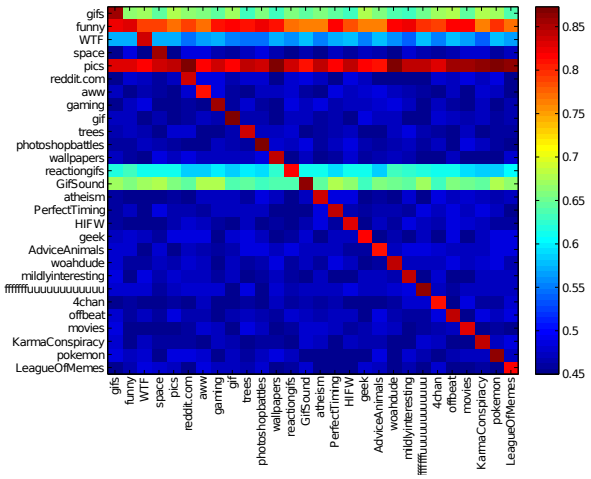

Figure 3: Community effect on successive submissions of the same content. Rows and columns are Reddit communities; an entry in row x, column y indicates the likelihood that a submission to community y will be (un)popular if the same content was previously submitted to community x and achieved a high popularity; a value of 0.9 indicates that the re-submission will be *less* popular 90% of the time.

Below we propose several features and predictors associated with the titles of submissions that can be used to model the output variable $y_{h,n}$. Later, we combine these elements in a simple way by training a linear regressor per community with weights associated with each of their predictions.

Modeling good and bad words. Each community may appreciate certain choices of words in titles and dislike others. We should be able to aggregate the effect of all the words in the title and associate it with the residual popularity $y_{h,n}$. In order to achieve this, we extend the supervised LDA framework (Blei and McAuliffe 2007) to learn 'topics' that are correlated with our response variable $y_{h,n}$.

A high level view of the model we propose segments words into three different topics, each of which contributes either in a favorable (a topic of 'good' words), unfavorable

1. For each title p (image h, sub. n, comm. c)
   a. Sample $\theta_p \sim Dirichlet(\alpha)$
   b. for each word position i in p
      i. $z_{p,i} \sim Multinomial(\theta_p)$
      ii. $w_{p,i} \sim Multinomial(\phi_{c,z_{p,i}})$
   c. $y_p \sim Normal(\eta_c^T \theta_p, \sigma^2)$

Table 2: Generative process for associating topics with residuals. For brevity we use p and c when referring to a title $p_{h,n}$ in community $c_{h,n}$.

(a topic of 'bad' words) or a neutral way to the title's popularity in the community. Each word in a given title falls into one of these topics. In order to aggregate the effect of the different words in the title, we model the residual popularity as a function of the proportion of the words in the title belonging to each of the three topics. Further, this function incorporates the notion of accounting for higher (lower) popularity when there are more words in the title which are received favorably (unfavorably) by the community.

More formally, our topic model is defined as follows: each title is associated with a topic distribution (a stochastic vector) from which the topic of each word is sampled. A linking parameter η_c (per community) relates the affinity scores of each topic to our response variable $y_{h,n}$ so that for each topic we learn which words are likely to generate positive, negative, or neutral values of $y_{h,n}$. We learn such topics *per community*, so that each community's choices of words are accounted for appropriately. Our generative process is shown in Table 2. The inference procedure for estimating the various latent parameters in this model is discussed later.

Modeling community and content specific words. An important aspect when studying submission titles is 'specificity'. Three different kinds of words come into play when phrasing a title—words which are specific to the image h, words which are specific to the community c, and 'generic' syntactic words which tie together the content and community specific words. Understanding how much content or community specificity is necessary for a good title is crucial. Further, this might be different for different communities.

In order to quantify the content and community specificity of a given title and associate it with the residual popularity $y_{h,n}$, we propose a topic model which again extends the supervised LDA framework of (Blei and McAuliffe 2007). Our model associates with each word in the title a latent variable $spec_{p,i}$ which indicates if the word is a generic word ($spec_{p,i} = 0$), a community specific word ($spec_{p,i} = 1$) or a content specific word ($spec_{p,i} = 2$). Further, we associate the residual popularity with the specificity by modeling $y_{h,n}$ as a function of the proportion of the words in the title belonging to each of these three categories.

Our model can be formally explained as follows: for each title, we sample a stochastic vector of length three corresponding to the proportion of generic, community specific, and content specific words used in the title. For each word in the title, we sample from this stochastic vector to determine which word distribution to use. In this way we learn

1. For each title p (image h, sub. n, comm. c)
   a. Sample $\theta_p \sim Dirichlet(\alpha)$
   b. for each word position i in p
      i. $spec_{p,i} \sim Multinomial(\theta_p)$
      if ($spec_{p,i} = 0$)
         $w_{p,i} \sim Multinomial(\phi^{generic})$
      else if ($spec_{p,i} = 1$)
         $w_{p,i} \sim Multinomial(\phi_c^{community})$
      else
         $w_{p,i} \sim Multinomial(\phi_h^{content})$
   c. $y_p \sim Normal(\eta_c^T \theta_p, \sigma^2)$

Table 3: Generative Process for specificity.

topics for each item h, each community c, as well as a background topic containing generic words. These topic distributions are tied to our output variable by modeling $y_{h,n}$ as a normal variable whose mean depends upon the proportion of content specific, community specific and generic words. and a linking parameter η_c^T (per community). The complete generative process is shown in Table 3.

Topic model inference. The models described above are motivated by the supervised topic models framework (Blei and McAuliffe 2007). We employ a collapsed Gibbs sampling approach (Griffiths and Steyvers 2004) to estimate the latent topics and topic distributions of titles. The hyperparameters α and β (note that β is a prior for ϕ) are set to the values of 0.01 and $0.1/K$ (where K is the number of topics) respectively. The parameters η and σ^2 are estimated by maximizing the likelihood after the latent variables are sampled. The newly estimated parameters are in turn used to sample the latent variables, and this process is repeated until the log-likelihood converges.

Other Linguistic Features

Part of speech tags. Certain communities may favor highly descriptive, adjective laden titles, while others may prefer simpler titles consisting of a few nouns. We use binary features to indicate the presence or absence of each part-of-speech, in order to measure how much they influence the success of a title in each community. Specifically, we consider determiners, pronouns, nouns, adjectives, adverbs, interjections, and prepositions.

Sentiment. We found that 'positive' sentiment contributes to a title's popularity in certain communities. In order to capture the sentiment of a title, we employed a hierarchical classification approach (Pang and Lee 2004) which first identifies if a given piece of text is subjective or not and then categorizes the subjective text into positive and negative sentiment classes. This enables us to classify each title into positive, negative and neutral sentiment classes. In this work, we used an off-the-shelf implementation of this algorithm.[3]

Length of the title. We found that the length of a title does not significantly impact the popularity of a submission's success, unless the title is either extremely long or extremely

[3] www.lingpipe.com

short. Therefore we use a simple binary feature that indicates whether a title is very short (fewer than four words), or very long (more than sixteen words).

Weighted Jaccard similarity. Finally, we consider the Jaccard similarity of the current title compared to previous submissions of the same content. Much as we did in (eq. 1), we learn two parameters μ, and μ' to measure the effect of this feature, depending on whether the post was submitted to the same, or to a different community. Specifically, the parameters μ and μ' measure these effects according to

$$\sum_{i=0}^{n-1} \overbrace{\mu\ \delta(c_{h,i} \neq c_{h,n}) \frac{J(p_{h,n}, p_{h,i})}{\Delta_{i,n}^{h}}}^{\text{similarity compared to submissions in different communities}} + \underbrace{\mu'\ \delta(c_{h,i} = c_{h,n}) \frac{J(p_{h,n}, p_{h,i})}{\Delta_{i,n}^{h}}}_{\text{similarity compared to submissions in the same community}}, \quad (4)$$

where $J(p, q)$ is the Jaccard similarity between the titles p and q.

Our final language model is a linear combination of all the above factors, including predictions made by our supervised topic models. Again, we fit the parameters of our language model by minimizing the least-squares criterion $\|y - \hat{y}\|_2^2$.

Experiments

Quantitative Evaluation

We use our models to predict three measures of a submission's success: the *rating* (upvotes - downvotes) that the image receives, the *attention* (upvotes + downvotes), and the *engagement* (total number of comments). In the previous sections we assumed that the output variable $V_{h,n}$ referred to the rating (and our analysis shall focus on this variable), though our methods are easily adapted to other variables.

Training proceeds as described in the previous section: we first normalize the output variables $V_{h,n}$ to account for the overall popularity of submissions in a certain community at a certain time, producing $A_{h,n}$. Our community model produces predictions $\hat{A}_{h,n}$, and finally we use our language model to fit the residuals $y_{h,n}$.

We evaluate our models by computing the coefficient of determination

$$R^2(x, \hat{x}) = 1 - \frac{\sum_i (x_i - \hat{x}_i)^2}{\sum_i (x_i - \overline{x})^2}, \quad (5)$$

which measures how accurately our regression fits the data (a value of 1.0 indicates a perfect fit). Since our language model fits the *residuals* of our community model, we report $R^2(A, \hat{A})$ when evaluating our community model, and $R^2(A, \hat{A} + \hat{y})$ when evaluating our language model.

To ensure that we are not overfitting to our dataset, we also evaluate our model when trained using two train/test splits: in the first split, we train our model for each image h using all but the two most recent submissions of the image h, and then evaluate it on the held-out submissions. Although testing on the most recent submissions most accurately captures the performance we would expect if we were to employ our algorithm 'in the wild' today, it may be biased in the sense that the most recent submissions are unlikely to be the most popular (since they are by definition content that has been submitted many times). For this reason, we also evaluate our model on a *randomly* selected test set of the same size. In both cases we report R^2 on the test set.

Baselines. In addition to our community and language model, we consider the following four baselines:

- *Simple decay model:* Models the success of the n^{th} resubmission using a simple exponential functin $\phi_h e^{-\lambda n}$.
- *Per-community decay:* Same as the above model, but adds a per-community decay rate. The success of the n^{th} resubmission is modeled as $\phi_h\ exp\{-\sum_{i=1}^{n-1} \lambda_{c_{h,i}}\}$
- *+forgetfulness:* Same as the above model, but adds the forgetfulness parameter $\Delta_{i,n}^{h}$. The success of the n^{th} resubmission is modeled as $\phi_h\ exp\{-\sum_{i=1}^{n-1} \frac{\lambda_{c_{h,i}}}{\Delta_{i,n}^{h}}\}$
- *Language-only model:* Uses our language model to predict $A_{h,n}$ *directly*, rather than using it to predict residuals.

The language-only model is intended to demonstrate a claim we made in our introduction, namely that trying to train a model to directly predict the quality of a title is not effective, but rather we need to explicitly control for the effect of content and language separately.

Results for each of these baselines (as well as our final models) are shown in Table 4. We see that each of our baselines gradually builds towards the performance of our final community model, indicating that each of the components in (eq. 1) is a meaningful contribution to our final prediction.

A simple decay model (our first baseline) is ineffective, achieving an average R^2 score across all tasks and testing conditions of only 0.071. Adding per-community decay terms improves this by nearly fourfold, to 0.352. Adding a forgetfulness parameter that gradually reduces the resubmission penalty further improves the performance by 40%, to 0.494. Adding the remaining terms from (eq. 1) further improves the performance by 12%, to 0.556.

Combining our language model with our community model further improves the performance by 15%, to 0.639. Thus, while our community model performs well, there is still significant variation that can be explained by modeling the linguistic features of the title. This confirms that a user's choice of title is impactful on the success of that submission.

On the other hand, trying to predict the success of our title using our language model *alone* (i.e., without combining it with our community model) leads to extremely poor performance, with an average R^2 score of only 0.139. Thus we confirm a claim made in our introduction: namely that when trying to model what makes a title successful, it is necessary to control for the quality of the content itself.

We see similar performance across all three prediction tasks we consider (rating, attention, engagement). We also see similar performance for different train/test splits, indicating that our model is not overfitting, and ought to generalize well to novel data.

	Entire dataset			Most recent posts			Random test set		
	Rating	Atten.	Engag.	Rating	Atten.	Engag.	Rating	Atten.	Engag.
Simple decay model	.128 (.001)	.018 (.001)	.134 (.001)	.070 (.018)	.022 (.016)	.102 (.022)	.046 (.022)	.018 (.021)	.097 (.016)
Per-community decay	.356 (.005)	.317 (.005)	.297 (.005)	.392 (.011)	.296 (.012)	.452 (.011)	.346 (.011)	.283 (.011)	.429 (.012)
+ forgetfulness	.522 (.002)	.412 (.002)	.563 (.002)	.534 (.004)	.423 (.004)	.554 (.004)	.476 (.005)	.447 (.004)	.519 (.005)
Community model	.543 (.000)	.496 (.000)	.604 (.000)	.596 (.001)	.484 (.001)	.684 (.001)	.528 (.001)	.463 (.001)	.602 (.001)
Language-only model	.189 (.009)	.128 (.010)	.208 (.009)	.108 (.022)	.114 (.019)	.186 (.022)	.081 (.020)	.109 (.023)	.127 (.020)
Community + language	**.648 (.000)**	**.601 (.000)**	**.708 (.000)**	**.622 (.001)**	**.592 (.001)**	**.698 (.001)**	**.618 (.001)**	**.583 (.001)**	**.685 (.001)**

Table 4: Coefficient of determination (R^2) for each model and for each train/test split. For each experiment we report our predictions of the rating (upvotes - downvotes), attention (upvotes + downvotes), and engagement (number of comments). Standard error values are shown in parentheses.

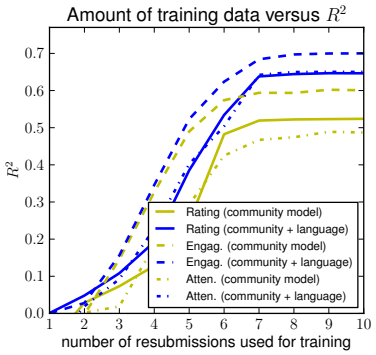

Figure 4: Performance of our community and language models using increasing amounts of training data.

In Figure 4 we show performance on the same tasks as those in Table 4, but vary the amount of data used for training. Here, we use only the first K submissions of each image in order to fit the model, whose performance we evaluate on the entire dataset. We find that around seven submissions is enough to train an accurate model, after which performance does not improve significantly.

Qualitative Evaluation

We now move on to discuss our findings of *what makes a good title*. Earlier, we suggested that a title should be targeted to a community, in the sense that it should be linguistically similar to other titles used within that community, yet at the same time it should be *novel* in order to capture readers' attention (Danescu-Niculescu-Mizil et al. 2012). We capture this effect in Figure 5, left. Here, the x-axis shows the fraction of the words in the title of the current submission which are community specific. This fraction is obtained from one of our supervised topic models (generative process shown in Table 3). The y-axis shows the probability of the submission being successful.

For most communities, we find a 'peak' in terms of how closely a submission should be targeted to that community. On either side of the peak, a submission is either poorly matched to its community (too few community specific words), or is too similar to content the community has already seen (too many community specific words). For a few communities (e.g. 'atheism', 'gaming'), there is no such dip: we might argue that such communities prefer content that is extremely closely targeted to their accepted standards.

Figure 5 (center) gives further evidence of this phenomenon. Here we compare a submission title to previous titles used *for the same image* h. Although using an original title (low Jaccard sim.) is itself not indicative of good performance, using a very *unoriginal* title (high Jaccard sim.) is a strong indication that the submission will do poorly.

Finally, Figure 5 (right) shows how individual parts-of-speech affect the success or failure of a particular title. To compute the probabilities we merely consider the *presence* or *absence* of each part of speech in a particular title. Here we see some general trends, e.g. nouns and adjectives impact the success of a title more than verbs and adverbs. However we also see that these effects differ between communities, for instance pronouns, adverbs, and determiners are highly negative in the atheism community.

In Figure 6 we visualize the outputs of one of our supervised topic models (Table 2). This topic model finds topics that are correlated with the output variable $y_{h,n}$ A naïve classifier for the same problem (that directly predicts $A_{h,n}$ from the title) is not successful, and leads to predictors that are composed largely of proper nouns which are specific to popular *content*. This is no longer an issue in our model: many of the words that our model finds to be effective are general words that could be applied to any title.

In Table 5 we look more closely at some of the titles used for a particular submission (the same one from Figure 1), and our model's explanation of *why* that title was good or bad. The main message from this figure is that the model correctly predicts that the most successful titles have positive attributes, while the least successful titles do not. For instance the title 'MURICA' (short for America) is penalized for not using language that is common for that community, but is rewarded for its originality (at least the *first* time it is submitted). In practice, we believe that such findings could be reported using an interface where a user proposes a title for their submission and our model outputs its expected success, along with its failings and qualities.

***In Situ* Evaluation.** In order to evaluate our models on Reddit directly, we carried out an experiment in which we pooled a sample of 85 images from our dataset and assigned two titles to each of them—one which our language model gave a high score (a 'good' title), and another which was given a low score ('bad' title). We posted the same image with these two titles on Reddit at approximately the same time in two different communities in order to deter-

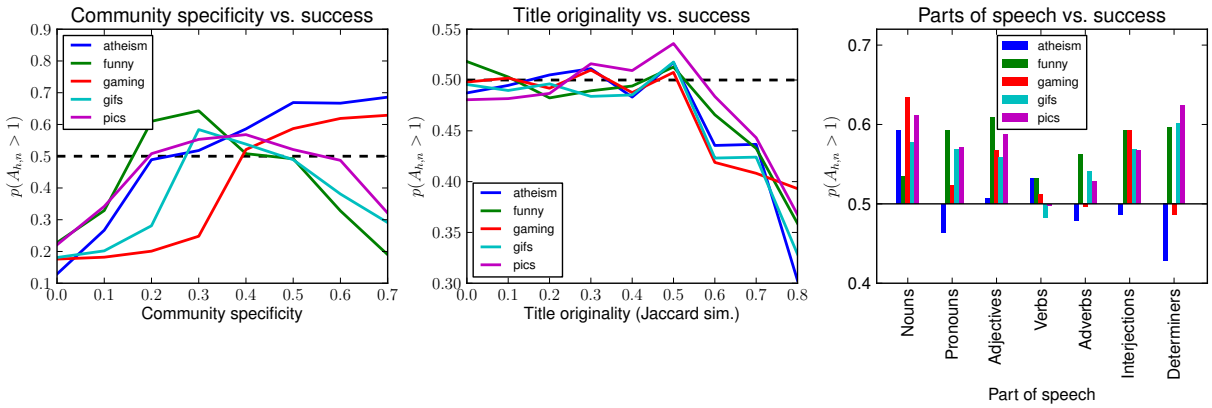

Figure 5: Three linguistic components that affect a submission's chance of being successful (i.e., more popular than average, or $A_{h,n} > 1$). Submission titles should be novel, but should still use language that is familiar to the community (left). A submission is more likely to be successful if its title is original compared to previous submissions of the same content (center). Different types of words should be used in different communities (right).

mine whether our notions of good and bad submissions as determined by our models match those of the real world.

After one day, we collected rating, attention and engagement metrics for these submissions. The R^2 values for each of these metrics with our community + language model were 0.565, 0.588, and 0.647 respectively. Overall, the total rating of the 'good' submissions (which is the sum of the ratings of all the 'good' submissions) was about three times higher than that of the 'bad' ones (10,959 vs. 3,438). Moreover, two of our 'good' submissions reached the Reddit front page and three others were on the front pages of their respective communities.[4]

Discussion

Although we have described how our dataset can be used to develop models of whether a title is well-matched to its content and community, we have yet to describe how this type of data might be used in the absence of such a submission history. Firstly, many of our findings are quite general, for instance the fact that a title should match a community's expectations—yet not be *too* dissimilar to other titles within that community—is a finding that can easily be applied without observing multiple submissions of the same content.

Secondly, we believe that there exist real-world applications where one *does* have access to data such as ours. For example, certain products, such as movies, books, and cars, are given different names when distributed in different markets. With a model such as ours, we could learn how to select a good name for a particular market, based on the success of previous names of that product in different markets.

We have concerned ourselves with *classifying* whether titles are good, though a natural next step is to automatically *generate* good titles. We believe that these goals are not orthogonal. Current work on title generation is generally concerned with whether generated titles are *accurate*, in that they meaningfully summarize the content they refer to (Tseng et al. 2006; Jin and Hauptmann 2002; 2001). Our methods could be combined with such approaches, to identify titles that are accurate, yet are also interesting.

A dataset such as ours could also be used to study other problems where content is a confounding variable. For example, how does the title of a submission influence the tone of a discussion about its content? Are popular users popular *because* they submit better content, or is their content more successful merely because they are popular?

[4]In Reddit terminology, a 'front page' refers to the page hosting the most popular content

Conclusion

When movies, books, or social media submissions become successful, it is difficult to assess *how much* of that success is due to the quality of the content, and how much is due to having used a good title, at the right time, in the right community. We have developed models that disentangle the interplay between these factors. We proposed a novel dataset of images posted to *reddit.com*, each of which has been submitted *multiple times*, with *multiple titles*. We developed *community models* to account for effects *other than* the title, such as the choice of community the image is submitted to and the number of times it has been submitted. We also developed *language models* that account for how the success of each submission was *influenced* by its title, and how well that title was targeted to the community. These models allowed us to study features of good title, and to understand *when*, *where*, and *how* a submission should be targeted.

Acknowledgements This research has been supported in part by NSF IIS-1016909, CNS-1010921, CAREER IIS-1149837, IIS-1159679, ARO MURI, Docomo, Boeing, Allyes, Volkswagen, Intel, Okawa Foundation, Alfred P. Sloan Fellowship and the Microsoft Faculty Fellowship.

References

Artzi, Y.; Pantel, P.; and Gamon, M. 2012. Predicting responses to microblog posts. In *NAACL*.

Asur, S., and Huberman, B. 2010. Predicting the future with social media. In *WI-IAT*.

Bandari, R.; Asur, S.; and Huberman, B. 2012. The pulse of news in social media: Forecasting popularity. In *ICWSM*.

Figure 6: 'Good' and 'bad' words discovered by our supervised topic model, i.e., words likely to lead to a high or a low residual $y_{h,n}$ (respectively). Results are shown for three communities. Word size is proportional to likelihood.

Title	Comm.	SW	SB	SC	LC	O	POS
I'm not sure I quite understand this piece.	WTF	+	.	– –	.	+	.
I believe this is quite relevant currently..	funny	+	+	–	++	.	.
The main reason the north won the civil war.	funny	.	+	.	++	–	+
'MURICA	funny	++	++	.	– –	++	–
...And that was how we won the war	pics	.	–	+	.	.	+
No tittle needed.	pics	– –	–	.	+	.	–
'MERICA !	pics	.	.	.	–	– –	– –
God bless whoever makes these...	funny	++	++	++	+	+	.
Freedom!	WTF	.	– –	.	–	.	– –
I see your Ronald Reagan riding a velocirap-tor, and raise you a bear riding Abe Lincoln.	funny	+	+	.	.	.	++

SW: Scored Well
SB: Scored Better than expected
SC: Specific to the Community
LC: uses words Liked by the Community
O: title is Original
POS: Parts-of-Speech are appropriate

Table 5: Some titles and our model's explanations for their success. Titles are selected from Figure 1. The first two values (SW, SB) indicate whether the title received a high rating, and whether it received a *better* rating than our community model predicted (i.e., whether it is a 'good' title). The remaining values are four of our language model's predictors. Each predictor's score is summarized using a percentile: top 10% (++); top 25% (+); bottom 25% (–) bottom 10% (– –).

Blei, D., and McAuliffe, J. 2007. Supervised topic models. In *NIPS*.

Boyd, D.; Golder, S.; and Lotan, G. 2010. Tweet, tweet, retweet: Conversational aspects of retweeting on twitter. In *Hawaii International Conference on System Sciences*.

Brank, J., and Leskovec, J. 2003. The download estimation task on KDD Cup 2003. *SIGKDD Explorations*.

Danescu-Niculescu-Mizil, C.; Cheng, J.; Kleinberg, J.; and Lee, L. 2012. You had me at hello: How phrasing affects memorability. In *ACL*.

Danescu-Niculescu-Mizil, C.; Gamon, M.; and Dumais, S. 2011. Mark my words! Linguistic style accommodation in social media. In *WWW*.

Gilbert, E. 2013. Widespread underprovision on reddit. In *CSCW*.

Griffiths, T., and Steyvers, M. 2004. Finding scientific topics. *PNAS*.

Hogg, T., and Lerman, K. 2010. Social dynamics of digg. In *ICWSM*.

Hong, L.; Dan, O.; and Davison, B. 2011. Predicting popular messages in twitter. In *WWW*.

Jin, R., and Hauptmann, A. 2001. Automatic title generation for spoken broadcast news. In *HLT*.

Jin, R., and Hauptmann, A. 2002. A new probabilistic model for title generation. In *COLING*.

Lee, J. G.; Moon, S.; and Salamatian, K. 2010. An approach to model and predict the popularity of online contents with explanatory factors. In *WI-IAT*.

Lerman, K., and Galstyan, A. 2008. Analysis of social voting patterns on digg. In *WOSN*.

Lerman, K., and Hogg, T. 2010. Using a model of social dynamics to predict popularity of news. In *WWW*.

Nocedal, J. 1980. Updating quasi-Newton matrices with limited storage. *Mathematics of Computation*.

Pang, B., and Lee, L. 2004. A sentimental education: Sentiment analysis using subjectivity. In *ACL*.

Petrovic, S.; Osborne, M.; and Lavrenko, V. 2011. Rt to win! Predicting message propagation in twitter. In *ICWSM*.

Suh, B.; Hong, L.; Pirolli, P.; and Chi, E. 2010. Want to be retweeted? large scale analytics on factors impacting retweet in twitter network. In *SocialCom*.

Szabo, G., and Huberman, B. 2010. Predicting the popularity of online content. *ACM*.

Tatar, A.; Leguay, J.; Antoniadis, P.; Limbourg, A.; de Amorim, M. D.; and Fdida, S. 2011. Predicting the popularity of online articles based on user comments. In *WIMS*.

Tsagkias, M.; Weerkamp, W.; and de Rijke, M. 2009. Predicting the volume of comments on online news stories. In *CIKM*.

Tseng, Y.-H.; Lin, C.-J.; Chen, H.-H.; and Lin, Y.-I. 2006. Toward generic title generation for clustered documents. In *AIRS*.

Wang, C.; Ye, M.; and Huberman, B. 2012. From user comments to on-line conversations. In *KDD*.

Yang, J., and Leskovec, J. 2011. Patterns of temporal variation in online media. In *WSDM*.

Yano, T., and Smith, N. 2010. What's worthy of comment? content and comment volume in political blogs. In *ICWSM*.