Talking to the crowd: What do people react to in online discussions?

Aaron Jaech, Vicky Zayats, Hao Fang, Mari Ostendorf and Hannaneh Hajishirzi
Dept. of Electrical Engineering
University of Washington
{ajaech,vzayats,hfang,ostendor,hannaneh}@uw.edu

Abstract

This paper addresses the question of how language use affects community reaction to comments in online discussion forums, and the relative importance of the message vs. the messenger. A new comment ranking task is proposed based on community annotated karma in Reddit discussions, which controls for topic and timing of comments. Experimental work with discussion threads from six subreddits shows that the importance of different types of language features varies with the community of interest.

1 Introduction

Online discussion forums are a popular platform for people to share their views about current events and learn about issues of concern to them. Discussion forums tend to specialize on different topics, and people participating in them form communities of interest. The reaction of people within a community to comments posted provides an indication of community endorsement of opinions and value of information. In most discussions, the vast majority of comments spawn little reaction. In this paper, we look at whether (and how) language use affects the reaction, compared to the relative importance of the author and timing of the post.

Early work on factors that appear to influence crowd-based judgments of comments in the Slashdot forum (Lampe and Resnick, 2004) indicate that timing, starting score, length of the comment, and poster anonymity/reputation appear to play a role (where anonymity has a negative effect). Judging by differences in popularity of various discussion forums, topic is clearly important. Evidence that language use also matters is provided by recent work (Danescu-Niculescu-Mizil et al., 2012; Lakkaraju et al., 2013; Althoff et al., 2014; Tan et al., 2014). Teasing these different factors apart, however, is a challenge. The work presented in this paper provides additional insight into this question by controlling for these factors in a different way than previous work and by examining multiple communities of interest. Specifically, using data from Reddit discussion forums, we look at the role of author reputation as measured in terms of a karma k-index, and control for topic and timing by ranking comments in a constrained window within a discussion.

The primary contributions of this work include findings about the role of author reputation and variation across communities in terms of aspects of language use that matter, as well as the problem formulation, associated data collection, and development of a variety of features for characterizing informativeness, community response, relevance and mood.

2 Data

Reddit[1] is the largest public online discussion forum with a wide variety of subreddits, which makes it a good data source for studying how textual content in a discussion impacts the response of the crowd. On Reddit, people initiate a discussion thread with a post (a question, a link to a news item, etc.), and others respond with comments. Registered users vote on which posts and comments are important. The total amount of up votes minus the down votes (roughly) is called *karma*; it provides an indication of community endorsement and popularity of a comment, as used in (Lakkaraju et al., 2013). Karma is valued as it impacts the order in which the posts or comments are displayed, with the high karma content rising to the top. Karma points are also accumulated by members of the discussion forum as a function of the karma associated with their comments.

[1] http://www.reddit.com

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 2026–2031,
Lisbon, Portugal, 17-21 September 2015.

subreddit	# Posts	# Comments/Post
FITNESS	3K	16.3
ASKSCIENCE	4K	8.8
POLITICS	7K	23.7
ASKWOMEN	4K	50.5
ASKMEN	4K	58.3
WORLDNEWS	12K	26.1

Table 1: Data collection statistics.

	Top1	Top3
ASKSCIENCE	9.3	25.9
FITNESS	1.4	12.3
POLITICS	0.3	7.4
ASKWOMEN	2.2	13.5
ASKMEN	3.9	11.9
WORLDNEWS	3.1	6.4

Table 2: Percentage of discussions where the top comment is made by the top k-index person (or top 3 people) in the discussion.

The Reddit data is highly skewed. Although there are thousands of active communities, only a handful of them are large. Similarly, out of the more than a million comments made per day[2], most of them receive little to no attention; the distributions of positive comment karma and author karma are Zipfian. Slightly more than half of all comments have exactly one karma point (no votes beyond the author), and only 5% of comments have less than one karma point.

For this study, we downloaded all the posts and associated comments made to six subreddits over a few weeks, as summarized in Table 1, as well as karma of participants in the discussion[3]. All available comments on each post were downloaded at least 48 hours after the post was made.[4]

3 Uptake Factors

Factors other than the language use that influence whether a comment will have uptake from the community include the topic, the timing of the message, and the messenger. These factors are all evident in the Reddit discussions. Some subreddits are more popular and thus have higher karma comments than others, reflecting the influence of topic. Comments that are posted early in the discussion are more likely to have high karma, since they have more potential responses.

Previous studies on Twitter show that the reputation of the author substantially increases the chances of the retweet (Suh et al., 2010; Cha et al., 2010), and reputation is also raised as a factor in Slashdot (Lampe and Resnick, 2004). On Reddit most users are anonymous, but it is possible that members of a forum become familiar with particular usernames associated with high karma comments. In order to see how important personal reputation is, we looked at how often the top karma comments are associated with the top karma participants in the discussion. Since an individual's karma can be skewed by a few very popular posts, we measure reputation instead using a measure we call the *k-index*, defined to be equal to the number of comments in each user's history that have karma $\geq k$. The k-index is analgous to the h-index (Hirsch, 2005) and arguably a better indicator of extended impact than total karma.

The results in Table 2 address the question of whether the top karma comments always come from the top karma person. The Top1 column shows the percentage of threads where the top karma comment in a discussion happens to be made by the highest k-index person participating in the discussion; the next column shows the percentage of threads where the comment comes from any one of the top 3 k-index people. We find that, in fact, the highest karma comment in a discussion is rarely from the highest k-index people. The highest percentage is in ASKSCIENCE, where expertise is more highly valued. If we consider whether any one of the multiple comments that the top k-index person made is the top karma comment in the discussion, then the frequency is even lower.

4 Methods

4.1 Tasks

Having shown that the reputation of the author of a post is not a dominating factor in predicting high karma comments, we propose to control for topic and timing by ranking a set of 10 comments that were made consecutively in a short window of time within one discussion thread according to the karma they finally received. The ranking has access to the comment history about these posts. This simulates the view of an early reader of these posts, i.e., without influence of the ratings of oth-

[2]http://www.redditblog.com/2014/12/reddit-in-2014.html

[3]Our data collection is available online at https://ssli.ee.washington.edu/tial/data/reddit

[4]Based on our initial look at the data, we noticed that most posts receive all of their comments within 48 hours. Some comments are deleted before we are able to download them.

ers, so that the language content of the post is more likely to have an impact. Very long threads are sampled, so that these do not dominate the set of lists. Approximately 75% of the comment lists are designated for training and the rest is for testing, with splits at the discussion thread level. Here, feature selection is based on mean precision of the top-ranked comment (P@1), so as to emphasize learning the rare high karma events. (Note that P@1 is equivalent to accuracy but allows for any top-ranking comment to count as correct in the case of ties.) The system performance is evaluated using both P@1 and normalized discounted cumulative gain (NDCG) (Burges et al., 2005), which is a standard criterion for ranking evaluation when the samples to be ranked have meaningful differences in scores, as is the case for karma of the comments.

In addition, for analysis purposes, we report results for three surrogate tasks that can be used in the ranking problem: i) the binary ranker trained on all comment pairs within each list, in which low karma comments dominate, ii) a positive vs. negative karma classifier, and iii) a high vs. medium karma classifier. All use class-balanced data; the second two are trained and tested on a biased sampling of the data, where the pairs need not be from the same discussion thread.

4.2 Classifier

We use the support vector machine (SVM) rank algorithm (Joachims, 2002) to predict a rank order for each list of comments. The SVM is trained to predict which of a pair of comments has higher karma. The error term penalty parameter is tuned to maximize P@1 on a held-out validation set (20% of the training samples).

Since much of the data includes low-karma comments, there will be a tendancy for the learning to emphasize features that discriminate comments at the lower end of the scale. In order to learn features that improve P@1, and to understand the relative importance of different features, we use a greedy automatic feature selection process that incrementally adds one feature whose resulting feature set achives the highest P@1 on the validation set. Once all features have been used, we select the model with the subset of features that obtains the best P@1 on the validation set.

4.3 Features

The features are designed to capture several key attributes that we hypothesize are predictive of comment karma motivated by related work. The features are categorized in groups as summarized below, with details in supplementary material.

- **Graph and Timing (G&T):** A baseline that captures discourse history (response structure) and comment timing, but no text content.
- **Authority and Reputation (A&R):** K-index, whether the commenter was the original poster, and in some subreddits "flair" (display next to a comment author's username that is subject to a cursory verification by moderators).
- **Informativeness (Info.):** Different indicators suggestive of informative content and novelty, including various word counts, named entity counts, urls, and unseen n-grams.
- **Lexical Unigrams (Lex.):** Miscellaneous word class indicators, puncutation, and part-of-speech counts
- **Predicted Community Response (Resp.):** Probability scores from surrogate classification tasks (reply vs. no reply, positive vs. negative sentiment) to measure the community response of a comment using bag-of-words predictors.
- **Relevance (Rel.):** Comment similarity to the parent, post and title in terms of topic, computed with three methods: i) a distributed vector representation of topic using a non-negative matrix factorization (NMF) model (Xu et al., 2003), ii) the average of skip-gram word embeddings (Mikolov et al., 2013), and iii) word set Jaccard similarity (Strehl et al., 2000).
- **Mood:** Mean and std. deviation of sentence sentiment in the comment; word list indicators for politeness, argumentativeness and profanity.
- **Community Style (Comm.):** Posterior probability of each subreddit given the comment using a bag-of-words model.

The various word lists are motivated by feature exploration studies in surrogate tasks. For example, projecting words to a two dimensional space of positive vs. negative and likelihood of reply showed that self-oriented pronouns were more likely to have no response and second-person pronouns were more likely to have a negative response. The politeness and argumentativeness/profanity lists are generated by starting with hand-specified seed lists used to train an SVM to classify word embeddings (Mikolov et al., 2013)

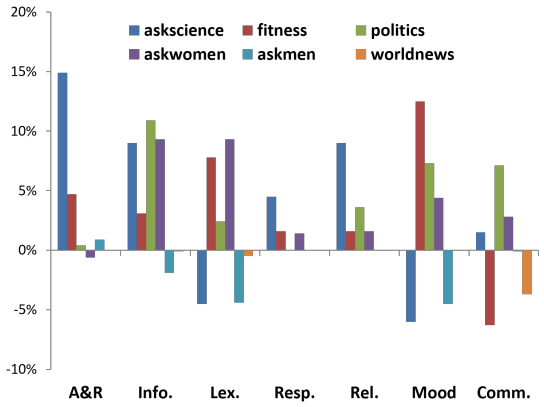

Figure 1: Relative improvement in P@1 over G&T for individual feature groups.

into these categories, and expanding the lists with 500 words farthest from the decision boundary.

Both the NMF and the skip-gram topic models use a cosine distance to determine topic similarity, with 300 as the word embedding dimension. Both are trained on approximately 2 million comments in high karma posts taken across a wide variety of subreddits. We use topic models in various measures of comment relevance to the discussion, but we do not use topic of the comment on its own since topic is controlled for by ranking within a thread.

5 Ranking Experiments

We present three sets of experiments on comment karma ranking, all of which show very different behavior for the different subreddits. Fig. 1 shows the relative gain in P@1 over the G&T baseline associated with using different feature groups. The importance of the different features reflect the nature of the different communities. The authority/reputation features help most for ASKSCIENCE, consistent with our k-index study. Informativeness and relevance help all subreddits except ASKMEN and WORLDNEWS. Lexical, mood and community style features are useful in some cases, but hurt others. The predicted probability of a reply was least useful, possibly because of the low-karma training bias.

Tables 3 and 4 summarize the results for the P@1 and NDCG criteria using the greedy selection procedure (which optimizes P@1) compared to a random baseline and the G&T baseline. The random baseline for P@1 is greater than 10% because of ties. The G&T baseline results show that the graph and timing features alone obtain 21-32%

subreddit	Random	G&T	All
ASKSCIENCE	15.9	21.8	**25.3**
FITNESS	19.4	22.1	**27.3**
POLITICS	18.5	24.7	**26.4**
ASKWOMEN	17.6	24.9	**28.0**
ASKMEN	18.2	**31.4**	29.1
WORLDNEWS	15.4	**24.5**	23.3
Improvement	-	42.9%	52.1%

Table 3: Test set precision of top one prediction (P@1) performance for specific subreddits.

subreddit	Random	G&T	All
ASKSCIENCE	0.53	**0.60**	**0.60**
FITNESS	0.57	0.61	**0.62**
POLITICS	0.55	0.61	**0.62**
ASKWOMEN	0.56	0.62	**0.65**
ASKMEN	0.56	**0.66**	**0.66**
WORLDNEWS	0.54	**0.61**	0.60
Improvement	-	12.5%	13.2%

Table 4: Test set ranking NDCG performance for specific subreddits.

of top karma comments depending on subreddits. Adding the textual features gives an improvement in P@1 performance over the G&T baseline for all subreddits except ASKMEN and WORLDNEWS. The trends for performance measured with NDCG are similar, but the benefit from textual features is smaller. The results in both tables show different ways of reporting performance of the same system, but the system has been optimized for P@1 in terms of feature selection. In initial exploratory experiments, this seems to have a small impact: when optimizing for NDCG in feature selection we obtain 0.61 vs. 0.60 with the P@1-optimized features.

A major challenge with identifying high karma comments (and negative karma comments) is that

subreddit	Pos/Neg	High/Mid	Ranking
ASKSCIENCE	44.5	63.7	61.3
FITNESS	74.7	43.9	57.5
POLITICS	95.5	59.1	58.0
ASKWOMEN	82.5	67.6	59.7
ASKMEN	87.0	66.2	60.6
WORLDNEWS	93.3	69.9	57.3
Average	79.6	61.7	59.1

Table 5: Accuracy of binary classifiers trained on balanced data to distinguish: positive vs. negative karma (Pos/Neg), high vs. mid-level karma (High/Mid), and ranking between any pair (Ranking).

they are so rare. Although our feature selection tunes for high rank precision, it is possible that the low-karma data dominate the learning. Alternatively, it may be that language cues are mainly useful for identifying distinguishing the negative or mid-level karma comments, and that the very high karma comments are a matter of timing. To better understand the role of language for these different types, we trained classifiers on balanced data for positive vs. negative karma and high vs. mid levels of karma. For these models, the training pairs could come from different threads, but topic is controlled for in that all topic features are relative (similarity to original post, parent, etc.). We compared the results to the binary classifier used in ranking, where all pairs are considered. In all three cases, random chance accuracy is 50%.

Table 5 shows the pairwise accuracy of these classifiers. We find that distinguishing positive from negative classes is fairly easy, with the notable exception of the more information-oriented subreddit ASKSCIENCE. Averaging across the different subreddits, the high vs. mid task is slightly easier than the general ranking task, but the variation across subreddits is substantial. The high vs. mid distinction for FITNESS falls below chance (likely overtraining), whereas it seems to be an easier task for the ASKWOMEN, ASKMEN, and WORLDNEWS.

6 Related Work

Interest in social media is rapidly growing in recent years, which includes work on predicting the popularity of posts, comments and tweets. Danescu-Niculescu-Mizil et al. (2012) investigate phrase memorability in the movie quotes. Cheng et al. (2014) explore prediction of information cascades on Facebook. Weninger et al. (2013) analyze the hierarchy of the Reddit discussions, topic shifts, and popularity of the comment, using among the others very simple language analysis. Lampos et al. (2014) study the problem of predicting a Twitter user impact score (determined by combining the numbers of user's followers, followees, and listings) using text-based and non-textual features, showing that performance improves when user participation in particular topics is included.

Most relevant to this paper are studies of the effect of language in popularity predictions. Tan et al. (2014) study how word choice affects the popularity of Twitter messages. As in our work, they control for topic, but they also control for the popularity of the message authors. On Reddit, we find that celebrity status is less important than it is on Twitter since on Reddit almost everyone is anonymous. Lakkaraju et al. (2013) study how timing and language affect the popularity of posting images on Reddit. They control for content by only making comparisons between reposts of the same image. Our focus is on studying comments within a discussion instead of standalone posts, and we analyze a vast majority of language features. Althoff et al. (2014) use deeper language analysis on Reddit to predict the success of receiving a pizza in the Random Acts of Pizza subreddit. To our knowledge, this is the first work on ranking comments in terms of community endorsement.

7 Conclusion

This paper addresses the problem of how language affects the reaction of community in Reddit comments. We collect a new dataset of six subredit discussion forums. We introduce a new task of ranking comments based on karma in Reddit discussions, which controls for topic and timing of comments. Our results show that using language features improve the comment ranking task in most of the subreddits. Informativeness and relevance are the most broadly useful feature categories; reputation matters for ASKSCIENCE, and other categories could either help or hurt depending on the community. Future work involves improving the classification algorithm by using new approaches to learning about rare events.

References

Tim Althoff, Cristian Danescu-Niculescu-Mizil, and Dan Jurafsky. 2014. How to ask for a favor: A case study on the success of altruistic requests. In *Proc. ICWSM*.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the International Conference on Machine Learning*, pages 89–96.

Meeyoung Cha, Hamed Haddadi, Fabricio Benevenuto, and P Krishna Gummadi. 2010. Measuring user influence in twitter: The million follower fallacy. *ICWSM*, 10(10-17):30.

Justin Cheng, Lada Adamic, P Alex Dow, Jon Michael Kleinberg, and Jure Leskovec. 2014. Can cascades be predicted? In *Proc. WWW*.

Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You had me at hello: How phrasing affects memorability. In *Proc. ACL*.

Jorge E Hirsch. 2005. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, 102(46):16569–16572.

Thorsten Joachims. 2002. Optimizing search engines using clickthrough data. In *Proc. SIGKDD*.

Himabindu Lakkaraju, Julian J McAuley, and Jure Leskovec. 2013. What's in a name? understanding the interplay between titles, content, and communities in social media. In *Proc. ICWSM*.

Cliff Lampe and Paul Resnick. 2004. Slash(dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 543–550.

Vasileios Lampos, Nikolaos Aletras, Daniel Preotiuc-Pietro, and Trevor Cohn. 2014. Predicting and characterizing user impact on Twitter. In *Proceedings of the Conference of the European Chapter of the ACL*, pages 405–413.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. NIPS*.

Alexander Strehl, Joydeep Ghosh, and Raymond Mooney. 2000. Impact of similarity measures on web-page clustering. In *Workshop on Artificial Intelligence for Web Search*.

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H Chi. 2010. Want to be retweeted? Large scale analytics on factors impacting retweet in Twitter network. In *Proc. SocialCom*, pages 177–184. IEEE.

Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The effect of wording on message propagation: Topic-and author-controlled natural experiments on Twitter. In *Proc. ACL*.

Tim Weninger, Xihao Avi Zhu, and Jiawei Han. 2013. An exploration of discussion threads in social news sites: A case study of the reddit community. In *Proc. ASONAM*.

Wei Xu, Xin Liu, and Yihong Gong. 2003. Document clustering based on non-negative matrix factorization. In *Proc. SIGIR*.