Analysis of Redditor Reliability

Kashev Dalmia[1], Ryan Freedman[1], and Terence Nip[1],
{dalmia3, rtfreed2, nip2}@illinois.edu
[1]University of Illinois at Urbana-Champaign

ABSTRACT
In this paper, we present a system by which to evaluate the reliability of the users of the popular social network reddit. In the past, reddit has had numerous identity issues. However, through the efforts of both reddit and the userbase itself, it is becoming a place where users come to read and discuss news. Thus, there is a growing need to evaluate the reliability of the suppliers of information on reddit. We first collect features of both reliable and unreliable users based on their contributions, and most importantly, the reaction of the community to their contributions. We then use machine learning techniques to train a regression model to assign a reliability score to an arbitrary reddit user.

1. INTRODUCTION
reddit, the self-proclaimed 'frontpage of the internet', is slowly becoming just that; many popular news sources, such as The New York Times and CNN, have begun to use reddit as a primary source of information in an effort to leverage the power of the large numbers of users which gather and try to provide and cross-validate information others have supplied.

In most cases, the information users share on reddit is rather inoccuous. For example, a subset of reddit users (or *redditors*) gather information from popular news sites and aggregate them on individual, topic-based forums (called *subreddits*) like `r/news` or `r/worldnews`, subreddits for current events within the United States and the world, respectively, in an effort to share and discuss events going on in the world around them.

However, reddit's track record is not necessarily completely spotless when it comes to dispersing correct, cross-validated information. For example, during the Boston bombings of the Boston Marathon in April of 2013, reddit falsely accused individuals of having been part of the bombings simply because of circumstantial evidence. Many news sites, including CNN, picked up the story and publicized it. However, it turned out that the individuals reddit picked out were completely innocent and were only identified as being potentially involved because of circumstantial evidence (such as having been missing for long periods of time).[**?**].

reddit is also noteworthy for allowing users to *upvote* or *downvote* user-submitted content, indicating either agreement/interest or disagreement/disinterest, respectively, with the material at hand. As a side effect (and as part of the site's algorithm to keep fresh content up longer), users ultimately promote material that they find to be interesting to them, and hide material that they find to be out of scope, irrelevant, or simply not interesting with respect to the subreddit they're in.[**?**] Additionally, every upvote and downvote is connected with a measurement of the community's net happiness/satisfaction with the content submitted. A user gains *karma* for every upvote they receive, and loses karma for every downvote they get.

2. RELATED WORK
TODO

Something something only a single subreddit [**?**]

3. DESIGN
In this section, we show the design of our system, starting from data gathering/aggregation, through to the final modeling/analysis of the data obtained. Additionally, we discuss some of the challenges, limitations, and design decisions that went into making it. Finally, we discuss some of the finer details of the system's implementation.

3.1 Gathering reddit Usernames
The first limitation of the reddit API is that usernames are an 'open secret'. If one has the username of an account, public information about that account can be retrieved. However, there there is no way to directly get usernames. Instead, one has to gather usernames by scraping posts of subreddits and getting the author's username of every post and comment. In doing so, we were able to collect over 150K usernames. Of the usernames we collected, we randomly selected around 3K to fully gather data on and run our regression model on.

One issue with this approach is that this makes it impossible to identify non- participants. If a user neither never comments on a post, nor posts a post themselves, there is no

way to know that that user exists. This is unfortunately an insurmountable limitation given the existing API. Instead, we chose only to find good and bad users who have demonstrated activity on reddit to train our classifier, and ignore non-participants.

3.2 reddit API Limits

reddit has an API limit of 30 requests per minute when one hasn't authorized their script, and 60 requests per minute when authorized. Given our timeframe, we eschewed the authorization method and decided to go with the less time-efficient method of not authorizing our script. In gathering data, however, we discovered this limit is not strictly enforced. In order to be good citizens and gather the data we wanted whilst simultaneously working not get our access revoked, we knew we had to design around these constraints. As such, to both speed up our ability to access user data, dynamically change the features that we used (see Section 3.4), and reduce load on reddit's servers, we used a MongoDB instance to cache the data we were gathering. Not were we able to store raw data from reddit API calls, we were also able to cache results from more computationally intensive features that we were generating.

3.3 Establishing Ground Truth

Establishing ground truth was done in two stages, the first of which was to find reliable users. Finding these users was relatively trivial as the site rewards positive behavior through karma, leading to increased visibility. From there, users could be filtered manually by their level of contribution. Moderators from various communities were also included in this set for their work in the reddit community. These users were used in our training set and were given a reliability score of $s_r = 1$.

The second stage was to find users who were unreliable or detrimental to the community. Eventually, we discovered subreddits dedicated to contributing content to reddit that was either not useful or not constructive in promoting reddit as a reliable site, and numerous posts that suggested their authors were accounts used to abuse the community. These were also used as part of our training set, and were given a reliability score of $s_r = -1$.

3.4 Feature Selection

Before having a trained regression model, there was no way to tell what characteristics of reddit users would be important. However, one can't get this regression model without these features. So, we decided to pick a wide variety of features generated from the raw data, and figured out what was important later by looking at our results. Here, we present features which we suspected would be important. A discussion of what features were actually important is in Section 4.

3.4.1 reddit Karma and Derivatives

Perhaps the most important feature of the reddit platform, and one of the reasons why reddit is an incredibly relevant resource in today's changing news landscape is its built in voting system. We hypothesized that users whose posts and comments are well received would have positive correlation with that users reliability. However, not all content on reddit is news-worthy. There are numerous subreddits in which the content is less than news-worthy, containing things like cat pictures and pornography, to name a few topics.. Thus, we established several sub-criteria in order to differentiate between karma earned by posting reliable news and karma earned by posting content that merely engaged the community's softer side, such as by posting picture of cats.

We broke down a user's net karma into karma earned in the top 100, 50, 25, and 10 subreddits as reported by redditlist [**?**]. This way, users who participate mostly in small, niche subreddits would be somewhat penalized for not participating in more mainstream, 'useful' subreddits as part of a measure of "impactfulness". For the purposes of this project, we picked a small list of 'trusted' subreddits, and picked good, reliable users from those communities as training examples, and used a users' net karma from the top n subreddits as features in our classifier.

3.4.2 reddit Gold and Gilded Content

Another interesting feature of reddit which differentiates it from other social networks is reddit gold. reddit gold is a premium membership which users may pay for. It comes with many perks, such as ad-free experience, a members-only subreddit, and deals from partner companies, amongst a number of other benefits [**?**]. What makes reddit's gold program so unique, however, is that users are encouraged not only to buy it for themselves, but also to give reddit gold membership as a gift to other users who produce content that they enjoy.

Content that motivates users to purchase reddit gold for others is marked as being 'gilded', and is marked as being such. This action can be done numerous times for any particular piece of content submitted/posted to reddit, to the point where a post can be gilded upwards of five or six times within the span of a few hours. Given the impactfulness and meaning behind a user's being given reddit gold, we suspected that this action would be extremely important in determining reliable content in that content which was reliable or incredibly enjoyable/not detrimental to the well-being of the site would be worth a user spending real money to reward.

3.4.3 Natural Language Processing

Finally, we suspected that the actual writing patterns of reliable users would separate them from unreliable ones. Specifically, we looked at three metrics that we suspected might correlate with a user's reliability.

1. **Unique Word Percentage**: We suspected that users who were reliable would use more unique words than those who were not.
2. **Readability**: We analyzed the Flesch–Kincaid readability [**?**] score of the comments users posted, which can be calculated using the formula in Equation (1).

$$206.835 - 1.015\left(\frac{\#\text{ words}}{\#\text{ sentences}}\right) - 84.6\left(\frac{\#\text{ syllables}}{\#\text{ words}}\right) \quad (1)$$

3. **Swearing Rate**: We took a list of 'bad' words which Google compiled [**?**], and calculated the percentage of those words that users used in comments.

Table 1: The features used to create the regression model

Feature Description	Importance %
Link Karma	57.92
Comment Karma	18.75
Total Number of Comments	11.97
Average Karma per Post	3.07
Total Number of Posts	2.62
% of Comment Karma - Top 100 Subreddits	0.85
% of Post Karma - Top 100 Subreddits	0.79
Has Verified Email	0.65
Unique Words / Total Number Words	0.55
Average Comment Karma per Comment	0.37
% of Comment Karma - Top 50 Subreddits	0.34
% of Comment Karma - Top 25 Subreddits	0.33
Flesch–Kincaid Readability of Comments	0.32
% of Comment Karma - Top 10 Subreddits	0.32
% of Comment Karma - Trusted Subreddits	0.28
% of Swear Words Used in Comments	0.24
Time Account Created	0.23
Is Reddit Gold	0.15
% of Post Karma - Trusted Subreddits	0.09
% of Post Karma - Top 50 Subreddits	0.07
% of Post Karma - Top 25 Subreddits	0.05
% of Post Karma - Top 10 Subreddits	0.02
Number of Gilded Posts	0.00
% of Comments Gilded	0.00
Number of Gilded Comments	0.00
% of Posts Gilded	0.00

3.5 Picking a Regression Model

In picking a regression model, two things needed to be considered, namely: (1) the robustness of the model, and (2) the ease with which we could garner insights from the model generated.

Ultimately, random forests [**?**] was chosen based on the notion that the results were more interpretable than other popular modeling algorithms such as neural networks, and would provide greater reliability than just a simple decision tree.

4. EVALUATION AND RESULTS

From our trained random forest regression model, we get a glimpse of both what useful, contributing redditors look like, and what bad, non-contributing redditors look like.

Our test data set was about 3,200 redditors, all of their post data (around 1GB), and all of their comment data (around 3GB). From this test data set, we used our our regressive model to get a reliability score between $-1 \leq s_r \leq 1$. Then, we re-correlate this score with input features to intuitively see what features are important or not, and what features indicated useful and not-useful redditors.

4.1 Distribution of Users

In general, we around 24% of users to have a negative score, and the remaining 76% to have a positive score. This falls in line with a YouGov study that revealed about a quarter of Americans admit to acting as 'trolls', or other sorts of argumentative, non-constructive behavior online [**?**]. A more detailed distribution of scores is in Figure 1. Most of the scores assigned were closer to 1, with a relatively uniform distribution of other scores.

4.2 Feature Efficacy

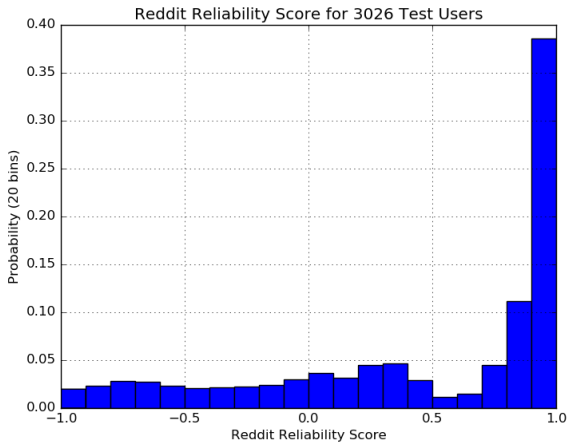

Figure 1: The distribution of the reliability score s_r of the sampled redditors, binned into twenty bins.

Though we gathered several different features, not all of them ended up being 'useful' in terms of creating a reliability score. The full table of the features that we used, as well as the percentage importance our model placed on them, appears in Table 1. It's worth noting that some featured are highly correlated, like karma from top 100 subreddits and total karma, so even though these features might be correlated with reliability, the model did not place importance on them.

4.2.1 What Did Work

The most important features unsurprisingly turned out to be the amount of post and comment karma a user had. As this is the built in voting system of reddit, we expected to see this. Plots of how post and comment karma vary with reliability appear in Figures 2 and 3, respectively. Both of these plots suggest that the voting system works to root out unreliable users; users with $s_r < 1$ don't get much karma. Of users with $s_r \geq 1$, the distribution is bimodal. We hypothesize that our model manages to differentiate between users who are casual, and post well received but non- useful content, and those who post useful, reliable content which is also well received by reddit on the whole. Similarly, the voting system seems to discourage participation from unreliable users. Figures 4 and 5 suggest that users who are unreliable do not end up posting much, but those who are more reliable are more likely to post far more content.

A more surprising feature was the importance of a user verifying their email. A split histogram of s_r of users, split on if they have verified their email or not (Figure 6), suggests that though the model only placed 0.65% importance on this feature, having an unverified email means a user is more likely to not be reliable. This stands to reason; people who act badly on the Internet often do so only when their anonymity is guaranteed [**?**], as it is on reddit, so they'd be less likely to link a non-anonymous account like email to their reddit account.

4.2.2 What Didn't Work

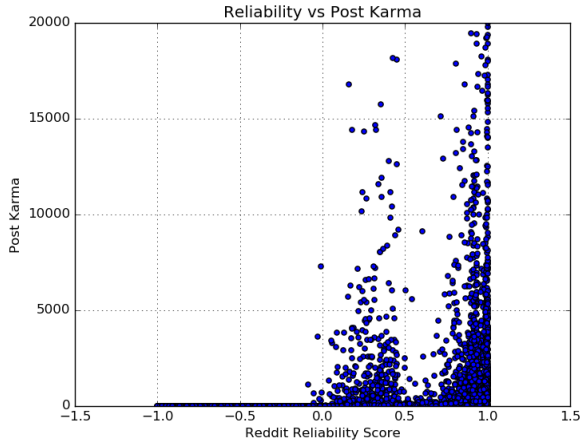

Figure 2: The reliability score s_r plotted against the post Karma the redditor has.

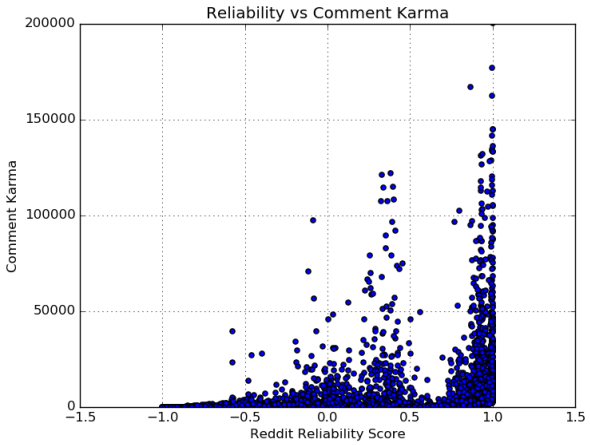

Figure 3: The reliability score s_r plotted against the comment Karma the redditor has.

Though we had hoped that features of the words that redditors used would indicate how reliable they were, we didn't find this to be the case. An examination of Figure 7 suggest that there is not relationship between reliability and Flesch–Kincaid readability of comments. Our model placed similarly low importance on the frequency of using bad words and readability of comments. This suggests to us that many kinds of speech patterns are accepted on reddit, and has little bearing on reliability. Our model also placed low importance on the amount of karma earned in subsets of subreddits.

We also thought that gilding would be an important feature, but it turned out not to be, as seen in Figure 8. Not enough users get gilded on reddit, which makes sense, as putting real money into the service is a much higher friction action than simply using it for free.

5. CONCLUSION

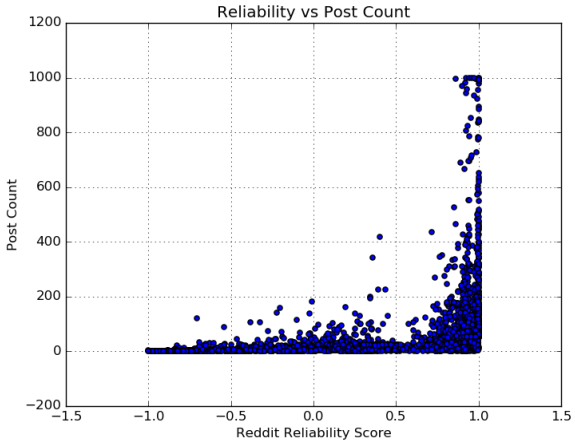

Figure 4: The reliability score s_r plotted against the number of posts the redditor has made.

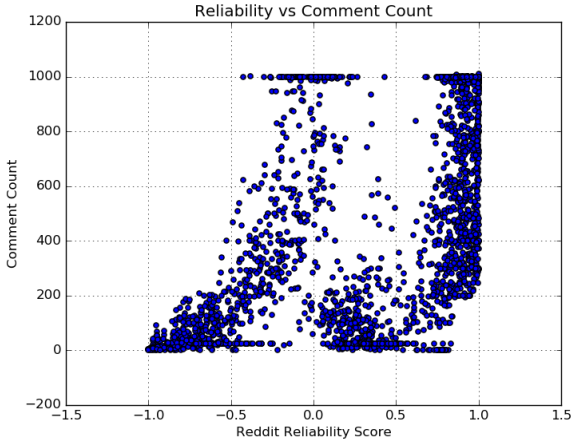

Figure 5: The reliability score s_r plotted against the number of comments the redditor has made.

In this section, we discuss some weaknesses in our project and discuss future improvements that may be made.

5.1 Weaknesses

Given the relatively unexplored landscape and the restrictions of the reddit API, there exist many weaknesses in our analysis of reddit.

First and foremost, a larger pool of users need to be analysed in order to get a better, more complete picture of what reddit is truly like. 3,200 users is an incredibly small subset of the number of users reddit has, and does not paint a full picture of how reliable the site is as a whole. A less biased pool of users would also be beneficial in trying to come up with a more robust model; in training, we artificially took our data and converted it from its natural bias of roughly 90% unreliable users into a 50% bias in order to allow our model to actually learn useful features. Having access to a pool of usernames that were deemed as being active and reliable would have helped immensely in decreasing the in-

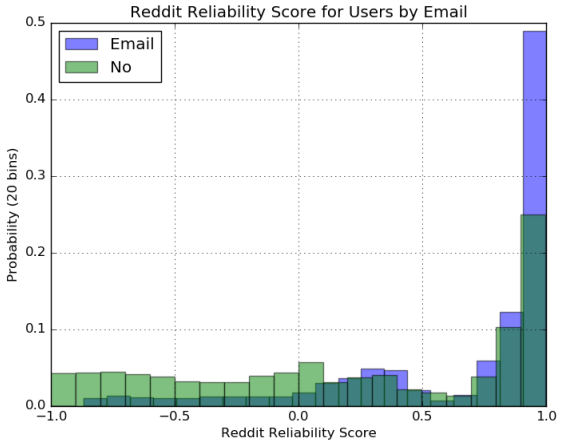

Figure 6: The distribution of the reliability score s_r of the sampled redditors, binned into twenty bins, separated by if they have a verified email address or not.

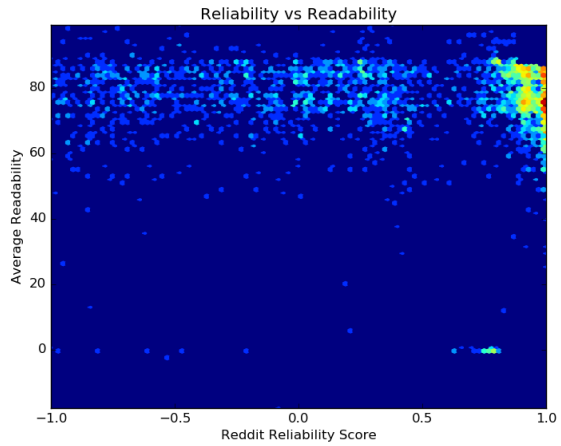

Figure 7: The reliability score s_r plotted against the Flesch–Kincaid readability of their comments.

herent bias and in turn, would have accelerated the data gathering/analysis process.

Secondly, more signals could be included in our classifier. Due to time constraints, there were features we would have liked to add into our classifier that we were unable to include, such as a temporal analysis of a user's reliability factoring into how reliable they may be at any given time. It is also worth noting that the data we have is solely limited to what reddit provides openly, and the model we have could be made more robust if greater access to data (such as subreddit subscriptions, etc) were provided.

Third, the training dataset was created rather subjectively instead of being created/scored by numerous individuals blindly. Given sufficient resources, the training dataset would have ideally been created via the use of Amazon Mechanical Turk, allowing for the crowdsourcing (and inherent cross-validation)

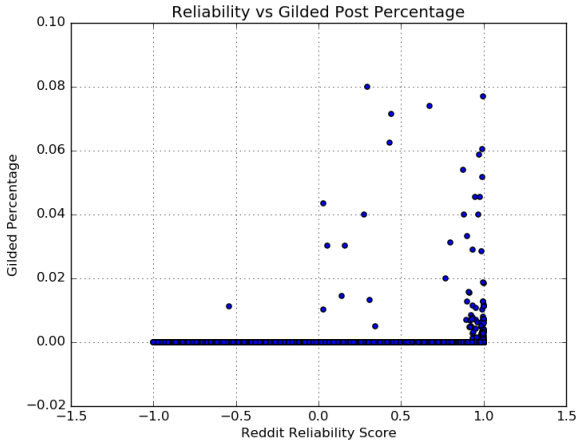

Figure 8: The reliability score s_r plotted against the percentage of gilded posts.

of the training dataset. It also would allow us to have a test set to work off of, which we unfortunately did not have the resources to generate given the difficulty involved in generating a set of reliable users.

5.1.1 Conclusion

Given the set of users we have obtained in conjunction with the variability that exists among the various subsets of reddit users there exists weaknesses in our analysis. However, given the current state of research of reddit and the increasing reliance of individuals on reddit as a source of reliable information our contribution noteworthy for the purpose of exploration with respect to reddit as an up-and-coming source of information.

5.2 Future Work

In addition to addressing the weaknesses presented previously, there exists quite a bit of future work that can be tackled.

To begin, the problem of establishing ground truth is a rather difficult one, and is one that we decided not to take on given the time constraints placed upon us. There is no objective way to establish ground truth, and the level of subjectivity required in establishing it would require not only greater temporal resources, but also financial resources that we do not possess. Given that users can be constructive and reliable in different ways, we chose perhaps the most obvious route of identifying reliability, but there are more nuanced senses of reliability that would help identify users as being reliable despite not being "mainstream".

The problem of voting on reddit is also an open issue; redditors not only vote because the content is important and reliable, but also because the content is personally interesting to them. This means that it may be the case that an up and coming, incredibly important news article submitted to `r/news` may have the same amount of karma and be gilded the same way as a cute cat picture on `r/aww`. Identifying the differences between these different types of karma and finding a proper measurement/way to include them in our

measurement of reliability is rather important, and should be addressed in the future.

Finally, building a more robust system to do a lot of the work that we did manually, and then extract likely true events from reddit, would be extremely useful. Work like this has been done for other real-time social networks like Twitter and Instagram in Apollo [**?**].

6. ACKNOWLEDGMENTS

The authors would like to thank Professor Tarek F. Abdelzaher, of the University of Illinois at Urbana-Champaign, for his support in this project.