# Predicting Reddit /r/relationships Post Popularity

Ho-Wei Kang, *UCSD, hok022@ucsd.edu*

***Abstract*—Reddit is an online forum where users can post and upvote (or downvote) other users' posts. Reddit is divided into interest groups called subreddits, wherein users discuss a particular topic. This analysis deals with submissions from the subreddit /r/relationships wherein users vent and seek advice on their relationship problems. The purpose of this analysis is to try to identify any trends in relationship problems (e.g. infidelity, breakups) and attempt to predict how many comments, upvotes, and downvotes a particular post might receive using linear regression and multi-class support vector machine models.**

## I. Introduction

Data for this analysis was gathered using the Reddit API via PRAW, a Python client library for the Reddit API. The aggregated dataset includes all submissions from January 1st, 2013 to November 25th, 2015 to the /r/relationships subreddit, and includes the following fields for each submission:

**author**
the Reddit username of the author of the submission

**created**
local epoch time the submission was created

**created_utc**
utc epoch time the submission was created

**title**
the title of the submission

**selftext**
the content of the submission

**num_comments**
the number of comments the submission received

**ups**
the number of upvotes the submission received

**downs**
the number of downvotes the submission received

**score**
the score the submission received ('ups' - 'downs')

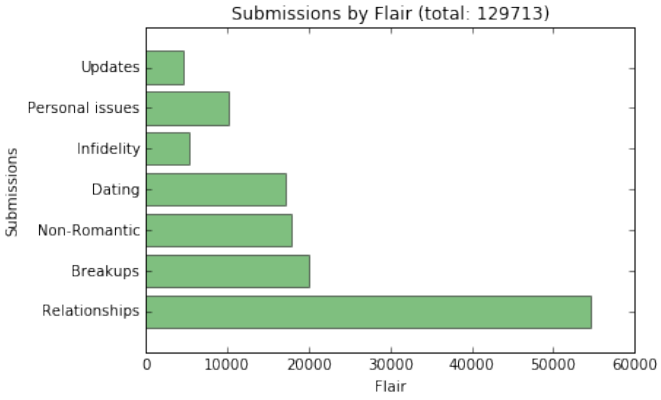
**over_18**
whether or not the post is marked for audiences over the age of 18

**flair**
the user-classified category of relationship problem which can be one of the following:

- *Relationships* issues in ongoing romantic relationships
- *Infidelity* issues of suspected or actual cheating
- *Breakups* break-ups and any issues with your exes
- *Dating* new people in your life, crushes, unclear relationships, or things shorter than 1 month
- *Updates* update to a previous /r/relationship submission
- *Non-Romantic* issues with family, friends, or coworkers
- *Personal Issues* personal problems
- *None* unlabeled submissions

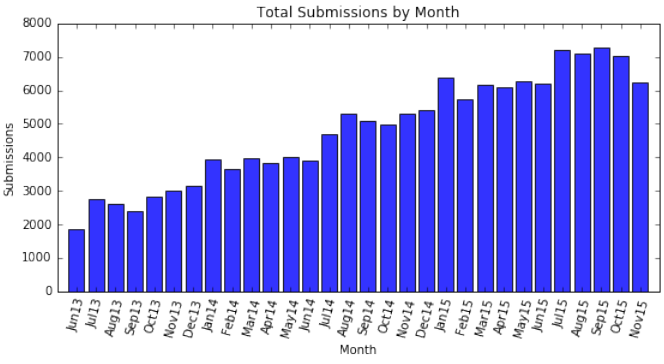
### A. Cleaning the Data

Flairs label the issues described in submissions, and it probably isn't too much of a stretch to say that the type of issue will be important in prediction tasks. Submissions without flairs also only comprise 15% of the data, so for the purposes of this analysis, submissions without flairs will not be considered. This necessitated discarding 23,821 submissions, including all submissions from before June 2013 (when flairs were introduced). Once the unlabeled submissions were removed, 129,713 submissions were left.




### B. Submissions Over Time

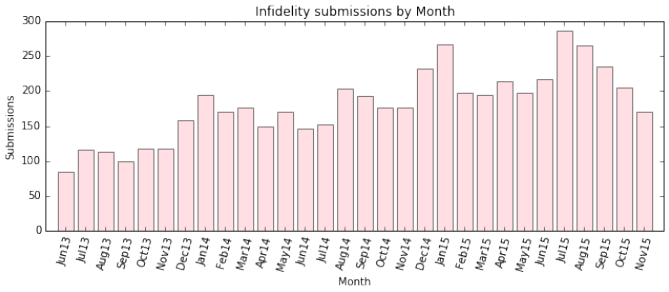
According to Alexa, Reddit is ranked the #33 most popular website in the world and #9 most popular in the US. Inside of Reddit itself, redditmetrics.com calculates /r/relationships as one of the top 70 fastest growing subreddits. By plotting the number of submissions a month /r/relationships has received over the last 2 years, we can see that this is largely true and that the number of submissions /r/relationships is receiving is monotonically increasing.



### C. Submission Flairs Over Time

Plotting monthly submission flairs over time did not seem to reveal anything significant. Most plots simply corroborated that yes, number of submissions on average have gone up over time as more and more users start using Reddit. The only plot that really stood out was the plot of monthly infidelity over time, which shows that in 2014 and 2015, January was the prime time for infidelity, possibly because of New Years. It would be interesting to see if infidelity submissions peak again January 2016.

Infidelity submissions also peak July 2015, but begin monotonically decreasing right afterwards. This could possibly be due to the long distance relationships of students failing over summer vacation, or it could possibly be just random. It would however be interesting to see if there are more student-aged posters during this time period.




### D. Submission Count and Scores Over Time

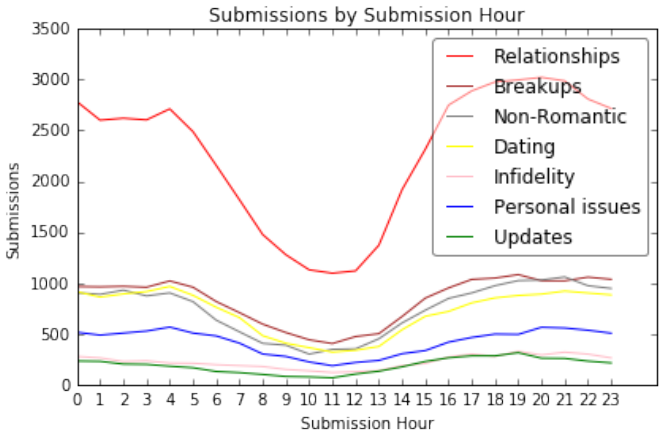



Fig. 1. It seems that most submissions to /r/relationships happen after hours from 17:00-4:00 in users' respective timezones. 12:00 noon seems to be the least popular time to post, since around that time is when most users are likely eating lunch, at school, or at work.

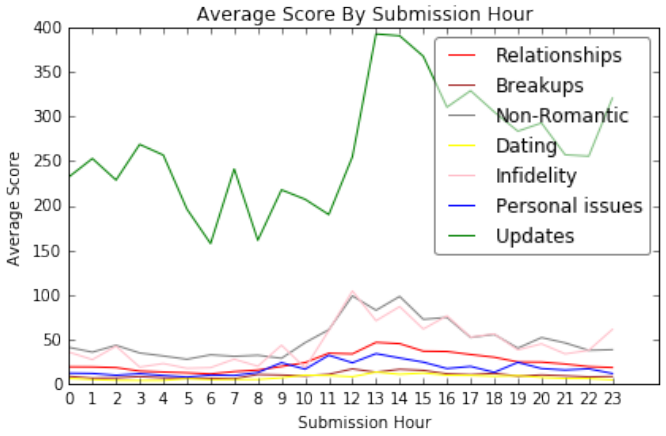



Fig. 2. Notice that in general, submissions with the 'Updates' flair do considerably better than all other types of submissions. Among the other types of submissions, Infidelity and Non-Romantic perform the best. Average submission scores seem to peak at around hours 13:00-14:00 which is when the fewest number of submissions are made.

### E. Age and Gender of Subjects

The age and gender of the subjects involved in a post likely affect the post's popularity. Posts involving subjects with large age gaps for instance might attract more upvotes on average. Or perhaps infidelity posts where the woman is being cheated on will receive more or fewer comments on average than posts where the man is being cheated on.

Age and gender proved to be tedious to extract since people had different ways of expressing age and gender. Most common was 'I [25M] have problems with my girlfriend [24F]', but some opt to put the gender before the age, the age/gender in parenthesis rather than brackets, or not put any space between the subject name and age gender combination at all. But using regular expressions and taking the word that comes before the age gender combination as the subject name, age and gender information was able to be extracted from 93.5% of the submissions.

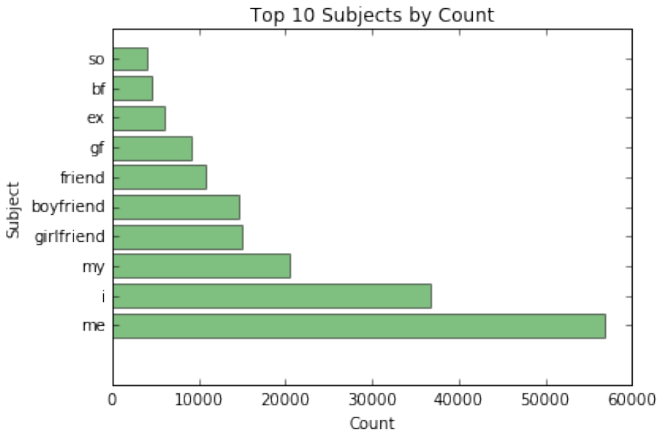



Fig. 3. Notice that 'i', 'me', and 'my' all refer to the original poster, or 'op' for short. To make features consistent, all subjects extracted that referred to the op were changed to 'op'.

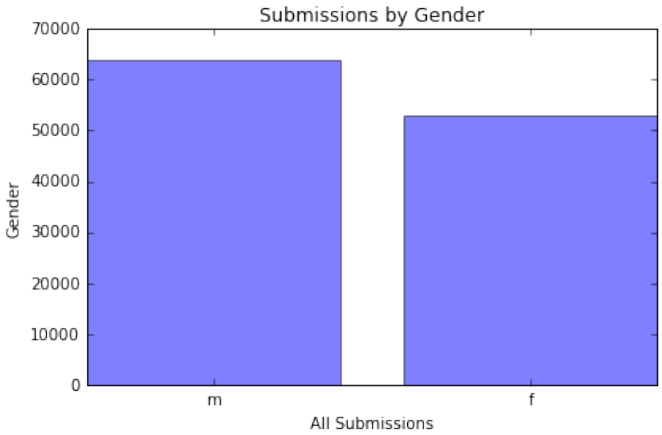



Fig. 4. The ratio of male to female posters in /r/relationships is roughly 1.2.

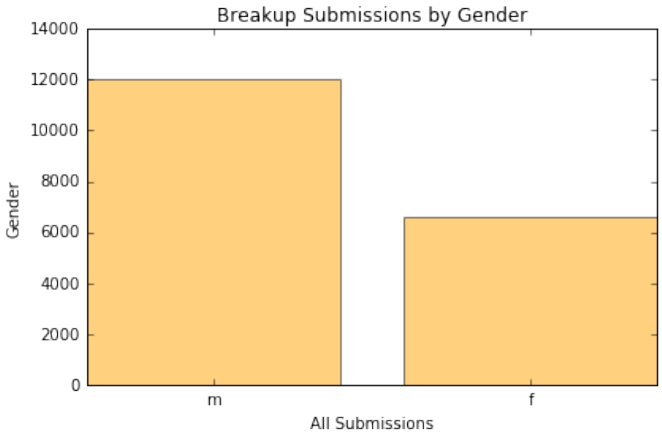



Fig. 5. Interestingly, the ratio of male to female posters of submissions with the 'Breakup' flair was 2.0.

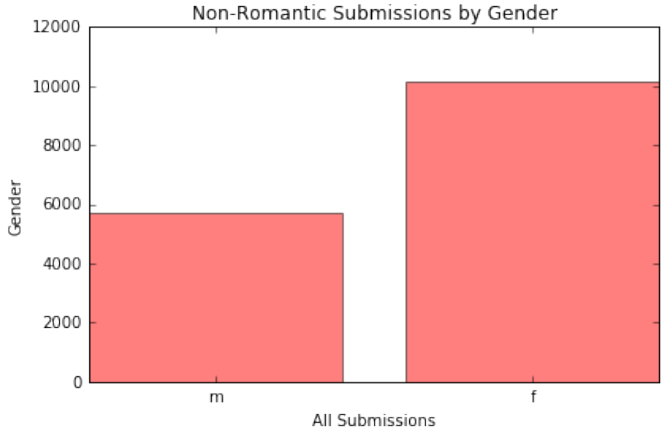



Fig. 6. Interestingly, the ratio of male to female posters of submissions with the 'Non-Romatnic' flair was 0.5.

## II. Predicting Popularity

The popularity of a Reddit submission is measured by the score submission receives (upvotes - downvotes). Predicting how popular a Reddit submission is therefore means predicting the submission's score.

### A. Using a Subset of the Data

Reddit has grown considerably over the years. As a reflection of this, the average score a submission on /r/relationships received in 2014 was 22, while in 2015 it has increased to 45. To make predictions more accurate, only submissions after and on December 1st, 2014 will be considered (61490 submissions total).

### B. Assessing Model Performance

Using root mean-squared error (RMSE) to measure model performance seems like a reasonable idea. Score predictions that fall further from their actual values will be penalized more than predictions that fall closer, which makes sense intuitively.

### C. Baseline Model

As a baseline predictor for comparison, the naive predictor that simply predicts the mean score for all submissions will be used. The purpose of this baseline predictor is to be used as comparison against any potentially better predictors. If the potentially better predictor isn't better than this naive one, then it isn't very good at all.

The mean score for all submissions in our data set is 44.57, which resulting in a root mean-squared error of 183.22.

### D. Linear Regression Model

For predicting an exact score, a linear regression model is a good fit as a linear regression model will predict an exact value. For the linear regression, the following features were considered:

**created_utc**
: The utc time the submission was used as-is. Useful and possibly linear as Reddit and therefore the average score will grow over time.

**created_utc % 86400**
: Time of day the submission was made, in seconds after midnight (86400 is the number of seconds in a day). Score is likely a degree 2 polynomial function of this feature as submissions at some point of the day are likely to have significantly more points than submissions made at the local mins of the day where fewer Reddit users are online and voting.

**length of title**
: Score is possibly a linear function of the length of the title. The intuition is that submissions with short titles are likely worse submissions, and that posters who spend the time to write a more quality title will likely have a better overall submission.

**title bag-of-words**
: The top 200 most occurring words in submission titles were used as features. To prepare, the entire string was converted to lower case, punctuation was removed, and then the entire string was split by whitespace characters. Each splitted string (word) was then stemmed, which is the process of reducing inflected words to their word stem or root. The purpose of this process is count the two distinct words 'finish' and 'finished' as one word: 'finish'.

**length selftext**
: Score is also possibly a linear function of the length of the actual submission, as submissions with more words are probably more controversial and dramatic, attracting more upvotes.

**selftext bag-of-words**
: The top 900 most occurring words in submission selftexts were used as features. The same process used in extracting the most common words from the title was also used in extracting the most common words from the selftext.

**over_18**
: Whether or not the post was marked as over18 was used as a feature.

**flair**
: The flair was used as well, with the flair weights divided up per category. This is because the weight of each flair is independent from one another (i.e. the weight that the flair Infidelity receives should not have an effect on the flair Non-Romantic receives).
$score \simeq ...\theta_1 \times is\_Breakups + \theta_2 \times is\_Infidelity...$

**op's gender**
: The op's gender was used, with weights divided up per gender. This is because the weight of op being male and female are indepent from one another.
$score \simeq ...\theta_1 \times is\_Male + \theta_2 \times is\_Female...$

**op's age**
: The OP's age was also used as a feature. 0 if no age was parsed.

## III. Related Literature

Jordan Segall and Alex Zamoshchin worked on a very similar prediction task with a very similar dataset. In addition to regression, their work also attempted to use naive bayes as well as a multi-class SVM. While this analysis largely tried to predict the exact score a Reddit post receives, Segall and Zamoshchin clustered the scores into score brackets and instead attempt predict which bracket a particular post would fall into. This seems like a smart idea, as often times posts will either explode and get hugely popular or die off in its infant stages in which case something such as a linear regression which tries to fit this step-like behavior to a line may not work very well. Their work also measured their results in RMSE, and using linear regression, they were not able to lower RMSE by much from their baseline. This was true for their SVM models as well, which shows that predicting Reddit post popularity is probably a non-trivial problem.

Alex Bragdon et. al did a very similar analysis as Segall, but with a much smaller dataset from Reddit (only about 38,000 posts). Unlike Segall's analysis or this analysis, the work done in Bragdon's paper did not take into account content of the

posts and only considered metadata. Like Segall, Bragdon clustered the scores of submissions and trained a multi-class SVM, but like Segall were unable to produce significant improvements over the baseline.

Terentiv and Tempest also realized that because of the strange distribution of scores (with 54% of their submissions with a score lower than 10), that a linear regression would be too skewed. Instead, they opted to use clustering algorithms to make score brackets, much like the work of the previous two cited works. What was unique with Terentiv et. al's work is that they performed semantic analysis on the first 10 comments a submission received and used those as features to attempt to predict which category a submission falls under. Unfortunately with no baseline model it was hard to tell how much better their model was from a naive predictor. Their training set was also only a humble 2,000 models.

## IV. Results

Without the bag-of-word features or the op gender or age features but using every other feature described in the previous section, a root mean-squared error of 169 on the test set was achieved, 14 points lower than the baseline of 183.

With the bag-of-word features added (but still without the op gender and age), which increased computation time by several orders of magnitude and took around 20 minutes to completely process, the root mean-squared error was further lowered to 157.

With the op gender and age but without the bag-of-word features, the root mean-squared error was only lowered to 167 from 169.

With all the features combined, the root mean-squared error stayed at 157.

This showed that the core features were important and that the bag-of-word features were very effective. It also showed however that the op gender and age features were practically useless, contrary to initial predictions in the beginning of this analysis.

The linear regression also showed that the flair was very important, with the weights of 'Update' and 'Breakups' submission flairs being very positive. The 'over_18' flag seemed to be a very high weight feature as well, with posts marked as being over 18 receiving on average a higher score than posts not marked as over 18.

Overall, this linear regression model performed 14% better than the baseline predictor. Based on related works, predicting Reddit post popularity is pretty hard. Perhaps a better solution would be to classify the submissions into a score bracket using a multi-class SVM, and then run a linear regression on the submissions within the brackets such that the linear regression is not so skewed from the uneven distribution of submission scores.

## References

[1] Jordan Segall and Alex Zamoshchin. Predicting Reddit Post Popularity (http://cs229.stanford.edu/proj2012/ZamoshchinSegall-PredictingRedditPostPopularity.pdf)

[2] Alex Bragdon et. al. Predicting Reddit Post Popularity (http://users.wpi.edu/ hsahay/assets/PredictingRedditPostPopularity.pdf)

[3] Andrei Terentiv and Alanna Tempest. Predicting Reddit Post Popularity via Initial Commentary