

Reddit Analysis

Adam McDaniel, Logan Wrinkle, Colby Smith, and Darius White

I. INTRODUCTION

Reddit is a social media website used for all types of communication. People use the platform in various ways, including: discussing politics, conferring over engineering practices and development, showcasing chess puzzles, sharing images and videos, holding public interviews, and direct-messaging other users. What's more, the simple structure of Reddit lends itself to web scraping and data collection: the website is divided into individual communities surrounding a single subject, and each community contains subscriber posted submissions which other users may reply to. With this structure, the social roles of users and attitudes of users towards others is more easily understood: negative post titles suggest a negative post subject, and negative comments on a post with a positive title indicate that the post itself off-topic or controversial. Aggregating posts, comments, and replies and computing their sentiments can easily illustrate the attitudes and culture of a given community.

II. METHODS

A. Implementation

A "collect" program was written to gather posts for a list of subreddits using the Reddit API. A the program downloads a hundred or so posts per subreddit per run, and all posts are taken from the "Hot" posts lists: posts that are the most upvoted, recent posts. Each subreddit's data is converted into a tree structure of posts and replies, and stored in a JSON file along with metadata about each submission.

Next, the "analyze" program reads this tree structure in from the JSON file, and creates a new JSON file where each post and comment is tagged with a predicted subject, a predicted attitude, and the confidence of the predictions from zero to one.

Subjects		Attitudes
Politics	Religion	Inquisitive
Science	Food	Praise
Animals	Sports	Condemnation
Music	Movies	Agreement
Joke	Technology	Complaint
Discussion	Personal	Mocking
Other (<i>unable to predict</i>)		Disagreement
		Annoyed
		Neutral (<i>unable to predict</i>)

Fig. 1. List of subjects and attitudes classified by the analyzer.

The analysis is performed by using zero-shot classification on each submission with the Hugging Face BART model. The model is supplied a given string to classify, and a list

of classes with which to classify the string's content. The model returns a predicted classification, and its confidence.

Finally, the "graph" program reads in the analysis JSON file and outputs several images of graphs using metrics extracted from the analyses. The graphing program does not include data with "Other" subject labels or "Neutral" attitude labels. Comments that cannot be classified with confidence, usually short comments with little to no communicative value which are very common on Reddit, cloud the data collected from more informative and communicative submissions.

B. Extracting metrics

Each of the attitudes classified by the analyzer were given a positivity and a negativity score from zero to one. To determine the "positivity" or "negativity" of each subreddit, the positivity and negativity scores of each individual submission are summed, and divided by the total number of submissions for that community. Similarly, to calculate agreeability or disagreeability of a subreddit, the proportion of submissions expressing agreement or disagreement is divided by the total number of posts for that community.

Positivity and negativity scores of attitudes

Attitude	Positivity Score	Negativity Score
Inquisitive	0.5	0.5
Praise	1.0	0.0
Agreement	0.8	0.2
Disagreement	0.25	0.75
Complaint	0.0	1.0
Mocking	0.0	1.0
Annoyed	0.0	1.0
Condemnation	0.0	1.0

Fig. 2. Assigned weights of positivity and negativity according to attitude.

Another metric extracted is the "divisiveness" metric. Divisiveness is measured by calculating the split of the agreement and disagreement of a submissions replies (polarization), and by counting the number of replies taking a position (assertiveness). This metric is calculated for an entire subreddit by calculating the divisiveness of each leaf in the tree of replies of each post, summing the divisiveness for each leaf in the reply tree, and then averaged across multiple posts.

III. RESULTS

A. Attitudes regarding subjects

After analyzing the data, several trends were noticed; by illustrating the attitudes of submissions according to their subjects, several predictable phenomena became immediately

apparent. For example, political submissions had the largest distribution of disagreement, condemnation, and complaints when compared to all other subjects.

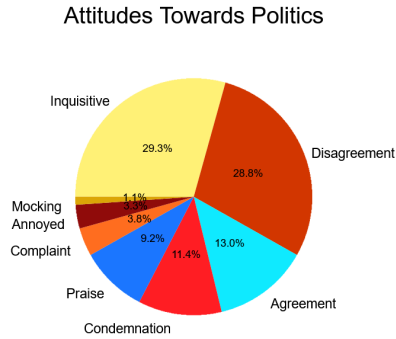


Fig. 3. Distribution of attitude labels for submissions labeled as political.

Discussion submissions, which are inherently curious in nature, were labeled as inquisitive 62.4% of the time. Submissions labeled as personal conversation (about oneself or another user), had the largest percentage of posts expressing annoyance at 12.5%. Music was the subject with the largest distribution of submissions labeled as praise at 49.0% of posts: music is a widely appreciated, non-controversial subject, and artists who post to music subreddits are likely to get praise from other musicians for their work. Jokes, as expected, were labeled as mocking far more often than any other subject: 25.1% of the time.

One unexpected trend is the disparity found between subjects labeled as science related versus technology related. It was expected that the two subjects would have similar distributions of attitudes, but this is not the case. Technology related posts were 20.7% likely to be labeled as expressing disagreement: *almost four times the disagreement of science related posts.*

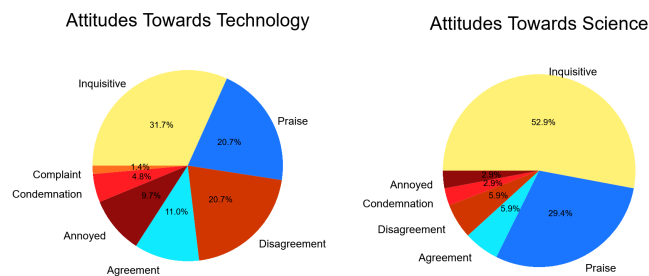


Fig. 4. Technology related submissions versus science related submissions.

Even more strangely, science related posts were 20% more inquisitive, and expressing 10% more praise. Posts about technology are also three times more likely to express annoyance. Science related posts have very few submissions expressing negative attitudes, while this is not true for technology related posts.

One explanation for this result is that posts about malfunctioning devices, which might express annoyance with the failure of the device, could be more likely to be labeled as technology related.

B. Positivity and negativity

Other trends, both expected and unexpected, appeared in the positivity metrics for each community as well. The most positive communities, as expected, were r/aww and r/mademesmile. Curiously, r/creepy was also one of the most positive subreddits analyzed, trailing just behind r/mademesmile. This was unexpected because it was assumed that the creepy content of the subreddit would classify the community as more negative than positive. Upon further inspection of the community after measuring this result, however, it was noticed that all of replies to posts were always immensely positive and encouraging: posters create creepy content, and other users praise their work.

The most negative communities were r/bad_cop_no_donut (a subreddit dedicated to discussing police brutality), r/rage, r/inceltear, and several political subreddits such as r/politics, r/democrats, r/republican, and r/conspiracy. Interestingly, r/democrats was closer to other non-controversial subreddits in positivity than r/republican. One unexpected highly negative subreddit was r/explainlikeimfive: this is attributed to disagreement in the comments about the possible explanations for the users question.

C. Divisiveness

The most divisive communities were r/programming, r/wtf, and r/anger. These make sense, as each of these communities often express recursive disagreement in the comments: r/programming expresses disagreement about programming solutions, r/wtf expresses disagreement about explanations of seemingly inexplicable content, and r/anger expresses disagreement about what each poster should or should not do to manage their anger according to their situation.

When creating a bubble chart with positivity and divisiveness as axes, clusters begin to arise:

These clusters illustrate that the positivity and divisiveness for subreddits which are similar by subject are strongly correlated with each other.

D. Humor

The most humorous communities were measured as: r/jokes, r/cringepecs, r/okbuddychicanery, and r/methwithoutcommunism, respectively. While r/okbuddychicanery is a humor subreddit, and the model classified posts and comments correctly, it is still highly surprising that this subreddit was measured as one of the most humorous; this subreddit in particular is based around humor which is presented in several nested layers of irony. It speaks to the quality of the BART model that these submissions are classified correctly.

Unsurprisingly, subreddits like r/explainlikeimfive, r/anger, r/republican, r/summonerschool, and r/programming were

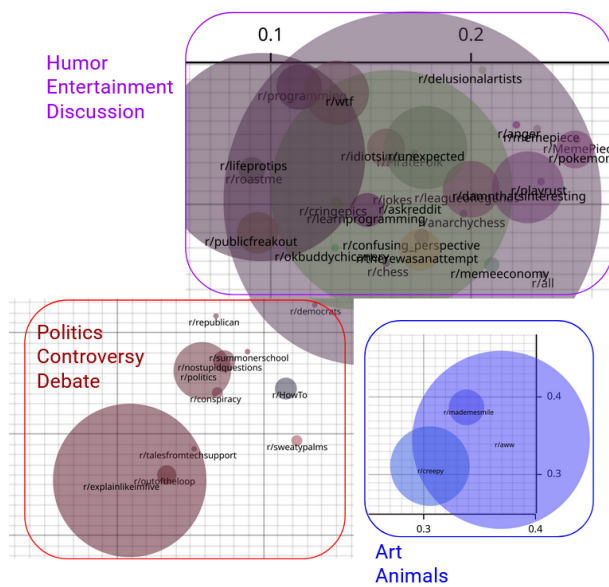


Fig. 5. Illustration of clusters according to divisiveness and positivity.

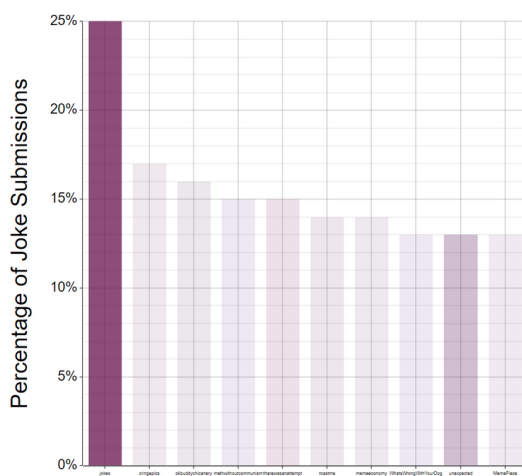


Fig. 6. A bar chart illustrating the percentage of joke content for each subreddit. Opacity indicates subreddit subscriber count, blue indicates subreddit positivity, and red indicates subreddit negativity.

the least funny subreddits. The content of these communities is dry because they're serious discussion communities. The average community falls between 8% and 9% humorous content, with the lowest being 1%.

E. Overall breakdown of content on Reddit

Any attempt to measure the distribution of subjects and attitudes of content on Reddit as a whole will inevitably succumb to selection bias: not all posts from all subreddits can be collected. There are millions of subreddits, and billions of posts. However, the overall distribution of subjects and attitudes of content on Reddit can be *approximated* by these attempts, and generally accurate conclusions can still be drawn.

Of all the subjects on Reddit, jokes are by far the most

Distribution of submissions by subject and attitude			
Subject	Distribution	Attitude	Distribution
Science	0.8%	Complaint	1.5%
Religion	1.0%	Mocking	2.5%
Movies	1.4%	Condemnation	4.2%
Sports	2.3%	Annoyed	6.1%
Music	3.4%	Agreement	10.6%
Technology	3.7%	Disagreement	19.2%
Politics	5.1%	Praise	24.4%
Food	6.3%	Inquisitive	31.4%
Personal	6.3%		
Animals	9.5%		
Discussion	22.9%		
Joke	28.6%		

Fig. 7. A table with the distribution of all collected Reddit posts according to their subject and attitude.

popular. While the majority of subreddits sampled are not joke subreddits, the joke subreddits have the largest number of subscribers. Discussion is also incredibly popular across the vast majority of the platform: this is likely because Reddit is the most conducive to public, long-form conversation when compared to other social media sites like Twitter or Facebook. Animals are the 3rd most popular subject; this is likely due to the huge subscriber count of the r/aww community, which consists mostly of animal pictures.

Total Submissions by Subject and Attitude

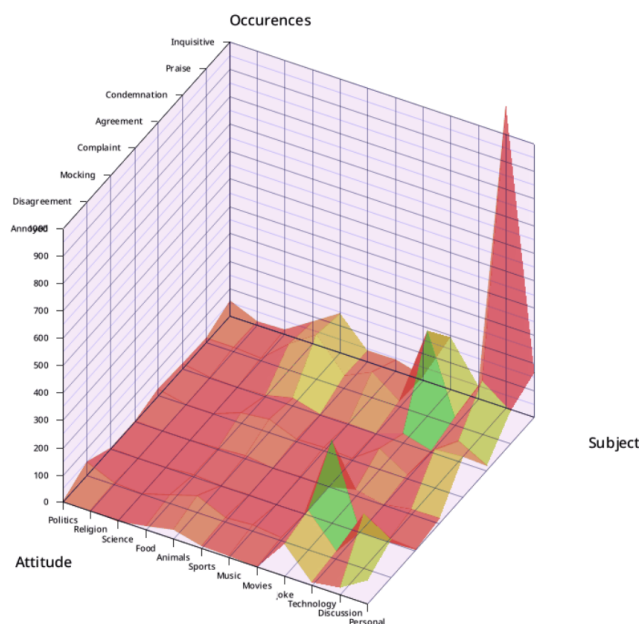


Fig. 8. A heat-map illustrating the occurrence of each combination of subject and attitude. An animated GIF of this image can be found in the project repository.

Presenting the occurrence of each submission's subject and attitude as a heat-map shows the data more clearly. By far, inquisitive discussion was the single most popular

combination of subject and attitude. There are also spikes at disagreement and discussion, agreement and discussion, mocking jokes, praising jokes, and inquisitive jokes. Other smaller, secondary spikes include praise towards animals, disagreement and politics, and praise towards music.

IV. ISSUES

The biggest issue was processing enough information for interesting results. The subject and attitude extraction was incredibly expensive even with CUDA parallelization and multiple machines running the BART model. Depending on the data size, a subreddit analysis could take over an hour to process a relatively small sample of 100 posts. In the beginning of the analysis phase, multiple group members could not run the code with CUDA support which slowed each subreddit's analysis to almost a full night of dedicated processing.

Looking at meta-issues about our actual plan for the project, we did not allot enough time to analyze the amount of data that we initially wanted. Because we planned and built our project sequentially and incorporated a milestone timeline, we reached the mid-point of November before fully analyzing data, which was insufficient given the resource cost of BART.

V. FUTURE WORK

Following the processing time constraints mentioned in the Issues section, we will work to increase the data we analyze across multiple dimensions. We will analyze more subreddits. We will also grab more posts per subreddit which should give a better representation of the subreddits attitudes and topics distributions. It will also benefit the richness of our data to include more metrics such as additional attitudes and topics, as well as new metrics such as named entity recognition and auto-summarization. Both of those are possible using BERT however selecting which variant of BERT will take additional research time.

We also plan to look across subreddits across multiple years to see how sentiment has shifted. Our team will also compare specific time frames with the general sentiment, for example an election or pandemic. This time-frame analysis will extend outside of the subreddit analysis and into a complete overview of Reddit's sentiment across time and specific events. The hypothesis would be that the entirety of Reddit would shift in one direction with individual nuance on a per subreddit basis.

Finally we will track user overlap between subreddits and how that correlates with those subreddits' sentiments. This could have interesting and unique results. Will users adopt the general sentiment of the subreddit they join? Will users maintain their sentiment trend regardless of the subreddit? It may also be possible that users tend to subscribe to subreddits which match their sentiment trend. This analysis has some privacy concerns since we are tracking users, however the goal of the research is not to attack or shame any individuals but to look at the general trends based on shared subreddit subscription overlap.

APPENDIX

Positivity vs. Divisiveness

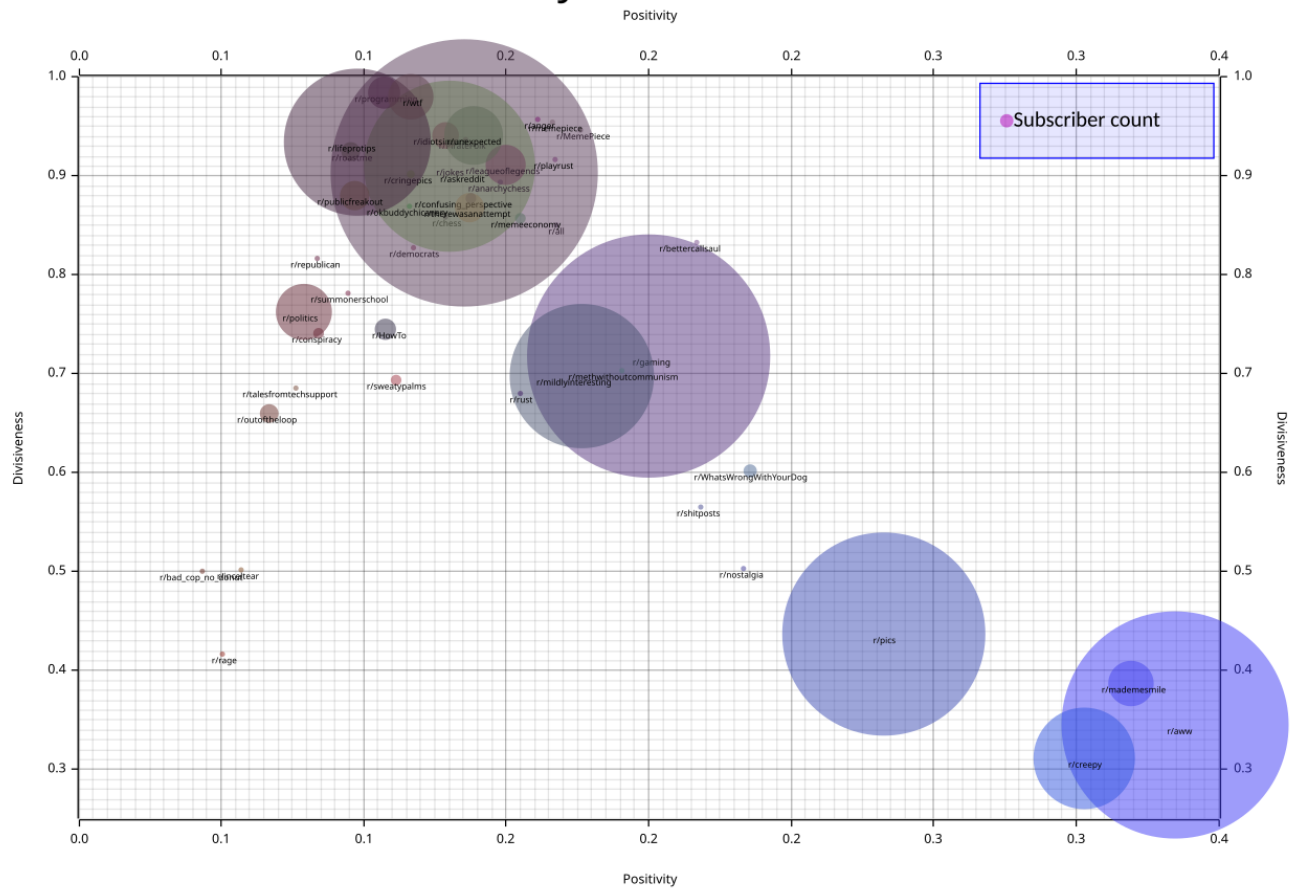


Fig. 9. This is the bubble chart where the clusters of positivity and divisiveness were discovered.

Attitudes of Submissions Towards Subjects

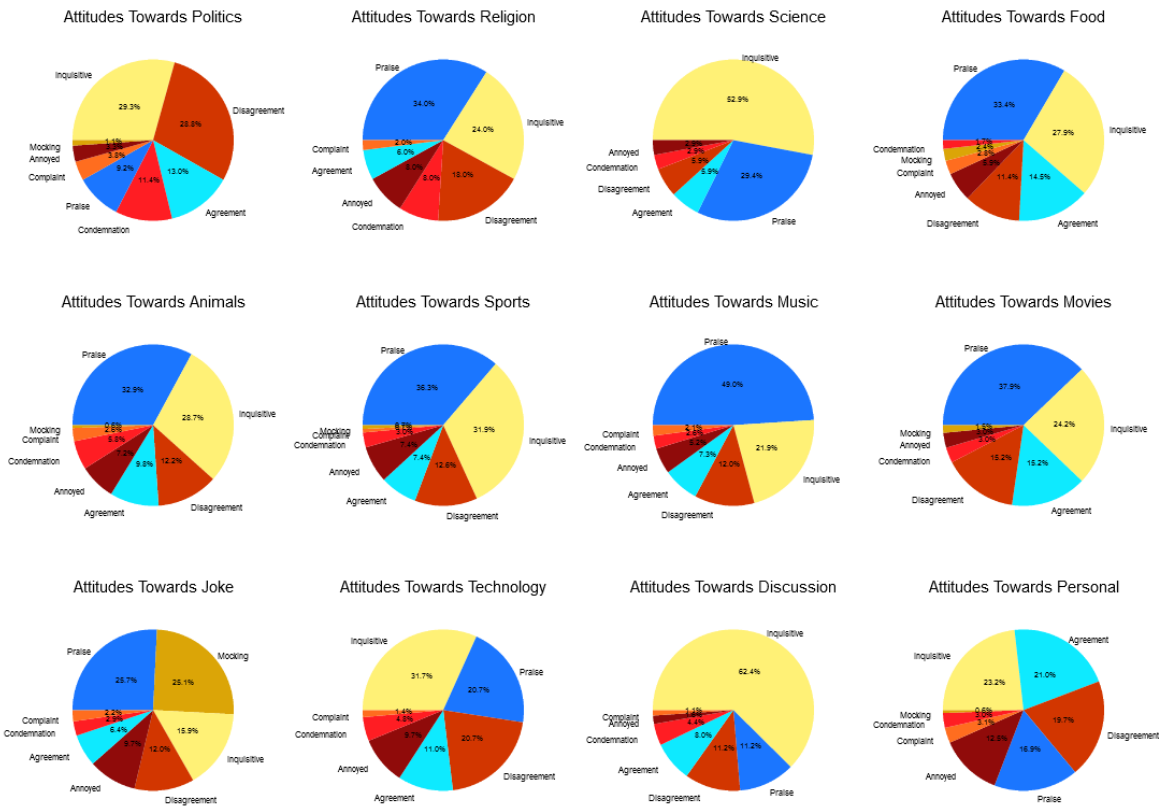


Fig. 10. These are the charts generated for the breakdown of Reddit according to subject. The PNG does not have full resolution on a PDF, but the SVGs in the project repository allows for seeing the data up close with high readability.

Attitudes and Subjects by Subreddit

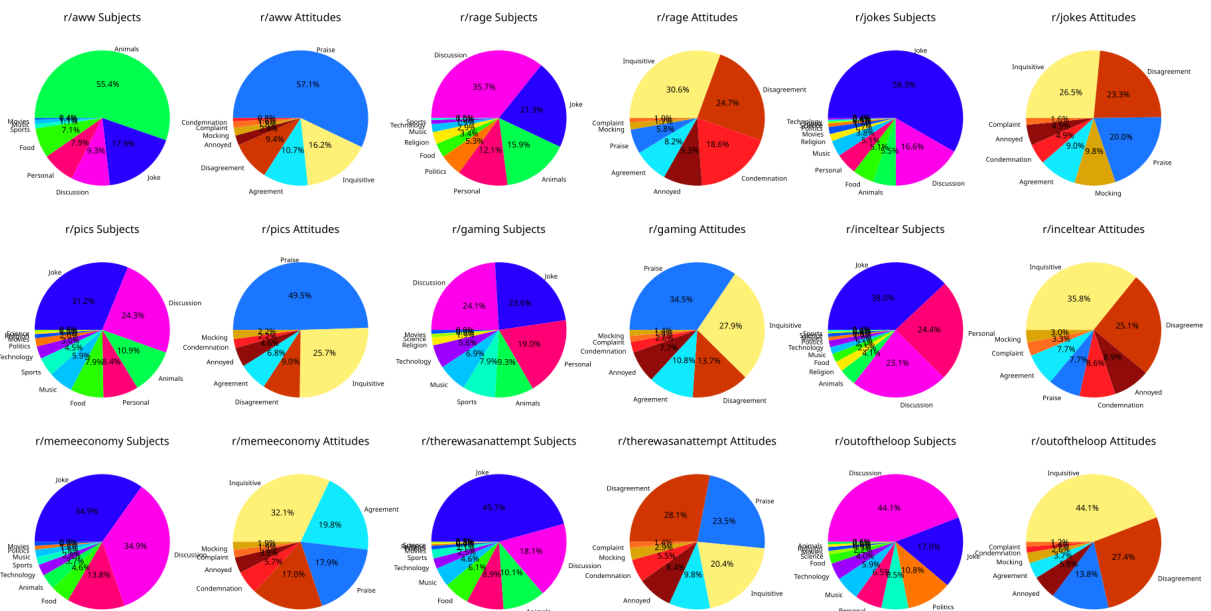
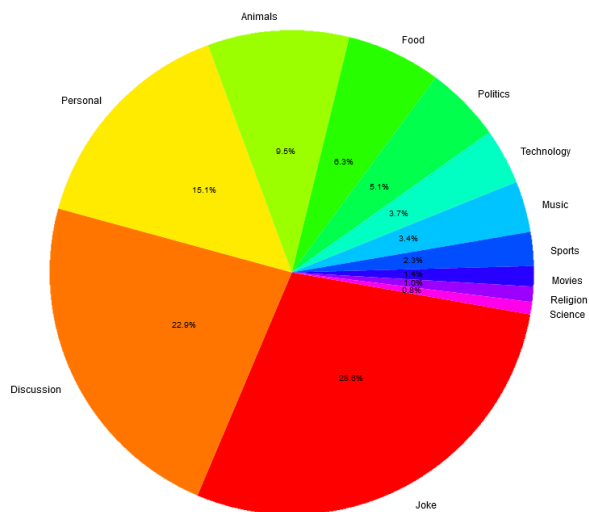


Fig. 11. These are the charts generated for the breakdown of various subreddits. The PNG does not have full resolution on a PDF, but the SVGs in the project repository allows for seeing the data up close with high readability.

Breakdown of Post Subjects



Breakdown of Post Attitudes

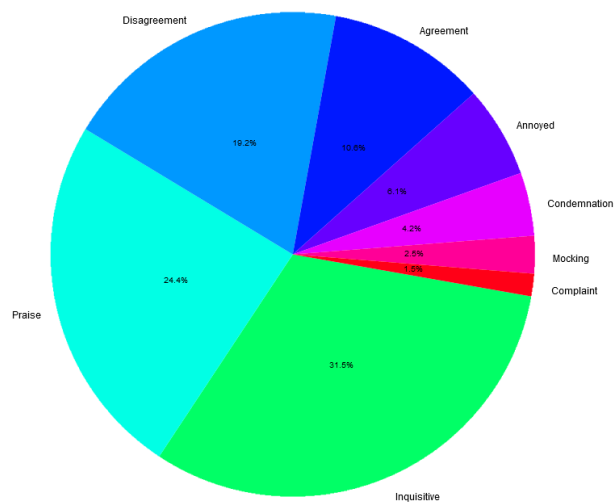


Fig. 12. These are the charts generated for the overall breakdown of Reddit. The PNG does not have full resolution on a PDF, but the SVGs in the project repository allows for seeing the data up close with high readability.

Humor Content Per Subreddit

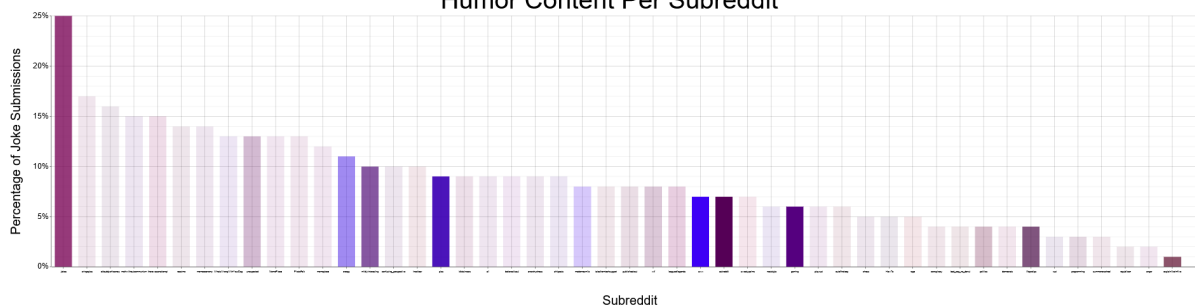


Fig. 13. This is the full chart of subreddits ranked by humor. The PNG does not have full resolution on a PDF, but the SVGs in the project repository allows for seeing the data up close with high readability.