Is lexical sentiment a good proxy for shared values in online communities?

Henry Anderson

# 1 Introduction

In the past fifteen or so years, digital communications have become the norm for modern societies: e-mail, text messages, and social media posts now comprise a significant portion of linguistic production and social interaction. More interestingly, the digital nature of these interactions opens the door to new forms of analysis that were previously inaccessible: modern computational methods can easily leverage corpora of many billions of words, hundreds of millions of interactions between individuals, and many millions of speakers (*speaker* here meaning *anyone who produces language in any context*). Furthermore, the inherently digital nature of these modern forms of linguistic data make it easier than ever to collect corpora of such sizes for analysis, and the wealth of metadata that comes nearly for free (timestamps, usernames, whether a message is a reply to another message, etc.) invite analyses that go beyond the purely linguistic and incorporate methodologies from anthropology, network analysis, and a wealth of other fields.

However, there are many issues plaguing the analysis of such data. While the tools to process and analyze them certainly exist, what they gain in breadth and quantity, they lose in depth. Even the best computers simply cannot perform the same level of deep and complex analysis as a human annotator working by hand. And while these tools are becoming increasingly accessible, they still require a highly technical skillset that is less common among the social sciences than the hard sciences (we can count computer science among this latter group)–but in the hard sciences, there is far less expertise in topics that inform social science research. Thus, there is an unfortunate dichotomy in much of the literature around digital language: the computational researchers focus on breadth, often at the expense of depth, while social scientists are often constrained by the difficulties in analyzing massive datasets.

Within the social sciences, there is also a tendency to privilege *face-to-face* language above textual, and especially digital, communications, with the latter being perceived as somehow "deficient" (e.g. by lacking intonational information and body language) compared to the former. Arminen, Licoppe & Spagnolli (2016) explicitly reject this approach, even going so far as to reject the terminology of *mediated communication* due to the implication that there exists some *unmediated communication*. Vandergriff (2013) provides evidence for this idea, demonstrating that there exist unique textual cues in digital communications that do not have direct counterparts in face-to-face communication, and thus should not be analyzed as attempting to mimic that form of communication.

Computational research often focuses on evaluating results not so much in the context of establish social

science literature as by objective *performance metrics*. A model for part-of-speech tagging, for instance, will be trained and run against a dataset whose POS tags are already known, and the accuracy of the model will be used to determine its viability. While this is sensible in the context of most model development–namely, with the ultimate aim of practical application–there is less focus given to a deep manual inspection of the results as is commonplace in the social sciences.

I myself have grown up always in the presence of computers, and always using digital media and digital communications. I thus find Arminen and Vandergriff's arguments refreshing and convincing, based on my own experiences and intuitions. However, I would further extend these arguments: I posit that analyzing entire communities in digital contexts should not be based on the idea that they are an imitation of face-to-face communities and interactions, but rather as a mode of communication with differing constraints, norms, and expectations that give rise to unique patterns of behavior that may not have parallels in in-person interactions and vice-versa. I find the analysis at the level of the community to be more interesting than at the level of individuals, given that linguistic and social behaviors can be enforced through community norms and values. The ways that these interact to develop a distinct identity for online communities, and for individuals within those communities, is of particular interest to me.

This leads to the research question that drives the work in this paper: to what extent can automated, computational methods, specifically domain-specific sentiment induction, identify shared values of online communities? I will investigate this question by replicating methodologies from natural language processing, and investigating the results not through performance metrics, but through manual evaluation in the context of domain and community insider knowledge. I hypothesize that such an approach is capable of identifying community values, though the extent to which this occurs remains to be seen. Particular attention must be paid to the *recall* performance of this approach: can it identify terminology indicative of meaningful community values, while *also* excluding terminology that is not meaningful?

It should be noted explicitly that the focus of this work is slightly different from the focus of LING 5347. Where the course focused on how *speakers* convey meaning with linguistic choices, this work focuses on how *communities* convey such information through *linguistic norms*. Communities are made up of individual speakers, though, and communal norms do not appear out of the aether–rather, they are arrived at by speakers.

# 2 Background

## 2.1 Sentiment Analysis

One of the many domains of natural language processing is the domain of sentiment analysis. In the broadest strokes, this encompasses assigning a *sentiment* value to words and documents. Simple (but none the less useful) forms might make use of a simple *positive-negative* spectrum. More sophisticated tools, such as Linguistic Inquiry and Word Count (LIWC; Pennebaker et al. 2015), offer a much broader spectrum of labels and may include multiple psychometric ratings of lexical items.

However, many sentiment analysis tools are constructed via static, curated dictionaries that are built for general-purpose use. An input document is checked, word by word, against the dictionaries, and the document's total score is updated for each word or phrase that the dictionary has a score for. This has some obvious drawbacks, especially when analyzing naturally occurring data that has not been cleaned: variations in spelling and orthography may cause significant false negative rates. Additionally, if the target text exists within some narrow domain, a dictionary–which is likely to be trained on a general corpus of the language–is apt to be inaccurate and miscategorize a number of words whose meaning and sentiment may not align with broader linguistic patterns. Loughran & McDonald (2011) showed that this is particularly problematic in the financial field, where typical language use is highly divergent from more common forms. This led, as the authors discovered, to extraordinarily high rates of word sentiment errors, necessitating the creation of a new dictionary specific to that field. Once such a dictionary was created, though, they found it performed very well–so there is still value in curated dictionaries.

But the broader point is that to analyze sentiment, a tool tailored to the document's particular domain is most appropriate. Such tools are not always available, however, especially for more niche domains. There is active research around how to address this using *unsupervised learning* techniques from machine learning[1], allowing the discovery of natural patterns that can be exploited. The appeal of such approaches is that they do not require any prior knowledge of the domain or corpus to be effective, and are readily deployable–but at the expense of the interpretability and accuracy associated with curated dictionary approaches.

Differential language analysis (DLA), an approach used by Andrew Schwartz and his collaborators (Schwartz et al. 2013), is a related framework that is conceptually and, often, technologically simple to implement and can be purposed for sentiment analysis. The approach requires no prior knowledge about populations, and discovers contrasting linguistic features that can differentiate between groups, e.g. students at the beginning of a course versus those same students at the end of a course, or a control group and ex-

[1] Unsupervised learning is any statistical or machine learning approach where there are no target variables for modeling. The aim is to discover the structure within a dataset, rather than to learn the relationship between a target variable and its hypothesized predictors.

perimental group. While not implemented in my code, DLA is a useful conceptual framework for comparing communities' language use: look for differences between them, and see how those differences might map to meaningful information about each community.

Hamilton et al. (2016) present a novel method for unsupervised sentiment analysis, tailored for smaller corpora where there is little to no assumed or prior knowledge about word sentiments. This is the approach that will be the focus of this paper; I have re-implemented their algorithm on their original dataset. The authors' approach is based on automatically learning vector representations of words based on their contexts (this is similar to, and intuitively an extension of, simple N-gram collocations), and inducing the sentiment of words in the corpus based on patterns of co-occurrence. A small set of seed words, whose polarity is known or assumed, is used as a starting point to induce the sentiment of the rest of the corpus. The specifics of Hamilton et al. (2016)'s work will be discussed in much more detail in the Methodology section.

## 2.2 Research in Online Communities

Any computational or analytical technique is immaterial and not particularly interesting to the social sciences when taken in isolation. Only in application, when used to learn something meaningful about the world, do any of these tools gain their value. How tools like sentiment analysis have been applied to real data is far more interesting and important than then the tools themselves.

In digital spaces, the analysis of *communities*, as opposed to merely *individuals*, is a field of study where many computational tools can be fruitfully applied. Digital communities like Reddit, Twitter, and even comments sections of YouTube videos generate staggering amounts of data on human interactions. The resulting linguistic data, quantitative data, and *network* data (i.e. patterns of interaction between users) all lend themselves naturally to a computational approach, and are often too massive to even consider analyzing by hand.

This has, in turn, led to a very interesting body of research and series of discoveries regarding online communities. Zhang et al. (2017) measure the *distinctiveness* and *dynamicness* of Reddit communities, meaning, respectively, how distinct the language used by each group is, and how stable language use is over time. They link these measures to *insider* and *outsider* classifications of users to determine what users are likely to return to the subreddit later on as regular participants. Danescu-Niculescu-Mizil et al. (2013) similarly tracked how the age and authority of users within communities correlates with linguistic choices, specifically the adoption of new conventions.

Research in this area often follows Vandergriff (2013) and Arminen, Licoppe & Spagnolli (2016), in that it considers digital and online communications to be full, rich modes of communication that present a

set of non-overlapping linguistic and non-linguistic possibilities and challenges when compared to in-person interactions.

## 2.3 Distributional Semantics

One of the most important ideas in modern NLP is the *distributional hypothesis* (first proposed by Harris (1954); built upon by Rubenstein & Goodenough (1965), Sahlgren (2008), among many others): the idea that there is structure in the *distribution* of linguistic features. In modern work, the term *distributional hypothesis* has become nearly synonymous with one specific ramification: words with similar meanings will tend to appear in similar contexts, i.e. have similar *distributions* (and vice versa). Or, as Firth (1958) put it: a word is characterized by the company it keeps. This is the foundation of *distributional semantics*, the study of word meaning using this hypothesis. Most notably, Mikolov et al. (2013)'s Word2Vec algorithm and Pennington, Socher & Manning (2014)'s GloVe algorithm have become standard tools for generating vector representations of words based on their contexts (specifically, based on their co-occurrences with other words), such that words with similar contexts have very similar vector representations.

These algorithms and other likes them are usually referred to under the heading of *word embeddings*, since they involve generating a very large matrix of word co-occurrences, then performing a mathematical *embedding* of that matrix into a lower-dimensional space. A crucial factor of these algorithms is that they are unsupervised, requiring no prior information about the language or corpus to learn relationships between words. This makes them excellent for working in narrow domains, assuming enough text is available to build a good model (learning word distributions requires very large amounts of data, due to the huge number of words available and the low probability of any single word occurring–large corpora are required to achieve a reasonable sample).

A toy example of word embedding results is that, in a large corpus of English, we would expect the words *king* and *queen* to both appear in proximity to words about royalty, heads of government, and names of specific countries with monarchies. Thus, the *distance* between these two words would be small. Conversely, the words *king* and *dinosaur* are unlikely to appear in similar contexts, and thus the *distance* between them is large.

Since the focus of this paper is on identifying community-specific values, word embedding algorithms are a perfect fit. Hamilton et al. (2016)'s method uses word embeddings to generate sentiment values–this will be discussed in more detail in Section 4. However, rather than using GloVe or Word2Vec, which are excellent choices for corpora of billions of words or more, they use a *pointwise mutual information* matrix, followed by *singular value decomposition*, to acquire their word vectors; in their experiments, this method outperformed

Word2Vec and other embedding algorithms.

# 3 Data

The dataset used is, as in Hamilton et al. (2016), the Reddit Public Comment Corpus. Reddit (https://www.reddit.com) is a massive, public discussion board that does not focus on any single topic. Rather, users may create *subreddits*, which are discussion boards focused on a single topic. Subreddits are typically referred to as *r/subreddit_name* (read as "R (slash) subreddit name"). I have selected a different set of subreddits from the ones Hamilton et al. focus on, the full list of which is presented in Table 1, along with the total number of posts and words (pre- and post-processing) for each. A brief description of each subreddit is in Table 2. These subreddits were selected based on the following three criteria:

1. Number of posts: subreddits with fewer than 100,000 posts were discarded, to ensure enough linguist data existed in any selected subreddits.

2. Familiarity: only subreddits whose content I am personally familiar with were selected, as I can bring domain-specific prior knowledge to such post and reason about the results with greater context and certainty.

3. Linguistic markedness: only subreddits which I intuit would be linguistically marked compared to other subreddits were selected. This selection was based on my prior knowledge about the different subjects and communities, and thus is prone to some bias.

This is not a comprehensive survey of subreddits on the site, and it is likely that many of the ones selected, such as r/darksouls and r/pcmasterrace, will likely have significant overlap in their userbase. We might thus expect to see similar linguistic patterns between these communities to a much greater degree than we would between r/darksouls and r/guns, a subreddit devoted to firearm-related interests. While this may be an issue for a fully automated approach with minimal human intervention, the hope is that by leveraging narrower, but deeper, domain-specific knowledge, meaningful patterns will still surface.

Table 1: Table of the subreddit names and basic count statistics. The "Raw" counts are for the full set of posts in each subreddit, before any processing was done; the "Processed" counts are the counts after applying the preprocessing and filtering steps described in Section 4. A "word" is here taken to mean any string of non-whitespace characters surrounded by whitespaces. Much of the discrepancy in post count is due to short posts, under 5 words, which were discarded; the word counts, which are far more extreme, account for the removal of high-frequency stopwords that contain little semantic information.

| *Subreddit* | *Raw Posts* | *Raw Words* | *Processed Posts* | *Processed Words* | *%Posts Kept* | *%Words Kept* |
|---|---|---|---|---|---|---|
| r/2007scape | 2,694,974 | 85,099,342 | 746,470 | 10,119,447 | 27.70% | 11.89% |
| r/4chan | 3,922,974 | 106,955,645 | 741,811 | 8,822,004 | 18.91% | 8.25% |
| r/atheism | 32,276,496 | 1,284,994,468 | 10,885,681 | 236,486,570 | 33.73% | 18.40% |
| r/darksouls | 2,905,068 | 108,375,236 | 1,036,108 | 17,715,770 | 35.67% | 16.35% |
| r/linux | 3,068,964 | 124,159,305 | 1,123,942 | 25,292,306 | 36.62% | 20.37% |
| r/pcmasterrace | 13,289,598 | 450,074,627 | 3,884,058 | 55,706,149 | 29.23% | 12.38% |
| r/politics | 45,809,992 | 1,920,709,203 | 16,692,897 | 393,522,380 | 36.44% | 20.49% |
| r/runescape | 3,532,152 | 117,617,607 | 1,047,984 | 15,548,826 | 29.67% | 13.22% |
| Total | 107,500,218 | 4,197,985,433 | 36,158,951 | 763,213,452 | 33.64% | 18.18% |

Table 2: Brief descriptions of each subreddit.

| *Subreddit* | *Description* |
|---|---|
| r/2007scape | Dedicated to Old School RuneScape, a version of the Massively Multiplayer Online Role-Playing Game (MMORPG) RuneScape. Jagex, the company who develops and maintains the game, opened servers in 2013 that ran a version of the game's code as it existed in August 2007. These servers are maintained alongside the main servers, and development on this version of the game has diverged from the main game. |
| r/4chan | 4chan is another major discussion and image-sharing board that predates Reddit considerably, and has a reputation for extremely offensive material and discussion on many of its boards, as well as a very strong tendency towards trolling.[2] According to the description fo this subreddit, it exists to make Reddit a worse place, and much of the content currently on the board is typical of 4chan. |
| r/atheism | Dedicated to discussion of atheism, agnosticism, secularism, and irreligiosity. Many of the posts in this subreddit, as of writing, are oriented more at attacking religious individuals and institutions than on promoting non-religious ones. |
| r/darksouls | Dedicated to the Dark Souls series of video games. These games are known for being very difficult, and featuring a wide range of highly varied equipment players can use, as well as difficult boss fights. However, the games are very popular among the gaming community for their novel gameplay, combat mechanics, lore, and environment design. |
| r/linux | Dedicated to discssion of Linux operating systems. Since Linux is inextricably tied to the Free Software movement and culture, there is extensive discussion of Free Software as well, and a general disdain for closed-source, proprietary software. |
| r/pxmasterrace | Dedicated to PC gaming enthusiasts. The name comes from a sarcastic description of PC gamers who insist that gaming on consoles, e.g. Playstation and Xbox platforms, is objectively inferior. This subreddit, however, is not a humorous or tongue-in-cheek one, but focuses on genuine enthusiasm for PC gaming and associated topics, e.g. hardware enthusiasts and video gaming generally. |
| r/politics | Dedicated to discussion of US politics. |
| r/runescape | Dedicated to the current version of RuneScape (not Old School Runescape/2007Scape), the MMORPS developed by Jagex. |

# 4 Methodology

The methodology follows closely from Hamilton et al. (2016)'s methodology. All work was done using the Python programming language[3] The full Reddit Public Comments corpus was downloaded, then a script was written to count the number of posts in each of the subreddits contained in the corpus. Beginning with the highest post counts and working down, the subreddits in question were selected according to the parameters listed in Section 3.

Note that this work began with an attempt to utilize the code provided by Hamilton et al. (2016) on Github. However, the provided codebase is written in Python 2.7 (an old version of the language; the current version is 3.6.3 as of writing this), and continuously throws errors when I attempt to run it with the correct version of Python. As such, a very large amount of time–the vast majority, in fact–was spent re-implementing their algorithm from scratch.

The ultimate aim is to generate a set of sentiment values for each word, representing how *positively* or *negatively* they tend to be used within a given subreddit. At a very high level, Hamilton et al. (2016)'s approach measures an approximate *distance* between words based on their contexts: two words that are often used in very similar contexts should have a very small distance between them (e.g. *king* and *queen*), while words that are often used in very different contexts should have very large distances between them (e.g. *aeronautics* and *cuneiform*). A small set of *seed words* of known sentiment is then selected using *a priori* knowledge. The sentiment value of these seeds is then *propagated* through the rest of the corpus, with words near the seed words gaining a similar sentiment value.

## 4.1 Computational Methodology: Data Processing and Modeling

A basic NLP preprocessing pipeline was applied to the raw post data. The processing steps are intentionally simplistic–due to the large quantity of data, computational speed is a necessity in order for the code to run in a reasonable amount of time. However, the quantities of data are assumed to smooth out any error introduced by this. More thorough processing, e.g. lemmatization/stemming and part-of-speech tagging, may well provide better results, but would make the runtime prohibitively long. Further, due to the size and very "dirty" nature of the datasets, simpler rules can help remove noise, e.g. rare misspellings, rare words, and other unusual features that may negatively impact the later processing steps. The processing steps, in order, are as follows:

1. Remove all metadata from the posts, leaving only the post text.

[3]All code, as well as data files containing the full final results, will be available on Github by the end of December 2017: https://github.com/andersonh-uta/PragmaticsFinal

2. Convert the text to lowercase.
3. Remove URLs, HTML tags, and non-alphanumeric characters (performed via regular expression search).
4. Tokenize the text by splitting at whitespaces.
5. Remove stopwords[4], single-character tokens, and tokens that only consist of numeric characters[5].
6. Find multi-word phrases, up to 4-grams.[6]
7. Discard any posts containing fewer than 5 tokens after the above processing steps have been applied.

Discarding posts below a minimum length is an important step here, since Hamilton et al. (2016) use word co-occurrences to obtain the final results. Extremely short posts, which comprise a large proportion of the original corpus, do not provide as much information about word co-occurrences as longer posts, and thus are considerably less valuable for this sort of analysis. Further, this step again decreases the computation time, which is a non-trivial consideration given that the starting dataset consisted of just over 100 million posts and nearly 4.2 billion words.

After the NLP preprocessing, a word co-occurrence matrix was constructed using skip-grams with a symmetric 5-word window–two words are considered to form a skip-gram if they occur within 5 words of each other. A positive pointwise mutual information (PPMI) transform was applied to this matrix using smoothed probabilities, per equation (1):

$$M_{i,j}^{PPMI} = max\left\{log_2\left(\frac{\hat{p}(w_i, w_j)}{\hat{p}(w_i)\hat{p}(w_j)}\right), 0\right\} \tag{1}$$

Where $\hat{p}(w_i)$ represents the smoothed probability of the word $w_i$ appearing in the corpus, and $\hat{p}(w_i, w_j)$ represents the smoothed probability of the skipgram $(w_i, w_j)$ appearing in the corpus. A singular value decomposition (SVD) was applied to the resulting matrix to generate a 300-dimensional vector for each token. The angle between each word's vector was computed, resulting in a very large and densely filled numeric matrix.

Intuitively, this matrix represents a notion of *distance* between words: each row and column correspond to some unique token in the corpus, and the values at the intersection of those columns represents how similar the distribution (i.e. context) of those two words are, and thus, represents some notion of how similar their meanings are.

All but the top 10 largest values in each row were replaced with zeros, such that the matrix only represents the distance from any word to its 10 nearest neighbors. All word sentiment values were initialized to $\frac{1}{|V|}$,

[4]Since the sentiment analysis is a primarily semantic one, such words, which have low semantic content, are removed to improve runtime and reduce noise.

[5]The stopword list from Gensim (Řehůřek & Sojka 2010), a third-party Python library for NLP and topic modeling, was used.

[6]Gensim's phrase-finding tools were used for this. These tools use the frequency of two words appearing in sequence; if they appear considerably more often than expected by random chance, they are considered to form a multi-word phrase.

Table 3: The list of positive and negative seed words used for the sentiment induction.

| *Positive Seeds* | *Negative Seeds* |
|---|---|
| good | bad |
| lovely | horrible |
| excellent | poor |
| fortunate | unfortunate |
| pleasant | unpleasant |

where $|V|$ represents the total size of the vocabulary for the given subreddit. The sentiment labels of the positive and negative seed words, using the same seeds as Hamilton et al. (2016) (copied in Table 3), were then propagated by iteratively updating the sentiment values until the total magnitude of the changes was below $1x10^{-6}$, using equation 2 from Hamilton et al. (2016):

$$\mathbf{p}^{t+1} = \beta \mathbf{T} \mathbf{p}^{t} + (1 - \beta)\mathbf{s} \tag{2}$$

Where:

- $\mathbf{T}$ is the transition matrix described in section 3.2 of Hamilton et al. (2016)
- $\beta$ is a controllable parameter, with large values (near 1) favoring *local consistency* (i.e., words' sentiments being similar to the sentiments of other words used in the same contexts), and low values (near 0) favoring *global consistency* (i.e., correct sentiment for the seed words). The value of $\beta$ was set to 0.9, as in the code provided by Hamilton et al. (2016).
- $\mathbf{s}$ is a vector set to $frac1|S|$ at indices corresponding to seed words, where $|S|$ is the number of seed words in the set (positive or negative).

This process was repeated twice for each subreddit, once using the positive seed words and once using the negative seed words. Total sentiment scores for each word were then calculated as:

$$sentiment_{total} = \frac{sentiment_{positive}}{sentiment_{positive} + sentiment_{negative}} \tag{3}$$

After reaching convergence, the sentiment scores for a single subreddit are normalized to have zero mean and unit variance. This ensures that positive sentiment corresponds to a positive value, while negative sentiment corresponds to a negative value.

Finally, each word's sentiment value was multiplied by the logarithm of its *document frequency* (the number of documents, per subreddit, in which the word appears). Intuitively, this is intended to account for frequency effects: infrequent words with very high sentiment scores are likely artifacts from the embeddings,

and not interesting; interesting words are those that have both high frequency *and* marked polarity.

The total runtime for the full codebase was in excess of 36 uninterrupted hours from start to finish.

### 4.2 Qualitative Methodology: Manual Inspection of Results

With the list of sentiment values delivered by the computational approach, each subreddit's list was sorted by the absolute value of the sentiment score, and the top weighted terms for each were manually investigated. Using my own knowledge about the communities and their subjects of interest/discussion, I manually culled these lists for words that I identified as being either particularly likely to align with the community's values, or particularly likely to be false positives. I reduced each list down to approximately 20 words, selecting about 10 terms which I identified as seeming to encapsulate what I understood the community's values to be, and about 10 which seemed to be semantically related but which I was unable to intuit about. I ignored words such as *yeah* and *terrible* that did not seem to have any content relation to the subreddit in question.

Focusing on r/atheism, whose top-ranked tokens were an interesting, though confusing, mix, 25 posts containing each word were randomly sampled from the corpus, totalling 475 posts. These posts were then manually read, paying close attention to the use of the selected tokens to see how they map to known facts about the communities in question. Random selection of posts is intended to ensure that the sampling is not biased by me seeking examples which confirm or deny the results. Since I have personal familiarity with many of the subjects covered by these subreddits, accounting for my own bias is especially important. For illustrative purposes, and because the analysis of r/atheism provided the most concrete and definitive results, I will focus the rest of the discussion on this subreddit in particular.

## 5 Results

Sample scatter plots, plotting the sentiment against the logarithm of the document frequency, are shown in Figures 5 and 5. Plots for all subreddits are in Appendix B. All figures have been set to the same X and Y axis bounds to make comparisons easier. It is interesting to note that each subreddit has a sharply defined maximum and minimum sentiment boundary. I suspect that this is likely an artifact of the algorithm, but this would require more investigation.

On many of the plots, there is a clear visible trend of thin vertical lines concentrated as specific sentiment scores. I suspect that this is an effect of the algorithm, due to the regularity of these clusters. Recall from Secton 4 that after computing the similarity matrix, we removed all but the 10 largest values from each row. This makes it entirely possible that there were a number of word clusters that were only directly connected to a small number of seed words, which would naturally cause these words to only gain sentiment from a limited

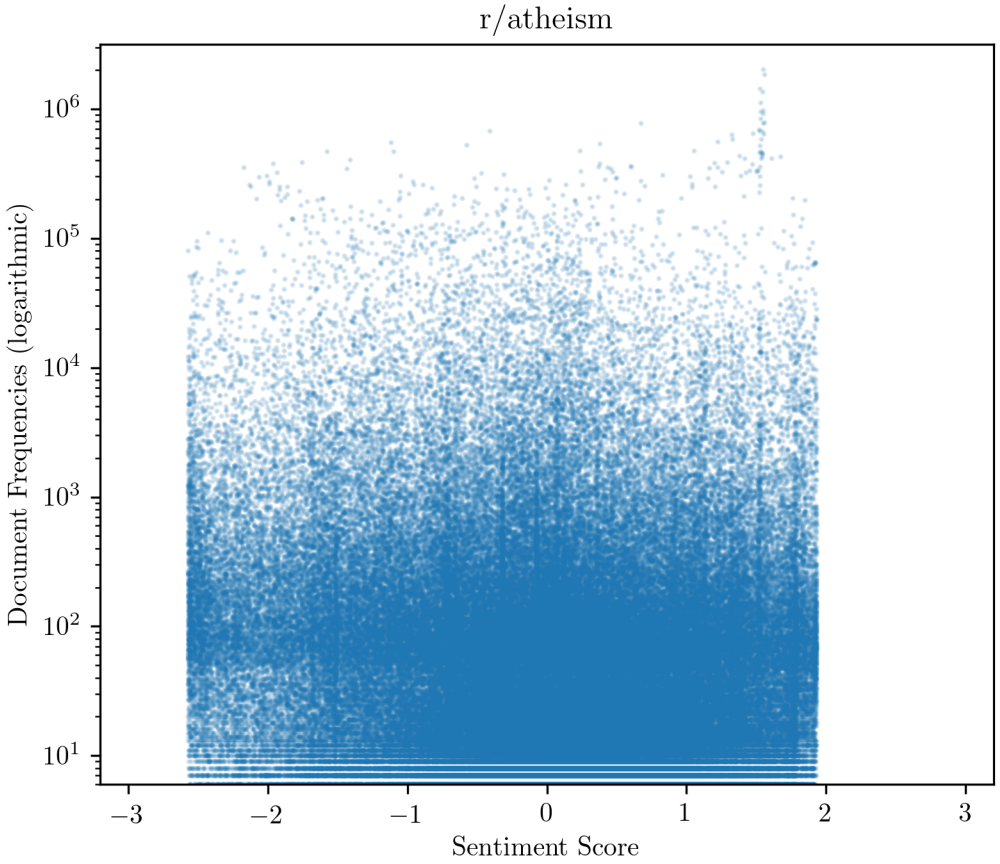


Figure 1: Plot of sentiment scores against document frequencies for the r/atheism subreddit.

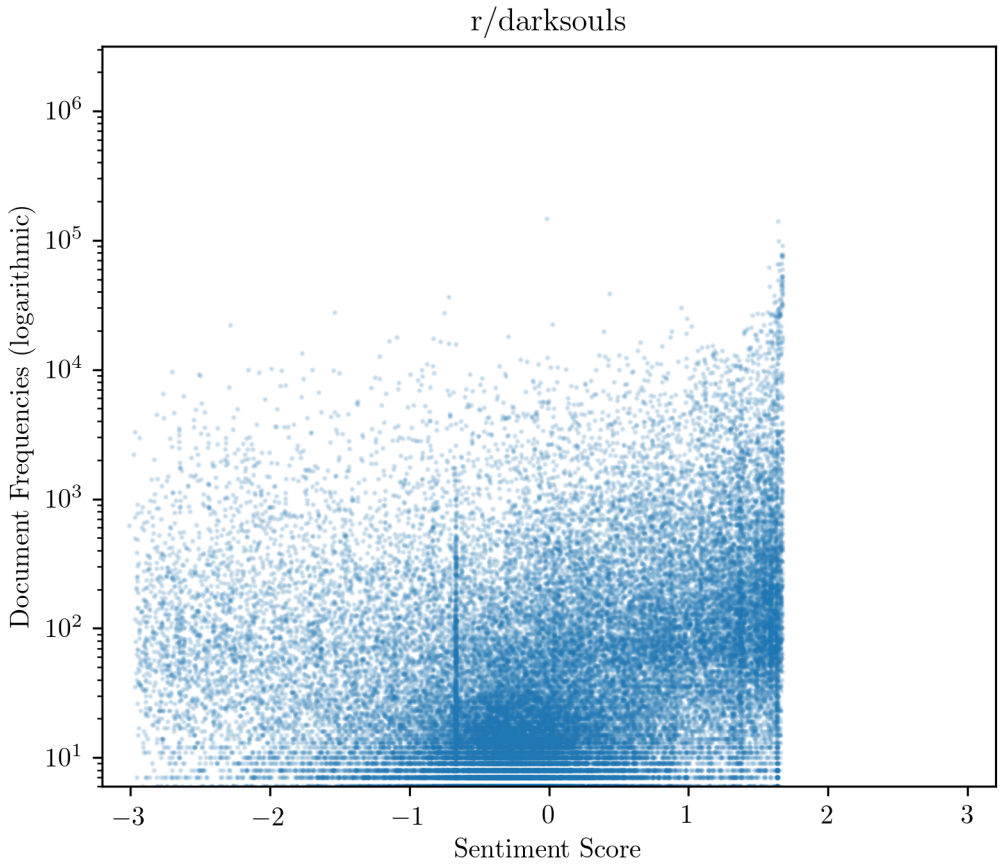


Figure 2: Plot of sentiment scores against document frequencies for the r/darksouls subreddit.

set of seeds, and thus all converge to similar sentiment values. This question merits further investigation in later work.

Tables showing the top 20 weighted tokens for each subreddit, along with their weight and sentiment, is in Appendix A. The tokens for r/atheism and r/darksouls are also shown in Tables 4 and 5, respectively. The manually selected tokens are listed in Table 6.

Table 4: Top 20 tokens in the r/atheism subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| poor | -2.5699 | 29.0158 |
| rights | -2.5058 | 28.7300 |
| worse | -2.5104 | 28.5265 |
| actions | -2.4984 | 28.3350 |
| act | -2.4303 | 28.2110 |
| horrible | -2.5624 | 27.7635 |
| bad | -2.1701 | 27.7222 |
| justify | -2.5405 | 27.5474 |
| worst | -2.5276 | 27.4786 |
| excuse | -2.5426 | 27.4007 |
| terrible | -2.5109 | 27.3824 |
| violence | -2.5271 | 27.3808 |
| blame | -2.5437 | 27.2619 |
| fucked | -2.4994 | 26.7905 |
| situation | -2.3348 | 26.7892 |
| action | -2.4765 | 26.7629 |
| responsible | -2.5301 | 26.7539 |
| making | -2.1289 | 26.5559 |
| potential | -2.5360 | 26.3993 |
| aren | -2.1227 | 26.3976 |

Table 5: Table of the top 20 tokens for the r/politics subreddit, sorted by highest weight.

| Token | Sentiment | Weight |
|---|---|---|
| hate | -2.7022 | 24.7787 |
| annoying | -2.7652 | 24.2943 |
| poor | -2.9664 | 24.0108 |
| worst | -2.8095 | 23.6353 |
| pain | -2.9352 | 23.4862 |
| sucks | -2.7605 | 23.0452 |
| horrible | -2.9773 | 22.9413 |
| stupid | -2.6185 | 22.8909 |
| face | -2.5070 | 22.8800 |
| bad | -2.2835 | 22.8432 |
| fair | -2.4996 | 22.7548 |
| terrible | -2.6849 | 22.6707 |
| bullshit | -2.7148 | 21.6689 |
| internet | -2.6511 | 21.6644 |
| awful | -2.8301 | 21.5477 |
| chosen_undead | -2.6511 | 21.5430 |
| worse | -2.5282 | 21.4756 |
| cheap | -2.4901 | 21.4603 |
| connection | -2.6509 | 21.3090 |
| useless | -2.5461 | 21.1824 |

Table 6: The manually selected tokens from each subreddit. "Expected" tokens are tokens I identified as conforming to my understanding of the community values based on their weight and on their sentiment value; "Unexpected" tokens are tokens which shows up with high weights, but which did not seem to conform to my understanding of the community's values. Underscores indicate that two words have been joined into a multi-word phrase.

| Subreddit | Expected | Unexpected |
|---|---|---|
| 2007scape | noob, poor, kid, free, scammed, trolling, immature, minor, easyscape, inb4 | giving, new_players, huge, logic, perfect,proper, properly, math, defend, software |
| 4chan | good_job, freedom, muh, pay, buy, poor, rights, suicide, black, reddit | account, fucking, money, bastard, stupid, hate, father, rich, got_banned, abortion |
| atheism | violence, rich, moral, morality, dangerous, genocide, slavery, oppression, insane | poor, rights, freedom, food, constitution, africa, guilty, separation_church_state, amendment, equal_rights |
| darksouls | connection, lost_izalith, ugly, dark_souls, damage, matchmaking, help, work, giant_skeletons, ng, boss, sword, armor | fair, chosen_undead, demon_ruins, duke_archives, bells, titanite_slabs, invaded, ease, titanite_slab, worthwhile |
| linux | linux, work, desktop, gnome, kde, mint, bloated | windows, unity, mac, osx, xp, vista, windows_xp, win7, widely_available |
| pcmasterrace | pcgiveaway, framerates_high, low_temperatures, temps_low, dust, bugs | monitor, overclock, oc, gigabyte, msi, 60hz, asus, 144hz, dell, psu |
| politics | hitler, koch_brothers, nazis, richest, richer, genocide, billionaires, trickle, upper_class, trickle_economics, government | poor, rich, middle_class, jew, book, backs, job_creators, sympathy |
| runescape | gwd, poor, pvm, gf, soloing, pvming, gg, xp, slayer, jagex, combat, runescape, game | boss, bosses, bossing, rich, corp, bandos, rares, million, millions |

# 6 Discussion

Before proceeding, a brief recap is in order. I am trying to determine the extent to which an automated approach, using the methodology described above, can reliable detect linguistic features (here, lexical features) that online communities use to encode and convey their values and interests. At this point, the automated analysis has been completed, and has returned a list of lexical items appearing in each subreddit, the number of distinct posts within that subreddit in which they occur, and their normalized sentiment value.

Now comes the investigation of these results, interpeting the sentiment values and weights through the lens of prior knowledge and experience with the communities and their primary interests and topics. Hopefully, the sentiment and weights of the tokens will provide enough information to infer collective community values and differentiate, at least qualitatively, each community. If this is the case, then it can be said with certainty that this sort of approach does, in fact, allow insight into community values. If this is not the case, the source may be twofold: 1) the methodology may be insufficient to perform this task, and alterations or extensions may be required, or 2) such information simply may not be retrievable from word distributions and sentiment.

## 6.1 Preliminary Observations

Examining the plots of document frequency versus sentiment, several interesting trends are apparent. All subreddits have a large, very dense cluster of low-frequency, near-0 sentiment terms–this is not unexpected, since the majority of words are likely used in fairly neutral ways most of the time. What is not captured, however, is words that are used frequently with highly disparate sentiment by different sub-populations; these would be scored near their average sentiment. Certain subreddits also have a noticeable shift: r/darksouls has a slightly negative-leaning bias in this central clump; r/darksouls also has a fairly distinct high-density cluster of very positive words occurring with document frequencies between approximately 100 and 1000. Similarly, r/linux's main cluster is significantly more negative than positive, though r/linux also has a much longer tail of high-sentiment words.

There is an interesting cluster of words in r/politics that may merit a closer investigation in the future. Its chart displays a tight vertical cluster of extremely high document frequency and extremely high sentiment words.

When sorting the tokens by my weighting scheme, there was considerable homogeneity among the sentiment values of high-weight items, almost universally being highly negative terms. r/linux is an exception, being dominantly positive; r/4chan is better mixed between positive and negative. On further inspecting the plots, this corresponds to imbalances in sentiment scores: most subreddits extend noticeable further in one

direction than the other, e.g. r/darksouls, which has maximum scores around +1.6, but minimum scores of nearly -3.0. Due to the weighting scheme's construction, these differences in sentiment value appear to be having a disproportionate effect on the weights. There may be a number of possible fixes, including scaling sentiment scores to a [-1,1] interval rather than zero mean and unit variance; however, this approach would need experimental validation, and the approach used here follows the method in Hamilton et al. (2016).

In spite of this imbalance, a number of expected terms appear highly-ranked. r/darksouls has the names of boss monsters and cetain equipment as being high-sentiment; given that the bosses and the equipment are major draws of the series, and major points of discussion among the player base, this seems to reasonably reflect the content of the forums and generally follows what was expected a priori in terms of high-sentiment items. However, many of the bosses and names of areas in-game are of *negative* sentiment, while the names of items (e.g. swords and armor) are *positive.* On its face, this is perhaps confusing: do Dark Souls players not like one of the core draws of Dark Souls? From my own experiences with the game and the player base, I find this unlikely. Many boss fights can be frustrating and difficult–I find it more likely that players are simply venting or complaining about the difficulty of a fight, or even berating (seriously or humorously) themselves and other players for being unable to defeat such bosses.

As the r/darksouls example shows, these results need to be contextualized to be interpreted and understood properly–a mere surface analysis does not lend itself to a full explanation.

## 6.2 Investigating Example Posts

Though there are many terms weighted highly that are perfectly expected, there are equally many confusing results. To pick just two such examples: r/atheism has *freedom* ranked at -2.35, and *food* at -2.32, both with very high document frequencies (64,479 and 56,615, respectively). Do users of r/atheism thus hate freedom and food? I find this exceptionally unlikely. It is important to recall what exactly these sentiment scores represent: the words *food* and *freedom* often appear *in similar contexts to the negative seed words.* Thus, we cannot interpret what r/atheism, collectively, thinks about the vast topic of *food.* We can only state that they tend to talk about food when talking about negative topics.

Of the tokens listed for r/atheism in Table 6, some were distributed as one would expect. *Violence* was most often use in the context of condemning violence and attempts to paint entire groups as violent as in:

> *Context: responding to a claim that Mormons are violent and have engaged in violence historically due to their religious beliefs.*
>
> This is true - but I still think you dishonestly made the quote as misleading as possible. The Mormons were at war - they formed militias, defended cities, attacked cities, all in the name of

> God. They received a fair amount of mob **violence**, and dished it out in return. They were not a peaceful people - but they might have been if they had never been harassed, we'll never know.

However, it also frequently appears with individuals accusing religion and religious people of comitting and perpetuating violent acts:

> [...] Religion is also used frequently to justify terrible things such as **violence** and oppression. The people who do this are evil with or without religion. People suck, and unfortunately religion and human nature are deeply intertwined.

Similarly, the term *rich* often occurs in condemnations of people who are wealthy but do nothing with their money, while espousing the virtues of charity; *genocide, slavery,* and *oppression* are used to accuse religious institutions of being responsible for, and culpable in, many social ills.

The more interesting group of terms is the ones that did not seem to mesh with my understanding of what r/atheism's users would tend to discuss and value. Upon examination, some words–such as *poor*–were used with a range of meanings spanning *poor people* to the adjectival *poor*. In the former sense, *poor people* are often being brought out as part of some broader argument. The argumentative context, it seems, is triggering negative polarity to land on the word.

> Sure.. everybody is free to wonder. But if it comes down to things like not giving out condoms because it is a "sin", or waging war against someone else because you think that is what God wants, or getting **poor people** to give you their money because it somehow gets them in the "good books", or teach people that they are going to go to hell because of the thoughts they are thinking .. then I think it's fair to have a problem with that.

The previously mentioned example of *freedom* shows a similar result: the most frequent contexts are proclaiming *freedom from religion*, often combatatively, and claiming that *freedom of religion* is used as a paper-thin excuse to behave poorly towards others.

The results for the rest of the selected tokens in r/atheism tell the same story, as do the tokens for the other subreddits: often, the sentiment of the given tokens seems to be derived from the broader conversational context, rather than on whether the referent of the tokens is positively or negatively valued. In r/atheism, a subreddit with many combative and argumentative posts–frequently heated–these high-weight tokens with negative sentiment seem to be better indicators of frequent topics of argument than of community values. In r/darksouls, the negative sentiment of bosses and game areas–which are considered to be among the biggest draws of the game–is mostly from people venting their frustration or debating about in-game lore.

Further, a number of tokens are akin to *poor*: where I initially read only a single referent in them (e.g., *poor people*, *the poor*, and similar NP-like uses), the polysemy in many of these words led to multiple contexts being conflated. I have chosen to leave these tokens in Table 6, as they are illustrative of the pitfalls and weaknesses of this approach that have come to light during the manual inspection. These two patterns were extremely prevalent in the post samples I manually investigated.

Also of note is that, for most tokens, most of their occurrences were mundane, and I could not read any information about community-level sentiment from their uses, even in context.

# 7 Conclusion

While at first the results of the sentiment induction seemed promising, with a number of high-weight tokens being in line with expected sentiment values and community values, on a closer inspection much of this fell apart. My initial instincts regarding both what tokens are and are not indicative of community values were frequently incorrect or misleading when examining actual posts. It is, in fact, impossible to accurately recover values and norms of the communities from the sentiments and rankings alone: frequently this results in misleading assumptions, incorrect conclusions, and a very large number of false positives–so many that the approach is almost useless for the intended purpose outlined at the beginning of this paper.

However, there were still interesting trends that were detected in the results, and which bear a closer look in future work. Many of the high-weight tokens were frequently involved in heated arguments and emotionally charged discussions. The approach here thus seems useful for identifying *topics of debate*, or perhaps *controversial topics* within the community. But these do not provide enough information to reasonably deduce community *values and norms*, as determined by comparing the high-weight, high-sentiment tokens against use cases and my own a priori knowledge of the communities.

Thus, I must unfortunately answer my original research question: automated approaches, such as the one described here, appear insufficient to determine community values and norms in the absence of extensive manual intervention and evaluation that is not feasible beyond fairly small scales. While topics of debate do seem identifiable, and while these are often related to the values I originally sought to identify, they are insufficient for my original purpose. This null result is still important, though: it was only on the manual inspection of the data, using personal knowledge of the communities in play, that I was able to both specifically identify the weaknesses of this methodology, the reasons for its failure, and some of its perhaps unexpected–but interesting–results that may serve to inform future work. Thus, the secondary aim of this paper–reinforcing the fact that automated, computational work *must* be accompanied by manual evaluation, inspection, and interpretation to achieve fully valid results–is in fact fulfilled. Had I foregone the

final inspections of original posts, randomly sampled, I would certainly have come to very different–nearly opposite–conclusions to the ones I ultimately reached.

However, this work leaves open a range of extensions and revisions. The incorporation of user-level data, e.g. accounting for the author of posts when modeling sentiment, seems one obvious way to control for potential noise, but this is also not a trivial task. Incorporating network information, such as what posts are in response to what other posts, may also refine the results either towards the original goal or towards better identification of topics of debate. Finally, refinements to the computational methodology, such as part-of-speech tagging of words (to address some of the issues with polysemy), are also promising opportunities for revisiting this work and further exploring some of the additional questions that arose during its course.

# References


Arminen, Ilkka, Christian Licoppe & Anna Spagnolli. 2016. Respecifying mediated interaction. *Research on Language and Social Interaction* 49(4). 290–309. https://doi.org/10.1080/08351813.2016.1234614. https://doi.org/10.1080/08351813.2016.1234614.

Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec & Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of world wide web conference*, vol. 22, 307–318. Association for Computational machinery.

Firth, J. R. 1958. A Synopsis of Linguistic Theory, 1930-1955. English. In, 1–32.

Hamilton, William L., Kevin Clark, Jure Leskovec & Dan Jurafsky. 2016. Inducing domain-specific sentiment lexicons from unlabeled corpora. English. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, 595–605. Austin, TX: Association for Computational Linguistics.

Harris, Zellig S. 1954. Distributional structure. *Word* 10(2-3). 146–162.

Loughran, Tim & Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of Finance* 66(1). 35–65.

Mikolov, Tomas, Kai Chen, Greg Corrado & Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Pennebaker, J.W., R.J. Booth, R.L. Boyd & M.E. Francis. 2015. *Linguistic inquiry and word count*. Austin, TX: Pennebaker Conglomerates.

Pennington, Jeffrey, Richard Socher & Christopher Manning. 2014. Glove: global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)*, 1532–1543.

Řehůřek, Radim & Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. English. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50. Valletta, Malta: ELRA.

Rubenstein, Herbert & John B. Goodenough. 1965. Contextual correlates of synonymy. *Commun. ACM* 8(10). 627–633. https://doi.org/10.1145/365628.365657. http://doi.acm.org/10.1145/365628.365657.

Sahlgren, Magnus. 2008. The distributional hypothesis. *Italian Journal of Disability Studies* 20. 33–53.

Schwartz, H. Andrew, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinskiand David Stillwell, Martin E. P. Seligman & Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: the open-vocabulary approach.

*Plos One* 8(9). e73791. https://doi.org/10.1371/journal.pone.0073791. https://doi.org/10.1371/journal.pone.0073791.

Vandergriff, Ilona. 2013. Emotive communication online: a contextual analysis of computer-mediated communication (cmc) cues. *Journal of Pragmatics* 51. 1–12.

Zhang, Justine, William L. Hamilton, Cristian Danescu-Niculescu-Mizil, Dan Jurafsky & Jure Leskovec. 2017. Community identity and user engagement in a multi-community landscape. In, vol. 11, 337–346. Association for the Advancement of Artificial Intelligence.

# A Top Weighted Tokens by Subreddit

Table 7: Top 20 tokens in the r/4chan subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| yes | 2.8353 | 25.9109 |
| thank | 2.8312 | 22.9517 |
| gt | -2.0398 | 22.8858 |
| account | 2.7890 | 21.6227 |
| amazing | 2.7505 | 21.3359 |
| sorry | 2.4363 | 20.9367 |
| truly | 2.7904 | 20.9298 |
| oh | 2.2356 | 20.7272 |
| funny | 2.1338 | 20.3549 |
| fucking | 1.9009 | 20.3438 |
| hope | 2.3590 | 20.2567 |
| good | 1.9838 | 20.2437 |
| money | -2.2342 | 20.0184 |
| beautiful | 2.6711 | 19.8433 |
| fun | 2.2721 | 19.5687 |
| good_job | 2.8093 | 19.5514 |
| genius | 2.8180 | 19.2830 |
| wish | 2.3486 | 19.1284 |
| freedom | 2.7186 | 19.1067 |
| muh | 2.8694 | 19.0298 |

Table 8: Top 20 tokens in the r/2007scape subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| noob | -2.9224 | 23.0142 |
| poor | -2.9253 | 22.3960 |
| giving | -2.6577 | 22.3183 |
| kid | -2.7942 | 21.9447 |
| bad | -2.2785 | 21.7925 |
| shitty | -2.7614 | 21.7028 |
| small | -2.5151 | 21.6702 |
| new_players | -2.8428 | 21.6359 |
| lack | -2.9168 | 21.5087 |
| huge | -2.5357 | 21.4673 |
| calling | -2.8890 | 21.3214 |
| logic | -2.7430 | 21.2195 |
| free | -2.3283 | 21.2064 |
| argument | -2.6440 | 21.1127 |
| bullshit | -2.7177 | 20.7494 |
| terms | -2.7642 | 20.6818 |
| noobs | -2.9157 | 20.6408 |
| horrible | -2.9101 | 20.5594 |
| sad | -2.5437 | 20.4119 |
| absolutely | -2.5006 | 20.3982 |

Table 9: Top 20 tokens in the r/atheism subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| poor | -2.5699 | 29.0158 |
| rights | -2.5058 | 28.7300 |
| worse | -2.5104 | 28.5265 |
| actions | -2.4984 | 28.3350 |
| act | -2.4303 | 28.2110 |
| horrible | -2.5624 | 27.7635 |
| bad | -2.1701 | 27.7222 |
| justify | -2.5405 | 27.5474 |
| worst | -2.5276 | 27.4786 |
| excuse | -2.5426 | 27.4007 |
| terrible | -2.5109 | 27.3824 |
| violence | -2.5271 | 27.3808 |
| blame | -2.5437 | 27.2619 |
| fucked | -2.4994 | 26.7905 |
| situation | -2.3348 | 26.7892 |
| action | -2.4765 | 26.7629 |
| responsible | -2.5301 | 26.7539 |
| making | -2.1289 | 26.5559 |
| potential | -2.5360 | 26.3993 |
| aren | -2.1227 | 26.3976 |

Table 10: Top 20 tokens in the r/darksouls subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| hate | -2.7022 | 24.7787 |
| annoying | -2.7652 | 24.2943 |
| poor | -2.9664 | 24.0108 |
| worst | -2.8095 | 23.6353 |
| pain | -2.9352 | 23.4862 |
| sucks | -2.7605 | 23.0452 |
| horrible | -2.9773 | 22.9413 |
| stupid | -2.6185 | 22.8909 |
| face | -2.5070 | 22.8800 |
| bad | -2.2835 | 22.8432 |
| fair | -2.4996 | 22.7548 |
| terrible | -2.6849 | 22.6707 |
| bullshit | -2.7148 | 21.6689 |
| internet | -2.6511 | 21.6644 |
| awful | -2.8301 | 21.5477 |
| chosen_undead | -2.6511 | 21.5430 |
| worse | -2.5282 | 21.4756 |
| cheap | -2.4901 | 21.4603 |
| connection | -2.6509 | 21.3090 |
| useless | -2.5461 | 21.1824 |

Table 11: Top 20 tokens in the r/linux subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| linux | 2.8229 | 34.2249 |
| like | 2.8167 | 34.0668 |
| use | 2.7869 | 33.6685 |
| people | 2.8150 | 32.4949 |
| good | 2.8646 | 32.1557 |
| time | 2.8164 | 31.9378 |
| way | 2.8162 | 31.8598 |
| want | 2.7771 | 31.8106 |
| know | 2.7733 | 31.6708 |
| windows | 2.7673 | 31.6178 |
| better | 2.8628 | 31.5822 |
| things | 2.8367 | 31.5124 |
| think | 2.7031 | 31.2648 |
| lot | 2.8607 | 31.1786 |
| new | 2.8061 | 30.5957 |
| ve | 2.7403 | 30.5233 |
| work | 2.6360 | 30.2493 |
| desktop | 2.8106 | 30.0704 |
| os | 2.8095 | 30.0179 |
| look | 2.8151 | 29.7927 |

Table 12: Top 20 tokens in the r/pcmasterrace subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| bad | -2.4820 | 28.6884 |
| shitty | -2.7391 | 28.1951 |
| power | -2.6284 | 28.1141 |
| worse | -2.7081 | 27.3939 |
| terrible | -2.7378 | 27.0821 |
| psu | -2.5777 | 27.0639 |
| poor | -2.7606 | 26.7350 |
| isn | -2.2803 | 26.4619 |
| sucks | -2.7436 | 26.3494 |
| crap | -2.6899 | 26.2731 |
| worst | -2.7344 | 25.9771 |
| horrible | -2.7684 | 25.8255 |
| fan | -2.5315 | 25.7572 |
| fans | -2.5820 | 25.7436 |
| actually | -2.1309 | 25.5011 |
| mean | -2.2635 | 25.4297 |
| suck | -2.7205 | 25.1095 |
| fucked | -2.7256 | 25.0225 |
| awful | -2.7570 | 24.9400 |
| mess | -2.7579 | 24.7562 |

Table 13: Top 20 tokens in the r/politics subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| poor | -2.6584 | 33.8246 |
| rich | -2.6584 | 33.4135 |
| wealth | -2.6583 | 30.9095 |
| wealthy | -2.6583 | 30.7385 |
| middle_class | -2.6583 | 30.6472 |
| violence | -2.6157 | 30.2394 |
| bad | -2.2756 | 30.1272 |
| poverty | -2.6578 | 29.7310 |
| killed | -2.5239 | 29.6294 |
| worst | -2.5611 | 29.5460 |
| horrible | -2.6762 | 29.5334 |
| excuse | -2.6189 | 29.2130 |
| terrible | -2.5673 | 29.2066 |
| killing | -2.4949 | 29.0108 |
| kill | -2.3877 | 28.9641 |
| dead | -2.4982 | 28.9088 |
| innocent | -2.5966 | 28.8104 |
| shot | -2.4674 | 28.6000 |
| justify | -2.6005 | 28.5863 |
| deaths | -2.6811 | 28.4208 |

Table 14: Top 20 tokens in the r/runescape subreddit, sorted by weight.

| Token | Sentiment | Weight |
|---|---|---|
| boss | -2.9986 | 27.9767 |
| bosses | -3.0043 | 27.4845 |
| bossing | -3.0055 | 26.3032 |
| gwd | -2.9992 | 25.7965 |
| poor | -3.0825 | 25.0554 |
| nex | -2.8472 | 24.8222 |
| pvm | -2.7767 | 24.7095 |
| real | -2.6907 | 24.6330 |
| lazy | -3.0730 | 24.6145 |
| type | -2.7773 | 24.4053 |
| bother | -3.0094 | 24.3962 |
| tl_dr | -3.0477 | 24.2142 |
| rich | -3.0767 | 24.0086 |
| solo | -2.7837 | 23.8576 |
| shitty | -2.9779 | 23.7382 |
| live | -2.8805 | 23.6269 |
| lack | -3.0020 | 23.6255 |
| kk | -2.8569 | 23.5669 |
| crap | -2.9811 | 23.4976 |
| legit | -2.9348 | 23.3709 |

# B Sentiment-Document Frequency Plots

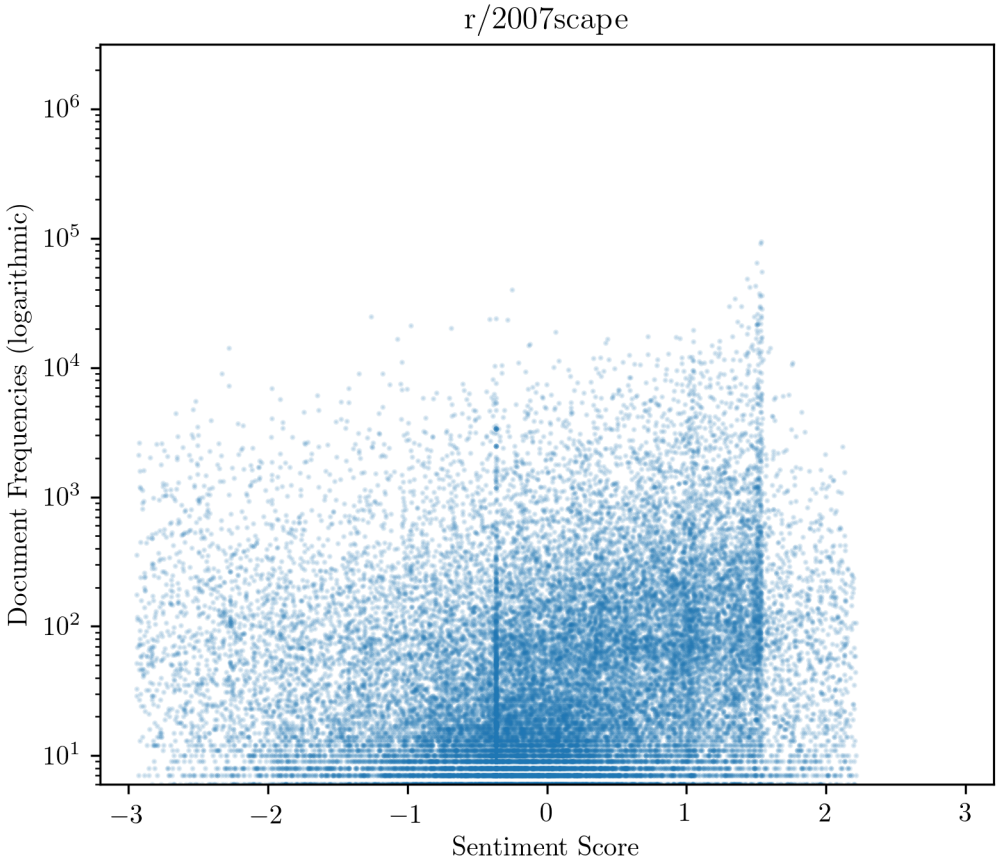


Figure 3: Plot of sentiment scores against document frequencies for the r/2007scape subreddit.

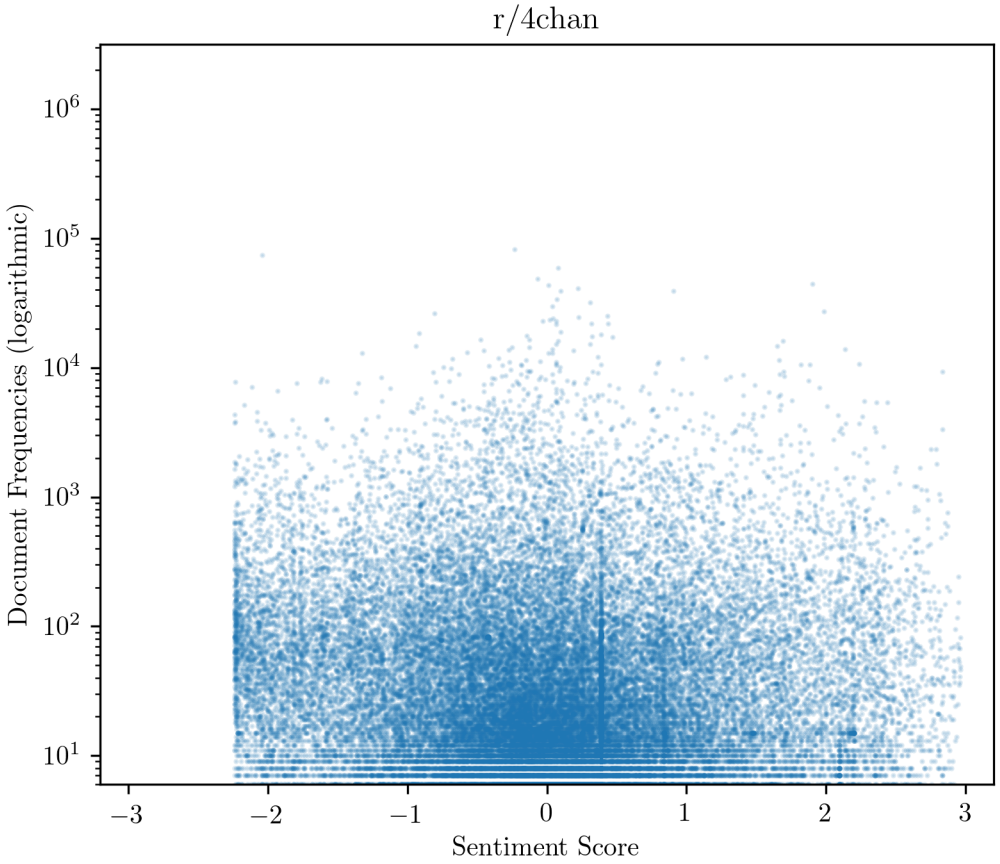


Figure 4: Plot of sentiment scores against document frequencies for the r/4chan subreddit.

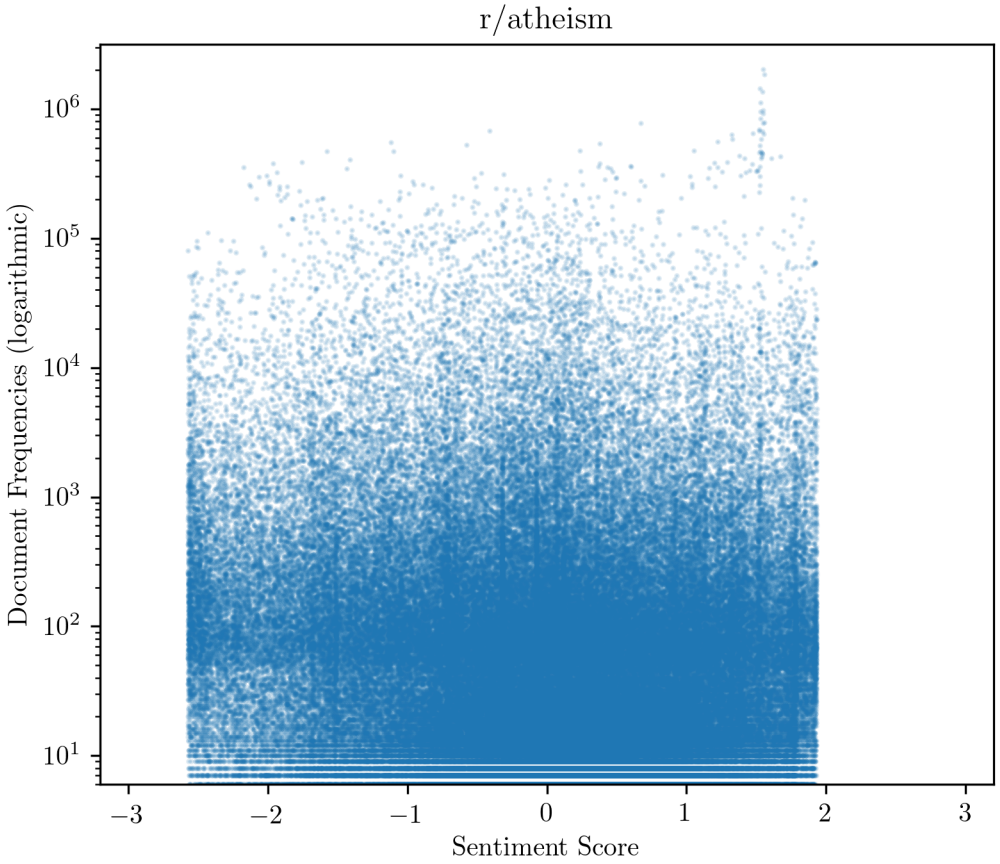


Figure 5: Plot of sentiment scores against document frequencies for the r/atheism subreddit.

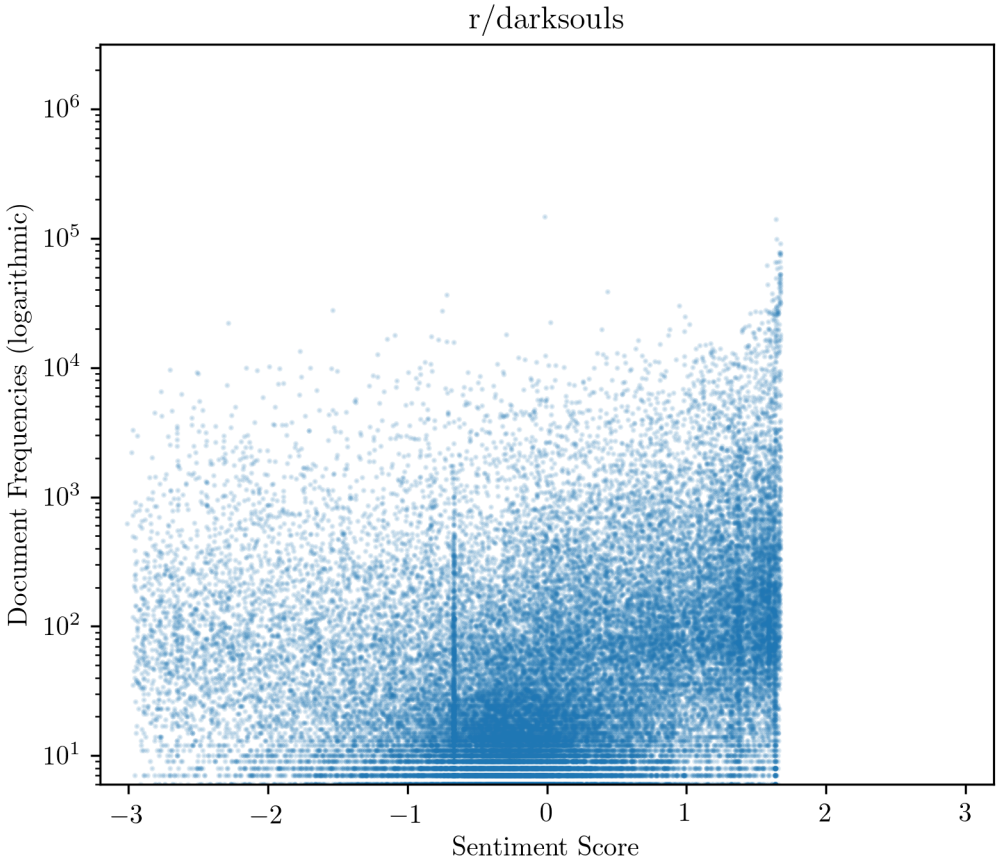


Figure 6: Plot of sentiment scores against document frequencies for the r/darksouls subreddit.

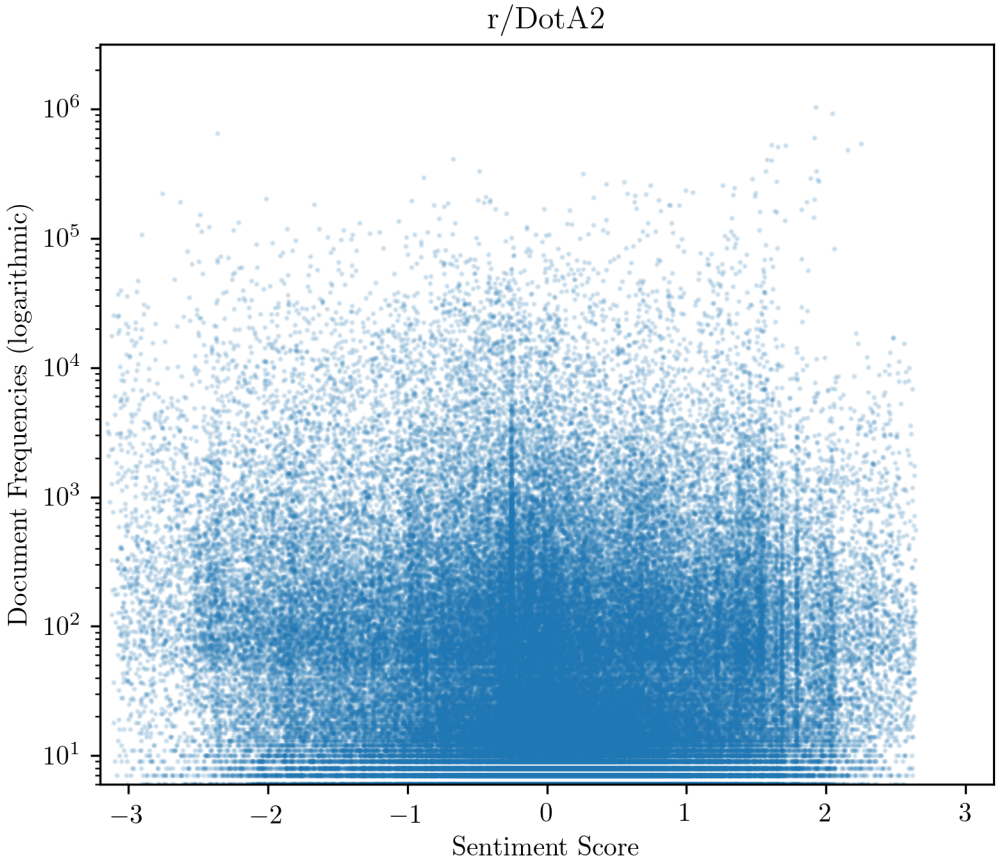


Figure 7: Plot of sentiment scores against document frequencies for the r/DotA2 subreddit.

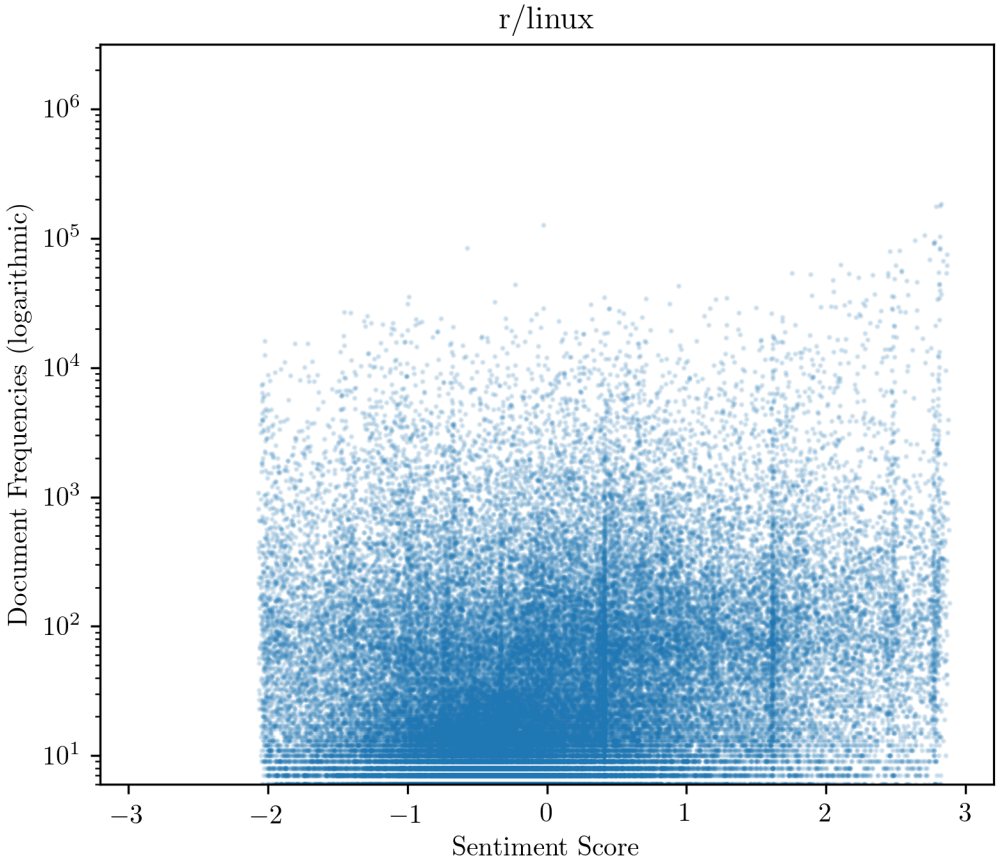


Figure 8: Plot of sentiment scores against document frequencies for the r/linux subreddit.

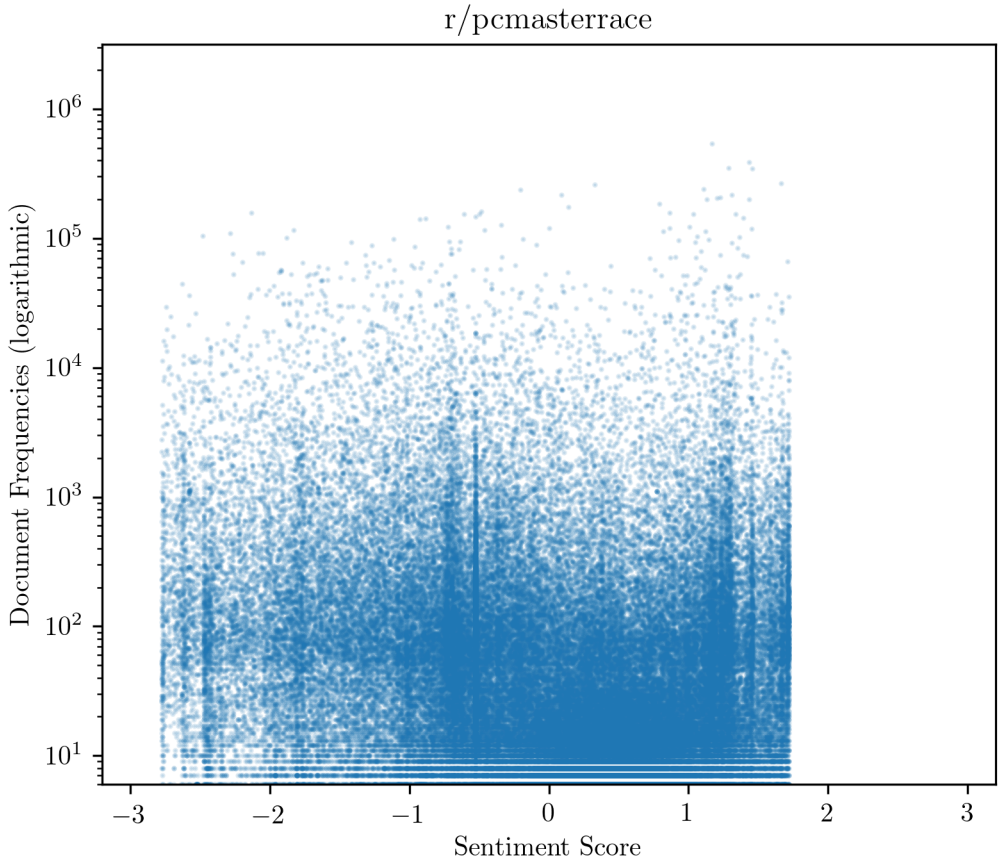


Figure 9: Plot of sentiment scores against document frequencies for the r/pcmasterrace subreddit.

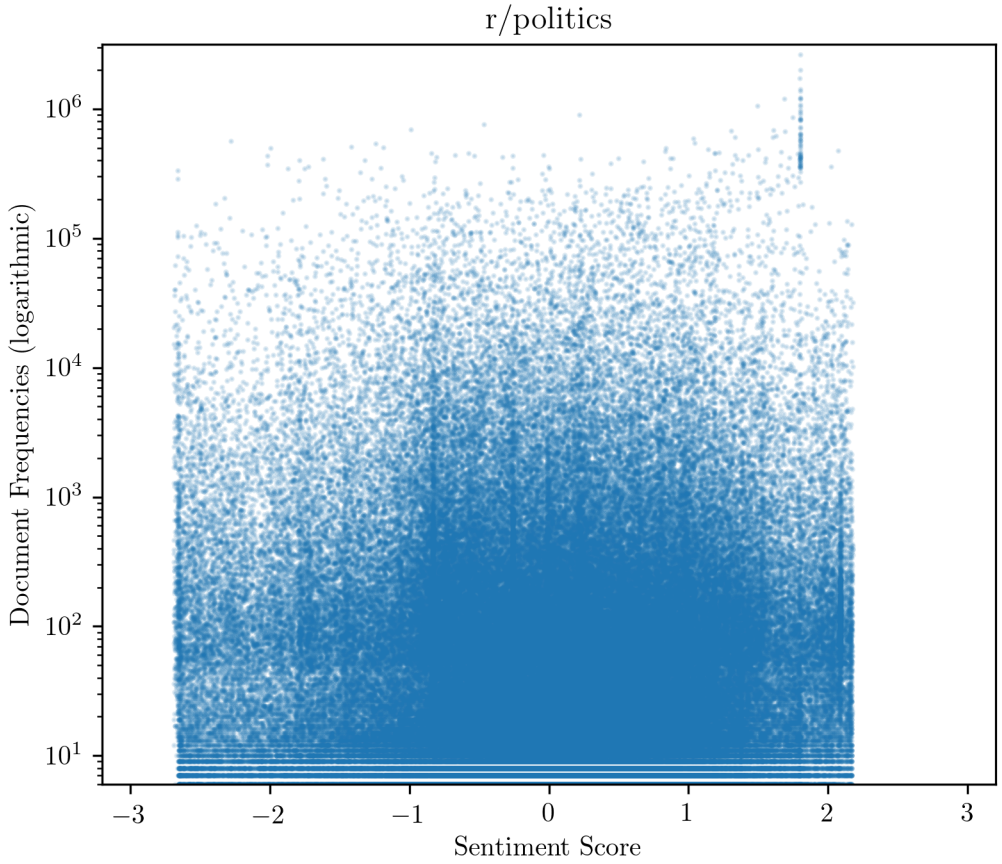


Figure 10: Plot of sentiment scores against document frequencies for the r/politics subreddit.

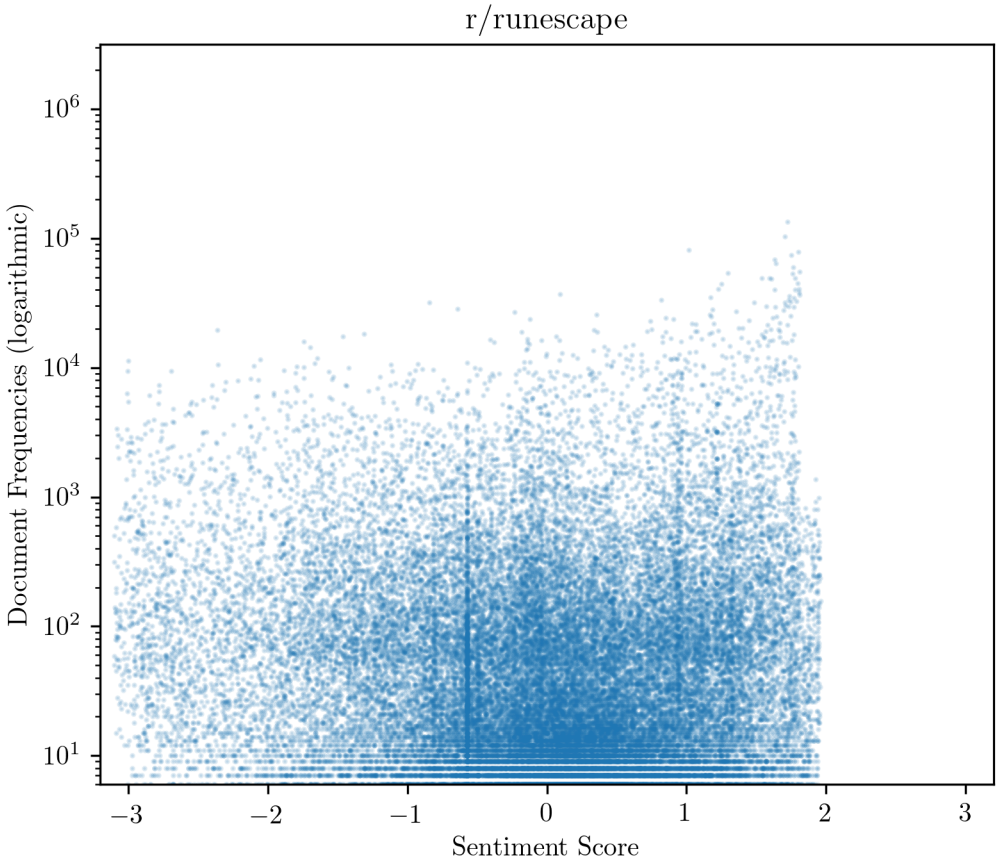


Figure 11: Plot of sentiment scores against document frequencies for the r/runescape subreddit.