# Ranking Subreddits by Classifier Indistinguishability in the Reddit Corpus

Faisal Alquaddoomi

UCLA Computer Science Dept.
Los Angeles, CA, USA
Email: faisal@cs.ucla.edu

Deborah Estrin

Cornell Tech
New York, NY, USA
Email: destrin@cornell.edu

***Abstract*—Reddit, a popular online forum, provides a wealth of content for behavioral science researchers to analyze. These data are spread across various subreddits, subforums dedicated to specific topics. Social support subreddits are common, and users' behaviors there differ from reddit at large; most significantly, users often use 'throwaway' single-use accounts to disclose especially sensitive information. This work focuses specifically on identifying depression-relevant posts and, consequently, subreddits, by relying only on posting content. We employ posts to r/depression as labeled examples of depression-relevant posts and train a classifier to discriminate posts like them from posts randomly selected from the rest of the Reddit corpus, achieving 90% accuracy at this task. We argue that this high accuracy implies that the classifier is descriptive of "depression-like" posts, and use its ability (or lack thereof) to distinguish posts from other subreddits as discriminating the "distance" between r/depression and those subreddits. To test this approach, we performed a pairwise comparison of classifier performance between r/depression and 229 candidate subreddits. Subreddits which were very closely related thematically to r/depression, such as r/SuicideWatch, r/offmychest, and r/anxiety, were the most difficult to distinguish. A comparison this ranking of similar subreddits to r/depression to existing methods (some of which require extra data, such as user posting co-occurrence across multiple subreddits) yields similar results. Aside from the benefit of relying only on posting content, our method yields per-word importance values (heavily weighing words such as "I", "me", and "myself"), which recapitulate previous research on the linguistic phenomena that accompany mental health self-disclosure.**

***Keywords–Natural language processing; Web mining; Clustering methods***

## I. Introduction

Reddit, a popular link-sharing and discussion forum, is a large and often difficult-to-navigate source of computer-mediated communication. Like most public discussion forums, it is distinct from other social networking sites such as Facebook or Twitter in that conversation largely occurs with strangers rather than members of one's explicit social graph (friends and followers, respectively). Unlike small topical discussion forums, Reddit is a vast collection of topical subforums (also known as "subreddits"), numbering just over one million as of January 2017 [1].

While Reddit is primarily a link-sharing website where users collaboratively filter content by voting, there is a significant portion of the site which is more social in nature. Many support subreddits exist in which the majority of posts are "self-posts", text written by users, rather than links to images or articles. A particularly interesting subset of these subreddits are the support and self-help subreddits where individuals spontaneously request and provide support to relative strangers. The ease of creating 'throwaway' accounts has encouraged the development of self-help subreddits where individuals can discuss possibly stigmatized medical conditions in relative anonymity. It has been shown that individuals who are anonymous tend to be less inhibited in their disclosures, and that Reddit users specifically make use of this feature when soliciting help (in the form of a self-post) more so than when providing it (in the form of a reply) [2].

The frankness and public accessibility of this communication makes it an attractive target for behavioral research, but as mentioned it can be difficult to navigate the vast number of subreddits, especially as existing ones change and new ones are introduced over time. It is infeasible to make use of user posting co-occurrence (a common and successful tactic for clustering subreddits) to study this subset of Reddit since users often do not maintain persistent accounts. This work presents a content-based subreddit ranking in which subreddits are ranked by the difficulty of distinguishing their posts from a "baseline" subreddit. We focus specifically on r/depression as the baseline subreddit in this work, since it is readily differentiable from average Reddit posts, as demonstrated in section V. We explore the task of finding subreddits that are similar to r/depression based on this initial strength and compare our ranking results with other content- and user-based subreddit similarity measures. As an added benefit, our method provides weightings on the feature (in this case, words) that differentiate two subreddits, making the model's decisions more interpretable as a result.

The remainder of the paper is structured as follows. Section II provides a brief discussion of two fields which intersect in our work: mental health disclosure in social media, and clustering of forums by user and post attributes. We discuss the specific dataset, the Reddit corpus, in section III. Section IV describes the methods we used to cluster subreddits, and section V presents the results of our method and comparisons to others. Section VI discusses these results, with some high-level observations about the differences between prior results and potential problems with our current methodology. Finally, section VII recapitulates the problem of clustering subreddits by post content, how we approached that problem, and what is left to do.

## II. Related Work

Since this work involves two separate topics, behavioral health as evidenced in online communities and subreddit

clustering, they are presented below in two distinct sections.

### A. Mental Health and Social Media

[2] examined the role of anonymity and how it affects disclosure in individuals seeking mental health support on reddit. They also automatically classified responses to these requests for help into four categories. While not directly relevant to the task of ranking subreddits by similarity, the context in which their study was conducted inspired this work, specifically in focusing on self-help subreddits in which individuals generally disclose anonymously. Their identification of the disparity between anonymous posters who are seeking help and often non-anonymous commenters providing aid influenced the decision to consider only self-post text in this work, as that is apparently more emblematic of people suffering from mental illness rather than individuals trying to help them. The ad-hoc process that they describe for collecting sets of related subreddits (a combination of knowledge from seasoned redditors and reading the information panel of their initial subreddits) motivated the need for an automatic method to find subreddits that requires only a "seed" subreddit from which to identify linguistically similar content. We hope that this work presents a possible solution in this context.

### B. Clustering Subreddits

As far as we can tell, there has been little academic investigation into the problem of clustering subreddits. Instead, a number of individuals have informally explored the problem in blog posts and postings to Reddit itself. Their approaches fall into two groups: 1) user-based, and 2) content-based.

User-based methods focus on the users as the evidence linking subreddits. [3] computed a set of active users for each subreddit and used the Jaccard coefficient (the intersection of the users in common between two subreddits divided by their union) as a similarity score. [4], whose results we compare to our own in section V, constructed a matrix of (normalized) user posting counts to subreddits, using the counts over all users posting to a subreddit as that subreddit's vector representation. Like the previous two approaches, [5], in an academic paper, also treated the same user posting a set of subreddits as evidence of their relatedness. They first built a graph weighted by this posting co-occurrence, then used a "backbone extraction" algorithm to eliminate edges that could be attributed to random chance.

Content-based methods focus on the text of comments and, to a lesser extent, posts to correlate subreddits. [6] used the top 100 words in the comments across 50 top subreddits (by commenting activity) to construct a (normalized) bag-of-words feature representation of each subreddit. They computed similarity by taking the Euclidean distance of all pairwise combinations of these subreddits, and performed clustering using affinity propogation. A second content-based method, [7], made use of term-frequency inverse-document-frequency (TF-IDF) and latent semantic indexing (with dimensions set to 2) on over 20 million comments to produce a plot of subreddits in a space where distance reflected their textual similarity.

## III. Data

The dataset consists of posts from Reddit, a popular online forum. Reddit posts, unlike Twitter, are not length-constrained, and unlike Facebook are typically public but not necessarily identifying. Redditors (Reddit users) overwhelmingly prefer pseudonyms, and the site allows one to easily create throwaway accounts for one-off sensitive posts, something that is difficult to do on other services. This combination of public, lengthy, and often sensitive posts is a good source of data for studying the language with which individuals candidly express their symptoms or other circumstances surrounding their illnesses. (Despite being a publicly-available dataset, we acknowledge the sensitivity of these disclosures; none of the results or other data included in this work identify the individuals by name.)

The dataset was obtained from a public database of Reddit posts hosted on Google's BigQuery service [8]. Posts from 12-01-2015 to 7-31-2016 were considered in this analysis, although the corpus has been regularly updated since then.

### A. Reddit Description

Reddit is made up of a large number of user-created special-interest fora, called subreddits, on which individuals post either links to content (images, news articles, etc. that are stored off-site) or self-posts, which typically consist of text entered by the poster. Subreddits are prefixed by an r/ in reference to their URL on the site, e.g., r/politics for https://reddit.com/r/politics. Each post on a subreddit is accompanied by a threaded comments section in which users can discuss the posted content.

Topics for subreddits include general interests, such as gaming or politics, or more specific interests such as particular television shows. Subreddits vary wildly in scale and activity, with some having thousands of subscribers and near-constant activity and others having been largely abandoned. Of particular relevance to this research are the social support/self-help subreddits, such as the ones around the management of chronic illnesses. This research in particular uses r/depression as a source of depression-relevant posts, although the method could be extended to other subreddits with a sufficient quantity of selfposts.

Content on the site is regulated through a community-driven mechanism of upvoting (or downvoting) both posts and comments on the site. Each user is able to provide one upvote or downvote for a particular element, and the aggregation of these votes (as well as other factors, such as age of the post or commenting activity) determines the order in which content is displayed, and thus its visibility. Elements that have a sufficiently negative score will be hidden by default, further reducing their visibility.

## IV. Methods

Our objective is to differentiate depression-relevant posts – posts which are specifically about depression - from non-depression-relevant posts. Note that this is a separate task from identifying posts that were written by a depressed person, since they could write about many topics without a necessarily detectable influence on their writing. The general strategy was to start with a simple approach, then gradually work up to more complicated approaches should the simpler ones not provide sufficient accuracy. There are three high-level tasks that we addressed:

1) Discriminating a post from r/depression from a post selected from the entire corpus at random.
2) Determining if there are other subreddits which are measurably similar to r/depression based on the inability of the classifier to distinguish them

3) Identifying what features were most significant in the discrimination.

Tasks 1 and 2 can be performed with any binary classifier, but task 3 requires a classifier that assigns importance values to the features.

For task 1, 10,000 self-posts were uniformly selected from r/depression and 10,000 were uniformly selected from the corpus at large (potentially including posts from r/depression, although r/depression makes up a very small proportion of the total posts in the corpus.) Each post was labeled as originating from r/depression or not, and the sets were concatenated into a total dataset consisting of 20,000 labeled posts. These 20,000 posts were split into a 60% training, 40% test sets consisting of 12,000 training posts and 8,000 test posts. The classifier was trained using the training set, then validated by attempting to predict the labels of the posts in the test set.

For task 2, subreddits were selected that had a sufficient number of self-posts ($\geq 5000$), which resulted in 229 candidate subreddits. 5,000 posts were selected uniformly from each candidate, and 5,000 posts were again selected uniformly from r/depression. The combined dataset of 10,000 labeled posts was constructed for each pairing of the 5,000 r/depression posts with the 5,000 posts from each candidate subreddit. The dataset was again split into training and test (6000 training, 4000 test) and the same process as described in task 1 was carried out for each pairing.

### A. Sample to Feature-Vector Encoding

Most classifiers cannot directly accept samples, in this case a series of characters of arbitrary length, as input. Instead, the samples must be reduced into a set of features before use. Each post was encoded into a feature vector, a fixed-sized set of word counts, prior to being input into the classifier. To construct this feature vector, the entire training corpus was converted to lowercase and all punctuation except apostrophes were converted into spaces. The text was split on the spaces to produce tokens. The counts of each token were summed, then the 5000 most frequent tokens over the full set of posts (that is, including both r/depression and the other set of posts) were chosen as the elements of the feature vector.

Each post was then subjected to a similar tokenization and counting process, creating a 5000-element feature vector per post. Words that were present in the post but not in the feature encoding were ignored, and words which were not present in the post were given a count of 0. These per-post word counts were then scaled using TF-IDF, which in this case was the occurrence of the word within each post divided by the number of times it occurred within the full set of posts. No stemming or other collapsing of the token space was performed, with the intent being to capture idiosyncrasies in word choice.

Scikit-learn [9] was used to perform the above steps, specifically the `CountVectorizer`, `TfIdfTransformer`, and `Pipeline` classes.

### B. Classification

We initially chose a naïve Bayes classifier as the simplest classifier to test the method. A naïve Bayes classifier considers each feature as an independent and identically distributed random variable and performs a binary classification on each sample into one of two possible classes (in this case, depression-relevant vs. not). After analyzing the performance on this classifier on the validation set, we moved on to a random forest classifier, which has many similarities to naïve Bayes, but also provides the importance values needed for task 3. (While feature importances can be derived from naïve Bayes' classifiers, according to [10] it is a good classifier, but poor estimator, so the importance values are apparently not robust.) A random forest classifier is an ensemble method which averages the performance of many decision tree classifiers to produce a more robust final estimate. Decision trees, as the name suggests, construct a tree of Boolean predicates on a feature (e.g., "feature #6 $< 563$"), with the leaves of the tree consisting of the final classification for a sample that satisfies each Boolean predicate. The random forest constructs many of these trees on subsets of the training data, then averages them to circumvent the tendency for a single decision tree to overfit to the training data.

### C. Comparison Methods

In the absence of a gold standard for subreddit clustering, we compare the rankings produced by our approach against several methods, described in detail in the following. The first two methods use the same feature representation for posts as described above, specifically 5000-element TF-IDF-scaled word counts. The last method's results were procured through the project's API by querying for subreddits related to 'depression'. We refer to the 5,000-post sample from r/depression as the **baseline set**, and each subreddit against which we are comparing r/depression as the **candidate set**.

*1) Averaged TF-IDF Cosine Similarity:* Cosine similarity is a popular choice in the field of information retrieval for determining the similarity of strings based on the angle between their feature representations [11]. In this case, we first compute a "subreddit vector" from its constituent posts in the sample, then determine the similarity of two subreddits by their angle. Specifically, for subreddit vectors A and B, the cosine similarity is defined as follows:

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} \tag{1}$$

Since our vectors all have positive components, the cosine similarity ranges from 1 (identical) to 0. The subreddit vectors are obtained by averaging the feature representations of each post in the baseline or candidate sample, respectively. We simply compute the cosine similarity between the baseline set's vector and each candidate set's vector to produce the final set of similarities, then order by descending similarity to produce the rankings.

*2) Topic Vector Similarity:* Prior to performing the similarity analysis, this approach first computes a 50-topic **topic model** over a co-occurrence matrix of the feature vectors for each post in the baseline set, performed using the software package gensim [12]. Specifically, we used a technique known as Latent Dirichlet Allocation (LDA) to produce a lower-dimensional 'topic' representation of the matrix. We apply this topic model of r/depression to transform each of the comparison subreddits' feature vectors into this lower-dimensional topic space. We employ gensim's `similarities.MatrixSimilarity` class to construct a data structure for efficiently comparing an input post's topic vector to every post in the baseline set. The comparison is performed via cosine similarity, but this time between the topic

vector of the input post and the topic vectors of each post in the baseline set.

The topic model is then applied to each feature vector from the candidate set, producing a topic vector, then the similarity of every topic vector from the candidate post is compared to the topic vector of every post from the baseline set. The results of all of these comparisons are averaged, producing an average similarity score for the baseline-candidate pairing. The remainder of this method is the same as cosine similarity: the similarities for each candidate subreddit are ordered to produce a final ranking.

*3) User-Centric Similarity:* We did not directly implement this method; instead, we utilized the project's website to issue a query for posts similar to r/depression and downloaded the result. As described in its accompanying blog post [4], this method first constructs a user-subreddit matrix consisting of times in which each user has posted in each subreddit. The user list was drawn from participants in 2,000 "representative" subreddits and compared against 47,494 subreddits. These counts are adjusted by computing the positive pointwise mutual information for each. In this case, the subreddit vectors are the user-count vectors for each subreddit; similarity is once again computed as the cosine similarity between the subreddit vectors.

Note that this method's returned subreddits do not completely overlap with the 229 candidate subreddits of the other methods, since they were drawn from 47,494 subreddits instead.

## V. Results

Surprisingly, the naïve Bayes classifier performed extremely well on task 1. With no hyper-parameter tuning we achieved 89.9% accuracy on the test set. The random forest classifier achieved similar performance (89.1% accuracy.) As mentioned previously, we opted for the random forest classifier since we had reason to distrust the feature importances from naïve Bayes.

### A. Classifier Performance

Figure 1 depicts the receiver operating characteristic (ROC) curve for the random forest classifier, which shows the proportion of true to false positives as the decision threshold of the classifier is varied. The confusion matrix in figure 2 demonstrates a relative scarcity of false-positive and false-negative errors compared to correct classifications in the test set.

To determine the feasibility of separating depression-relevant from non- posts, we also performed a principal component analysis (PCA) on the feature vectors of the samples in the test set. This was followed by a t-distributed stochastic neighbor embedding (t-SNE) of the first 50 principal components (derived from the 10,000 depressed vs. not set) to visualize the distribution of sample points in two dimensions, shown in figure 3. Teal points are from the depression set, blue points are randomly selected from Reddit at large. The figure reveals distinct clusters of depression-relevant versus non-depression-relevant posts, which supports the argument that the classification task is inherently feasible.

The scattering of non-depressed points through a section of the depressed cluster could be due to those points being erroneously classified as non-depressed. For instance, they may belong to r/SuicideWatch or other such subreddits which are shown in task 2 to be difficult to distinguish from r/depression.

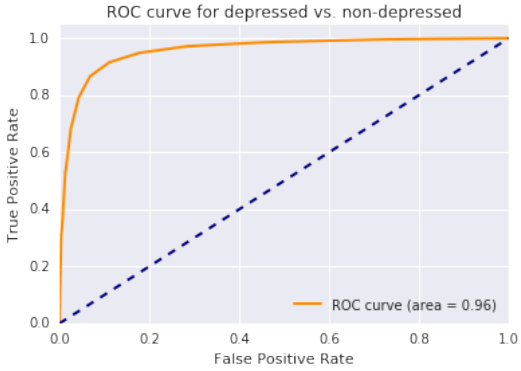


FIGURE 1. ROC CURVE DISPLAYING THE PERFORMANCE OF THE RANDOM FOREST CLASSIFIER IN DIFFERENTIATING POSTS FROM R/DEPRESSION FROM RANDOMLY-SELECTED REDDIT POSTS.

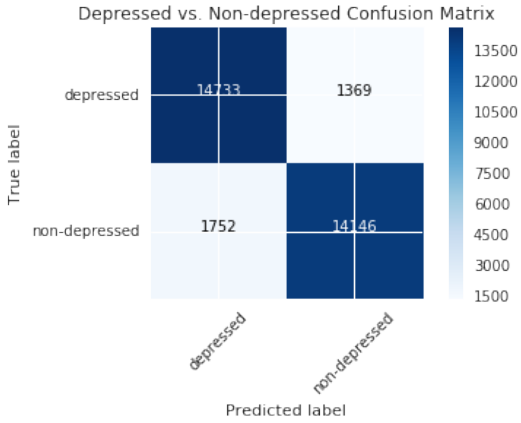


FIGURE 2. THE CONFUSION MATRIX IN CLASSIFYING R/DEPRESSION POSTS VERSUS POSTS RANDOMLY SELECTED FROM REDDIT.

### B. Pairwise Comparisons

The performance of the classifier in task 1 could potentially be explained by the prevalence of easily-differentiated non-depression-relevant posts in the Reddit corpus. To test the hypothesis that some text is easier to differentiate from r/depression posts than others, we constructed a candidate set of 229 sufficiently popular subreddits with over 5,000 posts. We repeated the analysis in task 1 for each candidate, using the accuracy of the classifier to determine the similarity of that subreddit to r/depression. Table I shows an excerpt of the top 20 subreddits ranked by difficulty of discriminating them from r/depression. The accuracy column, by which the list is sorted, is the proportion of posts which were successfully classified as their true subreddit.

The least-distinguishable subreddits (r/SuicideWatch, r/offmychest, r/advice, r/Anxiety) are all within the support/self-help community of subreddits that relate specifically to depression and anxiety. This supports the hypothesis that the classifier has learned which posts are more likely to mention depression.

*1) Alternative Rankings:* In the absence of a gold standard for subreddit clustering, we compare the rankings produced by our approach against several standard and popularly-available methods. Tables II, III, and IV show rankings for the cosine

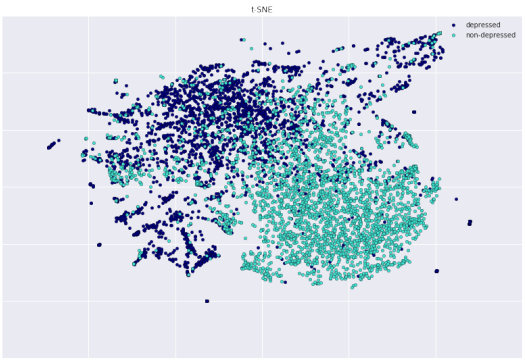


FIGURE 3. T-SNE 2-DIMENSIONAL PLOT OF THE FIRST 50 PRINCIPAL COMPONENTS.

TABLE I. TOP 20 SUBREDDITS THE RANDOM FOREST METHOD FOUND SIMILAR TO R/DEPRESSION.

| accuracy | subreddit |
|---|---|
| 0.628 | SuicideWatch |
| 0.703 | offmychest |
| 0.76825 | Advice |
| 0.7705 | Anxiety |
| 0.818 | teenagers |
| 0.8365 | CasualConversation |
| 0.84675 | raisedbynarcissists |
| 0.855 | askgaybros |
| 0.864 | asktrp |
| 0.871 | asktransgender |
| 0.8795 | opiates |
| 0.881 | trees |
| 0.88125 | relationship_advice |
| 0.88725 | NoFap |
| 0.888 | NoStupidQuestions |
| 0.8945 | breakingmom |
| 0.899 | BabyBumps |
| 0.901 | Drugs |
| 0.903 | Christianity |
| 0.90375 | sex |

similarity method, the LDA topic-vector method, and the user-centric method, respectively. For each of these tables, the distance column lists $1.0 -$ cosine_similarity to provide a consistent sorting order with table I.

In order to more rigorously compare these rankings to our method, we computed the Spearman's Rho [13] and Kendall's

TABLE II. TOP 20 SIMILAR SUBREDDIT RANKING FOR THE COSINE SIMILARITY METHOD.

| distance | subreddit |
|---|---|
| 0.008156 | SuicideWatch |
| 0.026798 | Anxiety |
| 0.028122 | offmychest |
| 0.038478 | Advice |
| 0.049564 | asktransgender |
| 0.056973 | stopdrinking |
| 0.060631 | teenagers |
| 0.062695 | NoFap |
| 0.070161 | raisedbynarcissists |
| 0.074363 | opiates |
| 0.077625 | CasualConversation |
| 0.078701 | BabyBumps |
| 0.078729 | askgaybros |
| 0.079949 | Drugs |
| 0.081216 | asktrp |
| 0.087126 | sex |
| 0.09335 | trees |
| 0.094424 | loseit |
| 0.096255 | breakingmom |
| 0.099262 | relationships |

TABLE III. TOP 20 SIMILAR SUBREDDIT RANKING FOR THE LDA TOPIC-VECTOR METHOD.

| distance | subreddit |
|---|---|
| 0.077287 | raisedbynarcissists |
| 0.077868 | relationships |
| 0.078384 | offmychest |
| 0.082861 | SuicideWatch |
| 0.089728 | Anxiety |
| 0.089788 | Advice |
| 0.09074 | tifu |
| 0.093103 | relationship_advice |
| 0.100608 | asktrp |
| 0.101775 | dirtypenpals |
| 0.10187 | stopdrinking |
| 0.102771 | exmormon |
| 0.102937 | breakingmom |
| 0.106659 | Drugs |
| 0.109762 | askgaybros |
| 0.113361 | asktransgender |
| 0.114258 | Christianity |
| 0.116465 | NoFap |
| 0.116918 | dating_advice |
| 0.117696 | legaladvice |

TABLE IV. TOP 20 SIMILAR SUBREDDIT RANKING FOR THE USER-CENTRIC METHOD.

| distance | subreddit |
|---|---|
| 0.195466212 | SuicideWatch |
| 0.204685824 | Anxiety |
| 0.214096225 | offmychest |
| 0.226656993 | socialanxiety |
| 0.245376634 | Advice |
| 0.270127495 | CasualConversation |
| 0.273800743 | BPD |
| 0.281158627 | bipolar |
| 0.295523869 | ForeverAlone |
| 0.312207559 | confession |
| 0.321152875 | BipolarReddit |
| 0.321237547 | raisedbynarcissists |
| 0.321867951 | relationship_advice |
| 0.321882484 | aspergers |
| 0.323138283 | ADHD |
| 0.338704493 | selfharm |
| 0.341794481 | OCD |
| 0.345224437 | ptsd |
| 0.345228268 | SeriousConversation |
| 0.349653894 | mentalhealth |

Tau rank correlation [14] coefficients over the top 40 subreddits for each method. Note that, since the user-centric method used a different set of candidate subreddits, subreddits not present in the 229 candidate subreddits were removed from that listing in the correlation. These coefficients and their respective P-values are listed in table V.

All p-values are significant ($\geq 0.05$), but strangely none of the correlations are particularly strong. This is likely due to the length of the sub-lists that were compared, as only the first ten or so entries are strongly correlated across the lists.

### C. Feature Importances

The random forest classifier assigns importances to each feature in terms of its ability to discriminate one label from the other. The list of words which best discriminated depression-

TABLE V. SPEARMAN'S RHO AND KENDALL'S TAU RANK CORRELATION COEFFICIENTS BETWEEN THE METHODS' LISTS.

| | Cosine | LDA | User-Centric |
|---|---|---|---|
| Spearman | 0.087 | -0.175 | 0.104 |
| P-Value | 0.198 | 0.093 | 0.174 |
| Kendall | 0.049 | -0.108 | 0.079 |
| P-Value | 0.219 | 0.109 | 0.157 |

TABLE VI. THE TOP 10 WORDS THAT DISCRIMINATE R/DEPRESSION FROM RANDOMLY-SELECTED POSTS.

| importance | words |
|---|---|
| 0.045848 | i |
| 0.040948 | feel |
| 0.038305 | depression |
| 0.032583 | myself |
| 0.022451 | don't |
| 0.021401 | just |
| 0.020019 | depressed |
| 0.01953 | me |
| 0.018206 | but |
| 0.017049 | friends |

relevant from non- posts reflects earlier research into the words that depressed people tend to use [15]. Specifically, they show a bias toward first-person personal pronouns (I, me, myself) in addition to the more obvious indicators of depression as a topic (e.g., depression, depressed).

Table VI is a selection of the 10 most important features in task 1, extracted from the 5000-element feature vector.

Figure 4 compares the importance of each word versus the rank of each word by importance. Importances, in accordance with Zipfs law, fall off at an inverse exponential rate.

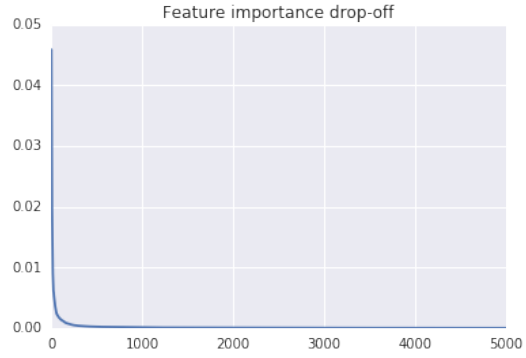


FIGURE 4. FEATURE IMPORTANCE DECLINES AT AN INVERSE EXPONENTIAL RATE IN ACCORDANCE WITH ZIPF'S LAW.

## VI. DISCUSSION, FUTURE WORK

While the random forest method does seem to present reasonable similarity rankings that align with the other known methods, there is an alternate interpretation of the difficulty in discriminating between two subreddits. It could simply be that the model is not sufficiently robust to identify the actual differences between the subreddits or the input is not sufficiently rich; thus, the framework considers the two subreddits to be the same when in fact it is an insufficiency of the model or feature representation. It would be of interest to explore models that can perform better on the differentiation task for pairs of subreddits.

An additional open question is whether the method described here is applicable to other domains, as it is well-known that depression-relevant text overexpresses personal pronouns as well as contains obvious signifiers such as "depression" or "depressed". It would be of interest to apply the method to other subreddits, or ideally across all subreddits to identify ones which are less readily distinguishable from the mean. This question is inherently related to the above regarding model robustness – a more robust model might accurately capture differences between subreddits that are more subtle than the ones between depression-relevant and irrelevant text.

Finally, it is appealing that this method relies solely on post text due to the tendency for users to seek support anonymously, but that advantage breaks down outside the support context. It may be useful to construct a hybrid model that makes use of both user- and content-centric clustering methods in a way that would address their mutual limitations.

## VII. CONCLUSION

In this work, we outlined the problem of exploring the relationships between self-help sub-forums on Reddit that are characterized by high self-disclosure, and consequently by anonymous posting behavior. We presented a method for ranking similar subreddits by the inability for a random forest classifier to distinguish between them, then compared its rankings to existing content-based and user-based subreddit similarity ranking methods. We present proposals to apply the approach to other corpora and to extend the framework with more sensitive classification on richer feature representations of the text, as well as hybrid user-content approaches that can circumvent anonymity by examining while still employing user data.

## REFERENCES


[1] redditmetrics.com: new subreddits by month. (accessed on 2018-02-01). [Online]. Available: \url{http://redditmetrics.com/history/month} (2017)

[2] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity." in ICWSM, 2014, pp. 71–80.

[3] J. Silterra, "Subreddit map," http://www.jacobsilterra.com/2015/03/10/subreddit-map/, 2015, (accessed on 2018-02-01).

[4] T. Martin, "Interactive map of reddit and subreddit similarity calculator," http://www.shorttails.io/interactive-map-of-reddit-and-subreddit-similarity-calculator/, 2016, (accessed on 2018-02-01).

[5] R. S. Olson and Z. P. Neal, "Navigating the massive world of reddit: Using backbone networks to map user interests in social media," PeerJ Computer Science, vol. 1, 2015, p. e4.

[6] A. Morcos, "Clustering subreddits by common word usage," http://www.arimorcos.com/blog/Clustering%20subreddits%20by%20common%20word%20usage/, 2015, (accessed on 2018-02-01).

[7] D. Wieker, "Subreddit clustering," http://dwieker.github.io/Reddit/, 2016, (accessed on 2018-02-01).

[8] F. Hoffa. 1.7 billion reddit comments loaded on bigquery. (accessed on 2018-02-01). [Online]. Available: \url{https://www.reddit.com/r/bigquery/comments/3cej2b/17_billion_reddit_comments_loaded_on_bigquery/} (2015)

[9] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[10] H. Zhang, "The optimality of naive bayes," AA, vol. 1, no. 2, 2004, p. 3.

[11] A. Singhal, "Modern information retrieval: A brief overview," 2001, pp. 35–43.

[12] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," in Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA, May 2010, pp. 45–50.

[13] C. Spearman, "The proof and measurement of association between two things," The American journal of psychology, vol. 15, no. 1, 1904, pp. 72–101.

[14] M. G. Kendall, "A new measure of rank correlation," Biometrika, vol. 30, no. 1/2, 1938, pp. 81–93.

[15] T. Brockmeyer et al., "Me, myself, and i: self-referent word use as an indicator of self-focused attention in relation to depression and anxiety," Frontiers in psychology, vol. 6, 2015, p. 1564.