Characterizing a New Taxonomy of Mental Disorders from Natural Language on Reddit

Scott L. Fleming
Department of Biomedical Data Science
Stanford University
Stanford, CA 94305
scottyf@stanford.edu
https://github.com/scottfleming/cs224n-project

Abstract

This project seeks to use a novel deep clustering algorithm on natural language from the social media website Reddit to characterize heterogeneity in the presentation of mental health disorders and associated patient experiences. Early results suggest that, while the methodology is promising, the current architecture, set of parameters, and dataset do not indicate clusters or subtypes that would corroborate associated literature from other data modalities. Qualitative analysis of clusters resulting from more traditional clustering methods like K-Means, however, suggests that a lack of clear correspondence between clusters and subreddit topics may indicate a true lack of consistency between subreddit labels and submission content. This carries important implications for researchers working on mental health data from Reddit.

1 Introduction, Background, and Related Work

The field of psychiatry has undergone something of a revolution in the last decade: while traditionally psychiatrists have understood mental health disorders based primarily on their presentation of symptoms in the clinic, as codified in the *Diagnostic Statistical Manual* (DSM), it has become increasingly clear that the delineations given by the current version of the DSM, namely the DSM-5, are neither coherent nor consistent [Cuthbert and Insel, 2013, Insel and Cuthbert, 2009]. Depression and anxiety, for example, frequently co-occur and the delineation between the disorders is not at all clear in many cases [Hirschfeld, 2001]. Indeed, 50% of individuals with at least one anxiety or mood disorder share a comorbid diagnosis of a second, while individuals within a single disease category in the DSM-5 may have just a single symptom in common. Given that mental health disorders are typically long-term, and given that clinical visits provide only a very small window into the patient's condition over time, there has been renewed attention around alternative ways to capture information about the mental health of psychiatric patients [Torous and Baker, 2016]. Efforts to better understand these conditions from multiple perspectives and data modalities, particularly those captured outside the context of a clinical visit, may elucidate a more consistent and biologically meaningful taxonomy of mental health disorders. A more consistent and biologically meaningful taxonomy could, in turn, have important implications for increasing the effectiveness of mental health treatment and care.

One alternative data modality that carries great potential in characterizing and delineating mental health disorders is natural language (text) on social media [De Choudhury, 2013, 2014, De Choudhury and De, 2014, De Choudhury, 2015]. The last decade has seen applications of natural language processing on social media data for mental health problems ranging from estimating disease prevalence [De Choudhury et al., 2013b,a, Bagroy et al., 2017] to forecasting both the onset and subsequent course of mental illness [De Choudhury et al., 2016, Reece et al., 2017]. Several recent papers have looked at Reddit specifically as a source of rich textual descriptions from patients with mental

32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

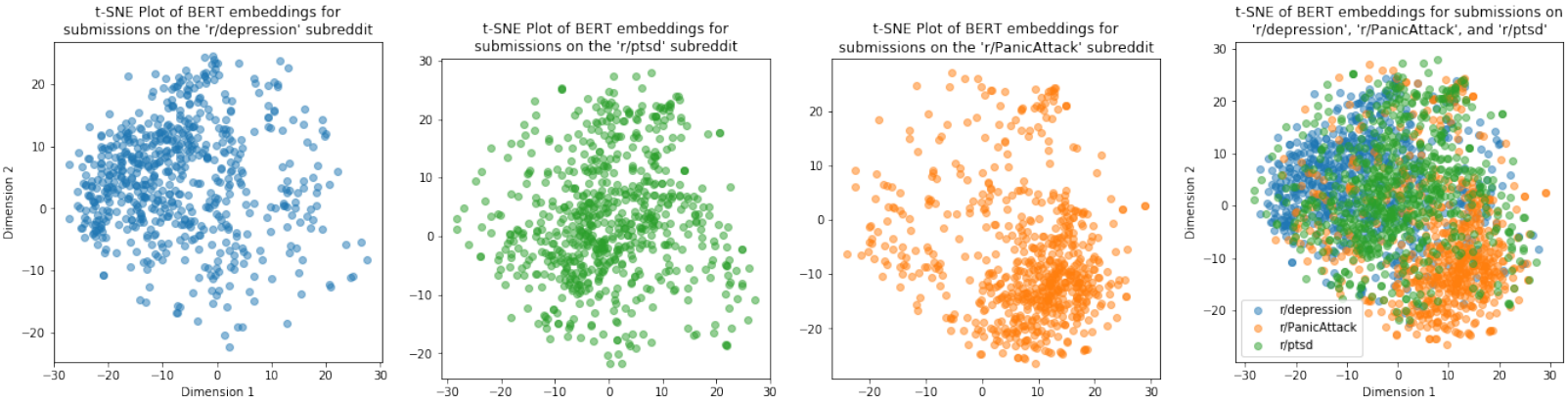

Figure 1: Visualization of BERT embeddings broken down by subreddit label, including depression (left), ptsd (center), and PanicAttack (right)

illnesses about their condition [Gkotsis et al., 2016, 2017, Ive et al., 2018]. Reddit provides a public forum for individuals to speak freely, with very little restrictions on what they say or how they say it. The Reddit site is organized into subreddits, each of which is organized around a theme. Many subreddits have formed around specific types of mental disorders, e.g. “r/depression", “r/ptsd", and “r/schizophrenia" (note that subreddit titles are typically referred to as “r/[subreddit name]"). Of note, users can communicate freely on one or more these subreddits and there are no restrictions of prior diagnosis, for example, for a user to post.

Prior work using natural language processing techniques on mental health subreddits has traditionally treated these subreddit thread topics/titles as ground truth labels (e.g. Gkotsis et al. [2017] and Ive et al. [2018]). In the context of clinical heterogeneity surrounding these mental health disorders, however, such an assumption seems potentially misguided. More specifically, the subreddit topic labels seem to follow a DSM-centric view of mental health disorders, perpetuating some of the same issues demonstrated in the DSM taxonomy itself. Additionally, it is not at all clear whether and to what degree users of Reddit post on the subreddit thread most appropriate to their post’s content. To that extent, it seems that taking the subreddit thread under which a post has been posted as the ground-truth label for that post is inappropriate.

Just how much systematic heterogeneity is there within and between subreddit posts? Do the inconsistencies in subreddit labels mirror some of the known inconsistencies within the DSM-5 taxonomy? If we were blind to the original subreddit labels, what sort of grouping word emerge naturally from the text itself? And, finally, would such a grouping corroborate recent findings around mental health disorder subtypes derived using different data modalities [Grisanzio et al., 2018]? In order to answer these questions, I propose an approach based on unsupervised learning to characterize natural clusters within a small subset subreddits related to mental health disorders.

Grisanzio et al. [2018] used cluster analysis to characterize 6 replicable psychiatric disorder subtypes within a cohort of 420 individuals, of which 100 had major depressive disorder (MDD), 53 had panic disorder (PD), 47 had posttraumatic stress disorder (PTSD), and 220 had no disorder (they were healthy controls). While these subtypes were based primarily on symptoms, as assessed by a standard self-report questionnaire (the Depression, Anxiety, and Stress Scale, or DASS), they nevertheless demonstrated meaningful between-subtype physiological differences. These differences include more objective measurements like brain activation, as measured through EEG, and cognition, as measured through standard computerized tasks. In the interest of (1) demonstrating the potential heterogeneity with and between subreddits as well as (2) characterizing alternative, data-driven taxonomies of mental health disorders corroborated by other modalities, I chose to focus on subreddits directly corresponding to the diagnostic labels analyzed by Grisanzio et al. [2018]. To my knowledge, this is the first demonstration of clustering methods for natural language on mental health discourse from social media.

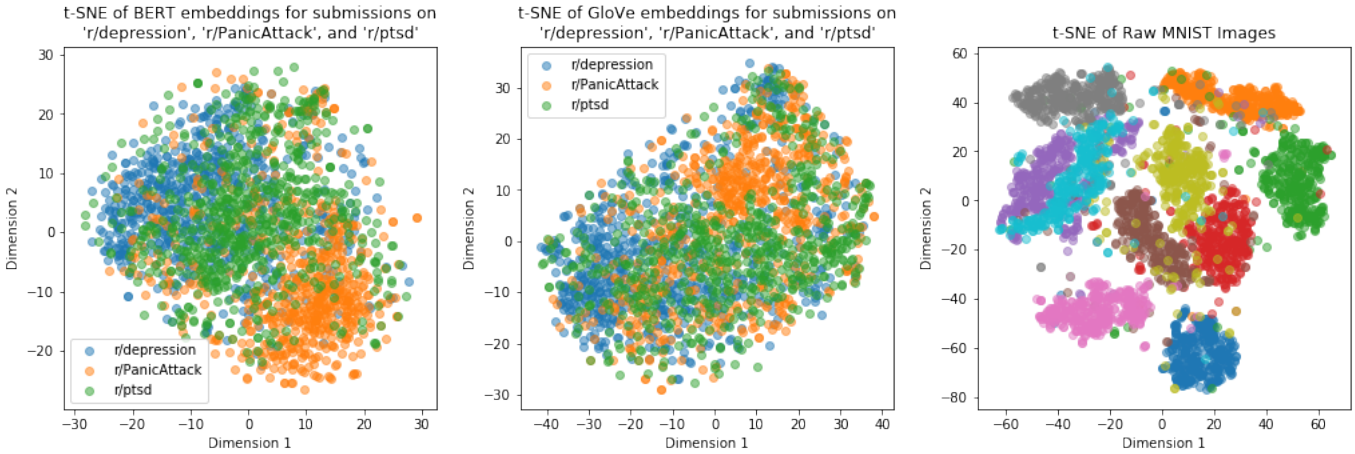

Figure 2: Visualization of BERT (left) vs. GloVe (center) embeddings for the reddit dataset. Comparison with a more well separated dataset, MNIST, is shown on the far right.

2 Approach

2.1 Dataset

In this project, I analyze the "r/depression", "r/PanicAttack", "r/ptsd" subreddits subreddits. These subreddits were chosen because they (1) were directly related to the MDD, PD, and PTSD diagnoses analyzed in Grisanzio et al. [2018]; (2) had large and active user bases ("r/depression" had 450k subscribers, "r/ptsd" had 21.2k subscribers, and "r/PanicAttack" had 4.4k subscribers at the time of this writing[1]); and (3) manual review of the content suggests that users' posts are indeed related to the subreddit theme. Using a publicly available API for querying Reddit data (`https://api.pushshift.io/`) I wrote code to extract all submissions (AKA original posts) made to "r/depression", "r/PanicAttack", and "r/ptsd" between January, 2006 and March, 2019. Given the substantial size imbalance between the "r/depression" subreddit relative to the others, I sampled the maximum number of submissions from each subreddit during time period such that the resulting subreddit labels would be balanced. This turned out to be 1500 overall submissions from each subreddit (rounding down to the nearest hundred).

2.2 Preprocessing

Given the loose restrictions on content, these reddit submissions were quite noisy. Misspellings, inappropriately concatentated words, and non-standard tokens were the rule rather than the exception. In order to handle these irregularities, I used a package called `ekphrasis` [Baziotis et al., 2017] which utilizes language patterns from a large social media corpus to perform preprocessing on text from social media. I adapted these tools (and some of their code examples) to normalize tokens like dates, numbers, emails, and times; segment words that were inappropriately concatenated; correct mispelled words; expand contractions; and convert emoticons to text. I additionally stripped submissions of all punctuations except periods to delineate the beginning and end of sentences.[2]

Once I standardized the text itself, I was able to use the `gensim` package data API to map every word in each submission to a dense/continuous representation, using GloVe embeddings as one approach and BERT as another.

[1] While there does exist a "r/panicdisorder" subreddit, it has only 687 subscribers.

[2] Delineating sentences was not necessary for the purposes of this milestone, but it will nevertheless be important in the context of the final project.

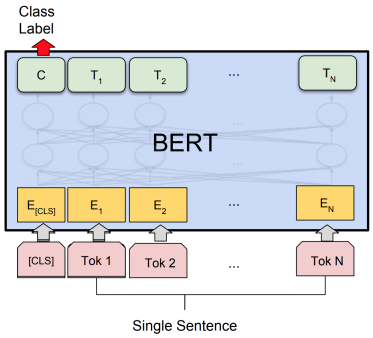

Figure 3: An illustration of the BERT architecture in the context of single sentence classification. My model used the second-to-last layer of the BERT architecture as word embeddings.

2.3 Featurization methods

2.3.1 GloVe Embeddings

Glove embeddings provide a way of mapping words to dense/continuous multidimensional representations [Pennington et al., 2014]. Learned in an unsupervised manner, GloVe embeddings are motivated by the idea that ratios of word-word co-occurrence statistics at a global level across an entire corpus carries rich information about the meaning of the word. More specifically, if we suppose that $P_{ij} = P(j|i)$ is the probability that the word j appears in the context of the word i in the corpus, then (at a high level) GloVe vectors try to find word vectors w_i, w_j, and $\tilde{w}_k$ such that

$$F\left((w_i - w_j)^T \tilde{w}_k\right) = \frac{P_{ik}}{P_{jk}} \tag{1}$$

. In my particular case, I used 100-dimensional GloVe embedding vectors trained on 2 billion tweets from Twitter (27 billion tokens total across a vocabulary of 1.2 million), which seemed like an appropriately similar source for the type of text that appears on Reddit. Following Arora et al. [2016], I simply used the average of all the GloVe vectors corresponding to all of the tokens in a submission to create a vector representation for that submission. These “submission vectors" were then passed on to the clustering model.

2.3.2 BERT Pretrained Embeddings

As an alternative and potentially more information-rich embedding approach, I also used embeddings from the uncased, pretrained BERT$_{BASE}$ model (hereafter referred to simply as “BERT") from Devlin et al. [2018]. The BERT model a multi-layer bidirectional Transformer encoder model described in Devlin et al. [2018]. Details of the model architecture and training are beyond the scope of this paper, but suffice it to say that the BERT model was trained on several large corpora of text for several different tasks, including sentence pair classification, single sentence classification, question answering, and single sentence tagging, among others. This pretraining enables contextual embeddings of words within a sentence in layers close to the output. For my specific model, to generate the embeddings of each word within a reddit post, I passed in each sentence from the post into the BERT model, then extracted the second-to-last layer’s output (see Figure 3 below) and averaged each of the word’s embedded representations to generate a single sentence embedding vector. All of the sentence embeddings within the post were averaged to generate an embedding of the entire post, which were than passed on to the clustering model. (Note that while the BERT model technically has a [CLS] class label token, these tokens are note as semantically meaningful without additional fine-tuning, which was against the spirit of my unsupervised learning task. Additional justification and reasoning surrounding this decision can be found in Xiao [2018]).

2.4 Clustering models

2.4.1 Deep Continuous Clustering (DCC)

My project's main goal was to perform clustering on the embeddings of subreddit submissions to see what groups naturally would emerge from the data. Implicit in this approach is the assumption that we do not know a priori what the correct number of clusters should be. One caveat with many traditional clustering algorithms is that they (1) poorly handle complex and high-dimensional data, (2) optimize a discrete and non-continuous objective function, and subsequently (3) require the user to specify the number of clusters *a priori*. These caveats are especially painful in our context, in which we decidedly do not know the appropriate number of clusters beforehand but nevertheless wish to cluster on complex, high-dimensional word embeddings. Shah and Koltun [2018] recently developed a clustering method that is able to elegantly solve each of these problems by using a neural network function approximator to optimize a continuous clustering objective.

The model proposed by Shah and Koltun [2018] assumes that the set of vectors representing the original data, $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]$ implicitly lie on a lower-dimensional manifold $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_N]$. The clustering problem then becomes one of simultaneously embedding the data into a lower dimensional space via an embedding function $f_\theta : \mathbb{R}^D \to \mathbb{R}^d$, while simultaneously maintaining a faithful representation of the data in this lower dimensional space, such that one could, for example, reproduce the original data $\mathbf{X}$ with reasonable fidelity. Thus, if we have a function $g_\omega : \mathbb{R}^d \to \mathbb{R}^D$ that performs the reverse mapping of f_θ, we constrain our lower dimensional representation to preserve the structure of the data in this lower dimensional space by optimizing the following objective:

$$\min_{\Omega} \|\mathbf{X} - G_\omega(\mathbf{Y})\|_F^2 \tag{2}$$

where $\mathbf{Y} = F_\theta(\mathbf{X}) = [f_\theta(\mathbf{x}_1), \ldots, f_\theta(\mathbf{x}_N)]$ and $G_\omega(\mathbf{Y}) = [g_\omega(\mathbf{y}_1, \ldots, g_\omega(\mathbf{y}_N)]$.

In order to push the lower-dimensional data representation toward an appropriate clustering in this lower-dimensional space, the Deep Continuous Clustering (DCC) objective proposes two additional terms to give the overall objective:

$$\mathcal{L}(\mathbf{\Omega}, \mathbf{Z}) = \frac{1}{D}\|\mathbf{X} - G_\omega(\mathbf{Y})\|_F^2 + \frac{1}{d}\left(\sum_i \rho_1\left(\|\mathbf{z}_i - \mathbf{y}_i\|; \mu_1\right) + \lambda \sum_{(i,j)\in\mathcal{E}} \omega_{i,j}\rho_2\left(\|\mathbf{z}_i - \mathbf{y}_i\|; \mu_2\right)\right) \tag{3}$$

which we wish to minimize. The matrix $\mathbf{Z} \in \mathbb{R}^{d\times N}$ in this case serves as a set of representative candidates $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$ in the lower-dimensional space such that we do not optimize over $\mathbf{Y}$ by changing its value directly, but rather change the value of our representatives $\mathbf{Z}$ and penalize representations that drift far from the original lower-dimensional embeddings $\mathbf{Y}$. This is captured in the data loss part of the penalty shown above, namely $(\sum_i \rho_1\left(\|\mathbf{z}_i - \mathbf{y}_i\|; \mu_1\right))$. We additionally push the candidate representations $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_N]$ together (e.g. to encourage the clustering behavior we desire) by penalizing $\mathbf{z}_i, \mathbf{z}_j$ for being far apart when they are in fact close together in the original data space. More specifically, if we let $\mathcal{E}$ represent the graph of neighbors in the original data $\mathbf{X}$ such that the edge $(i, j) \in \mathcal{E}$ if $\mathbf{x}_i$ is within the k-nearest neighbors of $\mathbf{x}_j$ and vice versa, then the last part of the penalty written above, namely $\sum_{(i,j)\in\mathcal{E}} \omega_{i,j}\rho_2\left(\|\mathbf{z}_i - \mathbf{y}_i\|; \mu_2\right)$, penalizes $\mathbf{z}_i$ and $\mathbf{z}_j$ from being far apart when $\mathbf{x}_i$ and $\mathbf{x}_j$ are close together.

In practice, I used a Stacked Denoising Autoencoder [Vincent et al., 2010] to intialize candidate values for Z and represented G_ω and F_θ with densely connected two-layer neural networks. The code for this algorithm was implemented by Shah and Koltun [2018] in Pytorch, but I had to perform nontrivial work to adapt it to the context of this project.

2.4.2 K-Means++

K-Means Clustering is a well-known clustering algorithm that takes in a specified number of clusters from the user, K, randomly selects putative cluster centers (or centroids), and iterativey refines the choice of these centroids and associated cluster assignments of data points to find the following

objective:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} \|\mathbf{x} - \mu_i\|^2 \tag{4}$$

where $S = \{S_1, \ldots, S_k\}$ is the set of k cluster assignments, S_i is the set of all datapoints associated with the i^{th} cluster, and μ_i is the cluster centroid of the i^{th} cluster. K-Means++ refines the initial seeding/selection of cluster centroids so as to accelerate convergence [Arthur and Vassilvitskii, 2007]. While perhaps not state-of-the-art, K-Means clustering provides a reasonable and oft-used comparison for other more novel clustering algorithms given its elegance and simplicity. I employed K-Means++ on the raw GloVe and BERT embeddings (i.e. without any autoencoding).

3 Experiments

I ran the aforementioned submissions from Reddit through the clustering network to evaluate both the number and quality of clusters. I used a learning rate of Following Vincent et al. [2010] and Vinh et al. [2010], I used both adjusted mutual information (AMI) and unsupervised clustering accuracy (ACC) to compare my clustering result against the original subreddit labels.

3.1 Experimental Details

For the SDAE initialization in the DCC algorithm, I used the default parameters given by Vincent et al. [2010] and Shah and Koltun [2018], namely a D-500-500-2000-d-2000-500-500-D SDAE architecture, where D is the original data dimension and d is the embedded dimension with d = 10, a minibatch size of 256, a dropout probability of 0.2, a learning rate of 0.1 for the BERT embeddings and a learning rate of 10 for the GloVe embeddings (tuned to be as large as possible such that the SDAE loss still converged), with a learning rate schedule of decreasing the learning rate by a factor of 10 every 80 epochs. Each layer in the SDAE was pretrained for 356 epochs, except for the last layer, which was trained for 712 epochs. Optimization of the DCC objective was performed using an Adam optimizer with the default learning rate of 0.001 and momentum 0.99, as described in Shah and Koltun [2018].

3.2 Evaluation Metrics

3.2.1 Adjusted Mutual Information (AMI)

The AMI [Vinh et al., 2010] takes into account both mutual information between a “ground truth" set of labels (e.g. the original subreddit under which the submissions were posted) and a candidate label (e.g. the one assigned by my clustering network):

$$AMI(\mathbf{c}, \hat{\mathbf{c}}) = \frac{MI(\mathbf{c}, \hat{\mathbf{c}}) - E[MI(\mathbf{c}, \hat{\mathbf{c}})]}{\sqrt{H(\mathbf{c})H(\hat{\mathbf{c}})} - E[MI(\mathbf{c}, \hat{\mathbf{c}})]} \tag{5}$$

where $H(\mathbf{c}$ is the entropy of labels $\mathbf{c}$ and $MI(\mathbf{c}, \hat{\mathbf{c}})$ is the mutual information between two partitions/clustering assignments $\mathbf{c}$ and $\hat{\mathbf{c}}$. Intuitively, the AMI is higher when the partitions $\mathbf{c}$ and $\hat{\mathbf{c}}$ agree. Note that while the AMI provides a useful and objective measure of cluster quality, it is *not* the objective which our clustering network is attempting to minimize.

3.2.2 Unsupervised Clustering Accuracy (ACC)

The unsupervised clustering accuracy (ACC) takes a cluster assignment and finds the best possible assignment of clusters to “ground-truth" labels by minimizing the following objective:

$$ACC = \max_{m} \frac{\sum_{i=1}^{n} \mathbf{1}\{l_i = m(c_i)\}}{n} \tag{6}$$

where l_i in this case is the subreddit label for reddit post i, c_i is the cluster to which that post was assigned by our algorithm, and there are m possible one-to-one mappings between clusters and labels. While we do not necessarily trust our “ground-truth" subreddit topic labels, this provides an insightful way of comparing the two.

Table 1: ACC and AMI for Experiments on Reddit Data

	AMI		ACC	
Algorithm	BERT	GloVe	BERT	GloVe
K-Means	0.1023	-0.0001	0.5313	0.4676
DCC	0.0383	0.0547	0.2331	0.3329

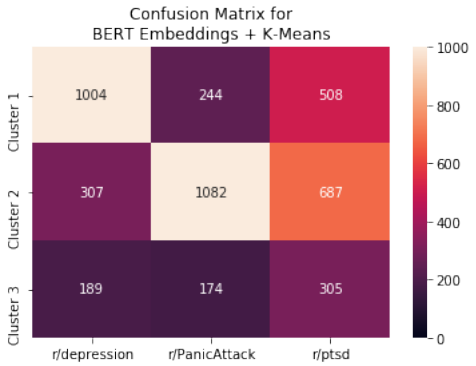

Figure 4: Confusion Matrix for K-Means clustering using BERT embeddings

4 Results

Results of evaluating my clustering against the “ground-truth" subreddit labels is given in Table 1. Notably, with K-Means as a clustering algorithm, the BERT model representations yielded better agreement with the subreddit labels relative to the GloVe-based representations based on AMI (0.1023 vs. -0.0001, respectively) and ACC (0.5313 vs. 0.4676) but the opposite was true for the DCC clustering algorithm. Overall, K-Means clustering obtained a higher AMI and ACC compared to DCC for all representations except BERT.

The distribution of cluster assignments relative to the subreddit labels (i.e. confusion matrix) is given in Figure 4. Cluster 1 appeared to align most closely with the ‘r/depression’, Cluster 2 with the ‘r/PanicAttack’ subreddit, and Cluster 3 essentially none of the above.

5 Discussion and Analysis

Admittedly, the results of running DCC on the BERT-based reddit submission embeddings were somewhat disheartening. It appears from monitoring the evaluation metrics, overall loss, and size of the largest cluster resulting from DCC during the training process (see Figure 5) that the DCC algorithm essentially collapsed all representations into a single cluster. Given the balanced nature of the dataset and the fact that there were three subreddit labels included, this meant that the clustering accuracy was capped at $1/3 = 0.33$. This despite the fact that the total loss decreased monotonically.

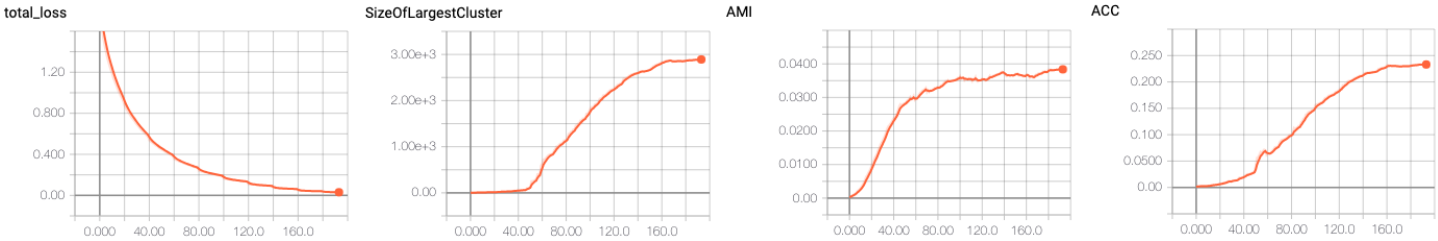

Figure 5: Loss, AMI, ACC, and the size of the largest cluster relative to the number of iterations during DCC training.

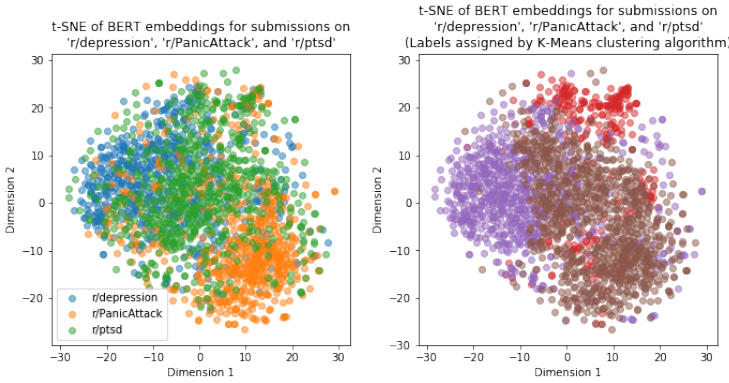

Figure 6: Clustering with K-Means

In evaluating why this collapse may have occurred, I visualized the distribution of the BERT- and GloVe-based representations, colored by subreddit label, using t-SNE [Maaten and Hinton, 2008]. The comparison is given in Figure 2.

From the visualization, it appeared that there was a reasonable ordering of the subreddit labels in BERT/GloVe-space and that this ordering was more accentuated in the BERT representation. Visualizing each subreddit individually, as in Figure 1, we see that while the subreddits each have a distinct location, there is substantial overlap. This, in comparison to the MNIST dataset, for example (also see Figure 2) on which Shah and Koltun [2018] evaluated their algorithm and on which their reported AMI and ACC were quite high.

When I fixed the number of clusters to match the number of subreddit topics we knew were actually present in the dataset, I saw substantial improvement in both the ACC and AMI, but it appeared that third cluster, Cluster 3 in Figure 4, was not assigned to any particularly meaningful label. This can be seen visually in Figure 6.

A more qualitative analysis revealed why this third cluster might be inappropriately assigned to any one of the subreddit labels or, alternatively, why the ptsd subreddit thread seemed to be split across all three clusters. One reddit post that originally in the ptsd subreddit but was assigned to the depression-related cluster, for example, read:

```
i have ptsd due to...  lets just say childhood abuse.  i also have
schizaffective thanks to my mother's genes .  its gotten to a point where
i am having flashbacks a day plus nightmares.  its causing depression and
extreme anxiety and isolation.  i'd rather die than keep living like this.
so i am going to the hospital.  i am scared.  i do not want to go but i can
not keep doing this.  i can not keep reliving my sexual assault over and
over.
```

Note that the author referred to feelings of depression and anxiety, such that, despite explicitly containing a reference to PTSD, it seems reasonable that this post was clustered with other reddit posts.

6 Conclusion

While the overlap between the clusters assigned by my clustering algorithm did not yield a perfect correspondence to the original subreddit labels, my qualitative analysis suggests that this mismatch is indicative of a broader pattern of mismatch between subreddit topic and actual submission content. This calls into question the use of subreddit topics as "ground-truth" labels for predictive analyses on mental health reddit data (e.g. [Gkotsis et al., 2017]). Unfortunately, the heterogeneity was strong enough that the algorithm chosen to identify the number of natural groupings in the dataset, namely Deep Continuous Clustering, was not able to find any groupings at all. Thus we did not corroborate the findings of Grisanzio et al. [2018]. Further work will be required to refine these analyses, including the inclusion of other subreddit topics and clustering algorithms more robust to noise.

7 Additional Information

External Collaborators: Daniel Rubin (Professor, Biomedical Data Science); Imon Banerjee (Instructor, Radiology); Nandita Bhaskhar (EE Ph.D. Student in Dr. Rubin's Lab); Shaimaa Bakr (EE Ph.D. Student affiliated with Dr. Rubin's Lab); **Project Type:** Custom; **Mentor:** Suvadip

Bibliography

S. Arora, Y. Liang, and T. Ma. A simple but tough-to-beat baseline for sentence embeddings. 2016.

D. Arthur and S. Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

S. Bagroy, P. Kumaraguru, and M. De Choudhury. A social media based index of mental well-being in college campuses. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 1634–1646. ACM, 2017.

C. Baziotis, N. Pelekis, and C. Doulkeridis. Datastories at semeval-2017 task 4: Deep lstm with attention for message-level and topic-based sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 747–754, Vancouver, Canada, August 2017. Association for Computational Linguistics.

B. N. Cuthbert and T. R. Insel. Toward the future of psychiatric diagnosis: the seven pillars of rdoc. *BMC medicine*, 11(1):126, 2013.

M. De Choudhury. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd International Workshop on Socially-aware Multimedia*, pages 49–52. ACM, 2013.

M. De Choudhury. Can social media help us reason about mental health? In *Proceedings of the 23rd International Conference on World Wide Web*, pages 1243–1244. ACM, 2014.

M. De Choudhury. Social media for mental illness risk assessment, prevention and support. In *Proceedings of the 1st ACM Workshop on Social Media World Sensors*, pages 1–1. ACM, 2015.

M. De Choudhury and S. De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

M. De Choudhury, S. Counts, and E. Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013a.

M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*, 2013b.

M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, and M. Kumar. Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems*, pages 2098–2110. ACM, 2016.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

G. Gkotsis, A. Oellrich, T. Hubbard, R. Dobson, M. Liakata, S. Velupillai, and R. Dutta. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, pages 63–73, 2016.

G. Gkotsis, A. Oellrich, S. Velupillai, M. Liakata, T. J. Hubbard, R. J. Dobson, and R. Dutta. Characterisation of mental health conditions in social media using informed deep learning. *Scientific reports*, 7:45141, 2017.

K. A. Grisanzio, A. N. Goldstein-Piekarski, M. Y. Wang, A. P. R. Ahmed, Z. Samara, and L. M. Williams. Transdiagnostic symptom clusters and associations with brain, behavior, and daily function in mood, anxiety, and trauma disorders. *JAMA psychiatry*, 75(2):201–209, 2018.

R. M. Hirschfeld. The comorbidity of major depression and anxiety disorders: recognition and management in primary care. *Primary care companion to the Journal of clinical psychiatry*, 3(6):244, 2001.

T. R. Insel and B. N. Cuthbert. Endophenotypes: bridging genomic complexity and disorder heterogeneity. 2009.

J. Ive, G. Gkotsis, R. Dutta, R. Stewart, and S. Velupillai. Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 69–77, 2018.

L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov): 2579–2605, 2008.

J. Pennington, R. Socher, and C. Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer. Forecasting the onset and course of mental illness with twitter data. *Scientific reports*, 7(1):13006, 2017.

S. A. Shah and V. Koltun. Deep continuous clustering. *arXiv preprint arXiv:1803.01449*, 2018.

J. Torous and J. T. Baker. Why psychiatry needs data science and data science needs psychiatry: connecting with technology. *JAMA psychiatry*, 73(1):3–4, 2016.

P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, and P.-A. Manzagol. Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, 11(Dec):3371–3408, 2010.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct): 2837–2854, 2010.

H. Xiao. bert-as-service. `https://github.com/hanxiao/bert-as-service`, 2018.