Quantifying Mental Health from Social Media with Neural User Embeddings

Silvio Amir[1] SAMIR@INESC-ID.PT
Glen Coppersmith[3] GLEN@QNTFY.COM
Paula Carvalho[1,2] PCC@INESC-ID.PT
Mário J. Silva[1] MJS@INESC-ID.PT
Byron C. Wallace[4] B.WALLACE@NORTHEASTERN.EDU

1. *INESC-ID Lisboa, Instituto Superior Técnico, Universidade de Lisboa*
2. *Universidade Europeia, LIU*
3. *Qntfy, Washington DC, United States*
4. *Northeastern University, Boston MA, United States*

Abstract

Mental illnesses adversely affect a significant proportion of the population worldwide. However, the typical methods to estimate and characterize the prevalence of mental health conditions are time-consuming and expensive. Consequently, best-available estimates concerning the prevalence of these conditions are often years out of date. Automated approaches that supplement traditional methods with broad, aggregated information derived from social media provide a potential means of furnishing near real-time estimates at scale. These may in turn provide grist for supporting, evaluating and iteratively improving public health programs and interventions.

We propose a novel approach for mental health quantification that leverages *user embeddings* induced from social media post histories. Recent work showed that learned user representations capture latent aspects of individuals (e.g., political leanings). This paper investigates whether these representations also correlate with mental health statuses. To this end, we induced embeddings for a set of users known to be affected by depression and post-traumatic stress disorder, and for a set of demographically matched 'control' users. We then evaluated the induced user representations with respect to: (i) their ability to capture homophilic relations with respect to mental health statuses; and (ii) their predictive performance in downstream mental health models. Our experimental results demonstrate that learned user embeddings capture relevant signals for mental health quantification.

1. Introduction

Mental illness is a critically important concern, significantly and adversely affecting a wide swath of the population directly and indirectly. An estimate by the Centers for Disease Control from 2008, suggests that 9% of US adults may meet the criteria for depression at any given time (CDC, 2010). While not as prevalent as depression, post traumatic stress disorder (PTSD) issues also cost hundreds of billions of dollars worldwide, according to a conservative estimate from the NIH.[1] The collective effect of mental health conditions, as measured by Daily Adjusted Life Years (DALYs), exceeds that of malaria, war, or violence[2] (Whiteford et al., 2013). At the same time, mental health problems are often difficult

[1] `https://www.nimh.nih.gov/health/statistics/cost/index.shtml`

[2] For a visualization of DALYs see `https://vizhub.healthdata.org/gbd-compare/`

to identify and thus treat. For example, perhaps half of depressive cases go undetected, in part due to the heterogeneous and complex expression of this condition (Paykel et al., 1997). Another exacerbating factor is that diagnosis generally requires individuals to actively seek out treatment. Yet the manifestation of this condition and prevailing social stigmas may disincline afflicted individuals to take these steps.

The internet may provide a comfortable medium for people to express their feelings anonymously and connect with healthcare professionals (McCaughey et al., 2014) and others affected by similar conditions (De Choudhury et al., 2016). Furthermore, individuals openly discuss mental health challenges on public social network platforms such as Twitter (Coppersmith et al., 2014a, 2015a). Prior work has demonstrated the potential of using social media to investigate mental health issues (Paul and Dredze, 2011), including depression (Schwartz et al., 2014), PTSD (Coppersmith et al., 2014b) and suicidal ideation (Coppersmith et al., 2016; De Choudhury et al., 2016), however models and techniques to identify and quantify mental health related signals from social media are relatively novel. Interest in these applications has motivated the creation of a shared task for the Computational Linguistics and Clinical Psychology workshop (CLPsych),[3] aimed at advancing the state-of-the-art in technologies capable of discriminating users affected by mental illness from controls, given their post history (Coppersmith et al., 2015b). A variety of methods have been proposed for this task, but none have achieved consistently superior performance, implying improvements may yet be realized by novel approaches.

Neural representation learning methods have been shown capable of learning good representations from raw data, freeing practitioners from the burden of manually designing and encoding task-specific features (Bengio et al., 2015; Goldberg, 2016). *Word embeddings*, e.g., aim to implicitly encode latent word semantics, and can be learned with predictive models that exploit word co-occurrence statistics and other regularities in unlabeled corpora (Bengio et al., 2003). These models have been recently extended to infer representations for larger textual units (Le and Mikolov, 2014), and even user representations (Yu et al., 2016; Amir et al., 2016). It has been shown that such *user embeddings* capture latent personal aspects and can be used in downstream applications, such as sarcasm detection (Amir et al., 2016) and content recommendation (Yu et al., 2016).

This paper investigates whether user representations induced via neural models can inform clinical models operating over social media. Specifically, we assess whether embeddings learned directly from user post histories capture aspects that indicate mental health status. We use the dataset created for the CLPsych shared task to address two research questions: (1) To what extent do user embeddings capture information relevant for mental health analysis over social media? (2) Can user embeddings be leveraged to discriminate between users diagnosed with mental illness and demographically matched controls? For the first question we compare different approaches to estimate user embeddings from post histories. In particular, we investigate whether the learned embeddings capture homophilic relations between users with respect to mental health. To answer the second question, we develop and evaluate models that leverage user embeddings to infer mental health status.

The main contributions of this paper are as follows: (i) we show that unsupervised user embeddings induced from posting histories capture user similarities, and are predictive of

[3] `http://clpsych.org`

mental health conditions; (ii) we develop a novel neural model that incorporates and refines these embeddings to improve the categorization of users with respect to mental health status. Furthermore, we show that the resultant fine-tuned user embeddings better align with mental health conditions.

The remainder of the paper is organized as follows. We next introduce the CLPsych shared task and dataset. In Section 3 we review related work on user modelling for social media analysis and neural embeddings. In Section 4, we describe the user embedding model used in our experiments, and discuss connections to prior approaches. Section 5.2 addresses the first research question by evaluating properties captured by the learned user embeddings. Section 5.3 reports results from classification experiments designed to address the second research question. We present our conclusions in Section 6.

2. Depression and PTSD on Twitter

In 2015, the CLPsych workshop held a shared task to foster progress in NLP technologies related to mental health analysis, over social media streams (Mitchell et al., 2015; Coppersmith et al., 2015b). A dataset was compiled comprising users that have publicly stated on Twitter that they were diagnosed with depression (327 users) or PTSD (246 users), and an equal number of randomly selected demographically-matched users as *controls*.[4] For each user in this dataset, associated metadata and posting history was also collected — up to the 3000 most recent *tweets*, per limitations of the Twitter API. For more details on the construction and validation of the data, see (Coppersmith et al., 2015b, 2014a,b).

Task participants then aimed to develop models to discriminate between users affected by mental illness from controls, given their posts and metadata. Specifically this entailed three binary sub-tasks: (i) `depression vs control`, (ii) `PTSD vs control` and (iii) `depression vs PTSD`. The proposed systems were based on a wide range of approaches including: rule-based systems leveraging lexical decision lists (Pedersen, 2015), linear classifiers exploiting features based on word clusters and topic models (Preotiuc-Pietro et al., 2015), supervised topic models (Resnik et al., 2015) and systems exploiting character-level language models (Coppersmith et al., 2015b). However, none of the proposed systems performed consistently better than the others across all the sub-tasks and evaluation metrics, highlighting the difficulty of this problem. Moreover, none of these systems used explicit representations of *users*, which is the innovation we propose here. We did not participate in this shared task, and thus could not obtain the official test data. Therefore, our results are not directly comparable to those of the participating teams. Nevertheless, we compared our proposed approach with the majority of the previously proposed methods.

3. Related Work

Most of the research in social media analysis has been concerned with deriving better models that operate on representations of the texts comprising individual user posts, both via manually crafted features and, more recently, representation learning approaches (Severyn and Moschitti, 2015; Astudillo et al., 2015). However for many problems it is crucial to also capture characteristics of the *users* communicating. These include, information extraction (Yang et al., 2016), opinion mining (Tang et al., 2015), sarcasm detection (Bamman

[4] This data was collected according to the ethical protocol of Benton et al. (2017), and follows the recommendations spelled out in Mikal et al. (2016).

and Smith, 2015) and content recommendation (Yu et al., 2016). The most straightforward approach to induce user representations is by scrapping "profile" information from social websites, to manually extract features based on social ties, demographic attributes, or posting habits (Bamman and Smith, 2015; Rajadesingan et al., 2015). However, this requires significant effort for data collection, and derived features will be platform-, task- and domain-specific. Furthermore, user profile information may be inaccurate or outdated.

Neural Embedding Learning

In recent years, models in NLP have moved from operating over sparse representations (scalars indexing into a pre-defined vocabulary) towards *distributed* continuous word embedding representations that encode latent semantics (Goldberg, 2016). The general approach to unsupervised induction of word embeddings entails associating words with parameter vectors, which are then optimized to predict other words that occur in the same contexts (Bengio et al., 2003). Skip-Gram (Mikolov et al., 2013), e.g., operationalizes this approach by sliding a *window* of a pre-specified size across texts. At each step, the center word is used to predict the probability of one of the surrounding words, sampled proportionally to the distance to the center word. Le and Mikolov (2014) later expanded this approach to learn representations for paragraphs (or, more generally, sequences of words). This is referred to as *Paragraph2Vec*, and two variants were proposed: (i) **PV-DM** tries to predict the center word of the sliding window, given the surrounding words **and** the paragraph (i.e., their respective embeddings); and (ii) **PV-DBOW**, tries to predict the words in a sliding window within a paragraph, conditioned only on the respective paragraph embedding.

Recently proposed methods to learn user representations use essentially the same approach, associating *users* with parameter vectors, and optimizing these to accurately predict observable attributes or the words used in previous posts written by said user (Amir et al., 2016; Yu et al., 2016). User embeddings induced by Amir et al. (2016) using only the previous posts from a user were shown to capture latent individual attributes (e.g. political leanings) and a soft notion of 'homophily' — i.e., similar users were generally associated with relatively nearby vectors. Further, the embeddings improved a downstream model for sarcasm detection in tweets. Similarly, Yu et al. (2016) improved a microblog recommendation system by including user representations. Our work aims to ascertain if embeddings learned only from previous posts can capture useful signals for clinical applications.

4. Mental Health Inference over Social Media

Our general approach to mental health inference over social media entails learning unsupervised user embeddings from user post histories to be used as features in downstream specialized models. However, generic features estimated with unsupervised models are suboptimal for downstream applications (Zhang and Wallace, 2015). Training neural models end-to-end (including representational parameters) can refine generic embeddings for specific tasks (Collobert et al., 2011). However, this strategy requires updating a large number of parameters, which is difficult when only small training datasets are available. Therefore, given the modest size of the CLPysch dataset, we adopted the Non-Linear Subspace Embedding approach (NLSE) due to Astudillo et al. (2015), which adapts generic representations to specific tasks using a small amount of labeled data.

4.1 Learning User Embeddings

We induced user embeddings with an approach similar to that recently proposed by Amir et al. (2016). The idea is to capture relations between users and the content (i.e., the words) they generate, by optimizing the probability of sentences conditioned on their authors. Formally, let $\mathcal{U}$ be a set of users, $\mathcal{C}_j$ be a collection of posts authored by user $u_j \in \mathcal{U}$, and $S = \{w_1, \ldots, w_N\}$ be a post composed of words w_i from a vocabulary $\mathcal{V}$. The goal is to estimate the parameters of a user vector $\mathbf{u}_j$, that maximize the conditional probability:

$$P(\mathcal{C}_j|u_j) \propto \sum_{S \in \mathcal{C}_j} \sum_{w_i \in S} \log P(w_i|\mathbf{u}_j) \tag{1}$$

However, directly estimating these quantities (e.g., with a log-linear model) would require calculating a normalizing constant over a potentially large number of words, a computationally expensive operation. Because we are only interested in the user vectors $\mathbf{u}_j$ and not the actual probabilities as such, we can approximate the term $P(w_i|\mathbf{u}_j)$ by minimizing the following Hinge-loss objective:

$$\mathcal{L}(w_i, u_j) = \sum_{\tilde{w}_k \in \mathcal{V}} \max(0, 1 - \mathbf{w}_i \cdot \mathbf{u}_j + \tilde{\mathbf{w}}_k \cdot \mathbf{u}_j) \tag{2}$$

where word $\tilde{w}_k$ (and associated embedding, $\tilde{\mathbf{w}}_k$) is a *negative sample*, i.e. a word not occurring in the post under consideration (authored by user u_j). By learning to discriminate between observed positive examples and *pseudo*-negative examples, the model shifts probability mass to more plausible observations (Smith and Eisner, 2005). Note that we represent both *words* and *users* via d-dimensional vectors — word vectors, $\mathbf{w}_i \in \mathbb{R}^d$ which are assumed to have been pre-trained via some neural language model; and user vectors $\mathbf{u}_j \in \mathbb{R}^d$ to be learned. We will refer to this approach as USER2VEC.[5]

We note that barring some minor operational differences, this model is equivalent to the PV-DBOW variant of *Paragraph2vec* — if users are viewed as paragraphs. The key differences are that: (i) USER2VEC predicts **all** the words in a post, whereas PV-DBOW slides a window along the paragraph and only predicts one word per step; and (ii) USER2VEC assumes that the word embeddings are pre-trained, whereas PV-DBOW aims to jointly learn the word and paragraph vectors.

4.2 Domain-Specific Representations with Embedding Subspaces

Neural embeddings are *distributed representations*, i.e. each concept is described by multiple *features*[6] and each feature can be involved in describing multiple concepts (Hinton, 1986). The subspace embedding approach is based on the assumption that, for a given problem, there is a subset of features responsible for encoding the most relevant information. Thus, instead of directly modifying the embeddings, the NLSE model uses the labeled data to extract a compact representation by learning linear projections into lower-dimensional subspaces (Astudillo et al., 2015). The resulting *embedding subspaces* capture domain- and

[5] This formulation is a simplification of Amir et al. (2016) model. Specifically, we omitted a term in Eq.1, encoding the marginal probability of S; and we allow the negative samples to be drawn from all the words in $\mathcal{V}$. These simplifications dramatically reduce training time without significant loss of quality on the resulting embeddings.

[6] Vector dimensions can be interpreted as abstract features.

task-specific aspects, while preserving the rich information encoded by the original embeddings. Furthermore, a dimensionality reduction of the feature space eliminates noise and reduces the number of free parameters, making downstream models easier to fit with small datasets.

More formally, given an user embedding matrix $\mathbf{U} \in \mathbb{R}^{d \times |\mathcal{U}|}$, in which column $\mathbf{U}_{[j]}$ represents user $u_j \in \mathcal{U}$, we induce new representations by factorizing the input as $\mathbf{S} \cdot \mathbf{U}$ where $\mathbf{S} \in \mathbb{R}^{s \times d}$, with $s \ll d$, is a (learned) linear projection matrix. The intuition is that by aggressively reducing the representation space, the model is forced to learn only the most discriminative aspects of the input with respect to the prediction targets. The NLSE model is similar to a feed-forward neural network with an embedding layer and a single hidden layer. The main differences are: (1) the factorization of the embedding layer into two components (the original embedding matrix and a linear projection matrix) and, (2) the dimensionality reduction induced by the subspace projection, with typical reductions greater than an order of magnitude. Similar to feed-forward networks, we can use backpropagation (Rumelhart et al., 1988) to jointly learn task-specific embeddings and the parameters for the classification layer. Using this approach, our proposed mental health prediction model can be formalized as:

$$\begin{aligned} P(\mathcal{Y}|u_j) &\propto \boldsymbol{\beta} \cdot g(u_j) \\ g(u_j) &= \sigma\left(\mathbf{S} \cdot \mathbf{U}_{[j]}\right) \end{aligned} \tag{3}$$

where $\sigma(\cdot)$ denotes an element-wise sigmoid non-linearity, and the matrix $\boldsymbol{\beta} \in \mathbb{R}^{|\mathcal{Y}| \times s}$ maps the embedding subspace to the classification space. Notice that at inference time this model reduces to a linear classifier with embedding subspace features.

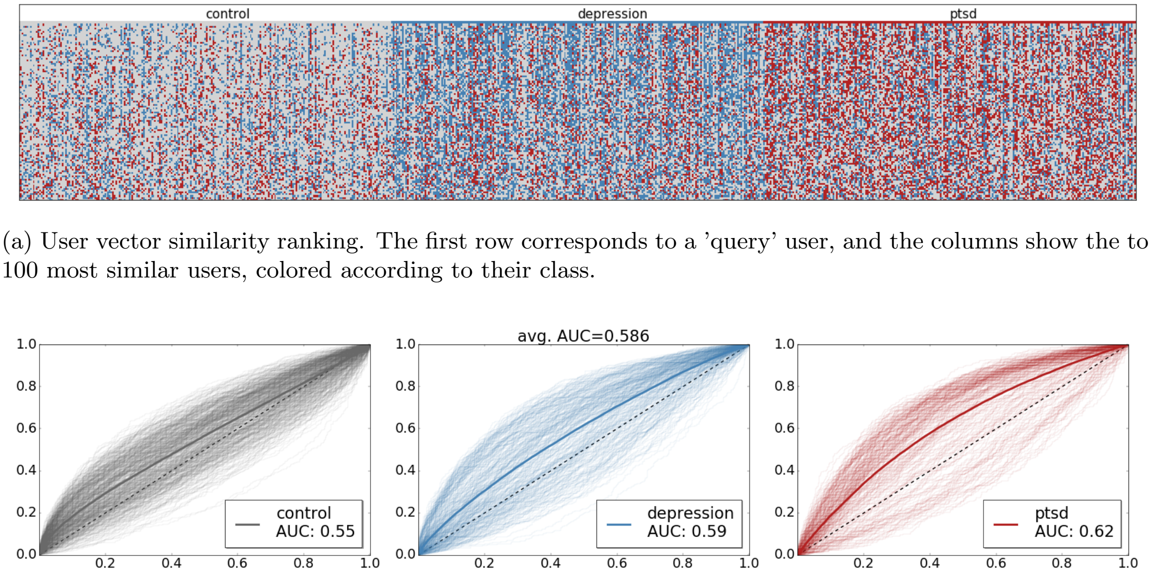
5. User Embedding Evaluation

In this section we evaluate our proposed methodology and address our research questions. Our main goal is to investigate whether user embeddings learned from Twitter post histories encode information relevant for public health applications. If user embeddings indeed correlate with mental health, then they could be used to recognize and characterize risk groups on social media, e.g., by identifying individuals that 'look' like patients affected by depression. This would potentially enable scalable, real-time estimates concerning the prevalence of mental health issues in particular sub-populations.

5.1 Experimental Setup

The experimental evaluation was conducted with the CLPysch shared task corpus described in Section 2. We first preprocessed all tweets with a conventional normalization scheme that included: lowercasing; reducing character repetitions to at most three; and replacing user mentions and URLs with canonical forms. Users with fewer than 100 tweets were discarded. We estimated embeddings for all remaining users in the corpus, leveraging the USER2VEC,[7] and *Paragraph2Vec*'s PV-DM and PV-DBOW models (Section 3). To learn USER2VEC embeddings, we first pre-trained a set of SKIP-GRAM **word** vectors from a large unlabeled corpus comprising the task data and an additional set of 53 Million tweets. Next, for each user $u_j \in \mathcal{U}$, we sampled a held-out set $\mathcal{H}_j \subset \mathcal{C}_j$ with 10% of the posting history. The rest of the data was used to estimate an embedding $\mathbf{u}_j$, by minimizing Eq. 1 via SGD

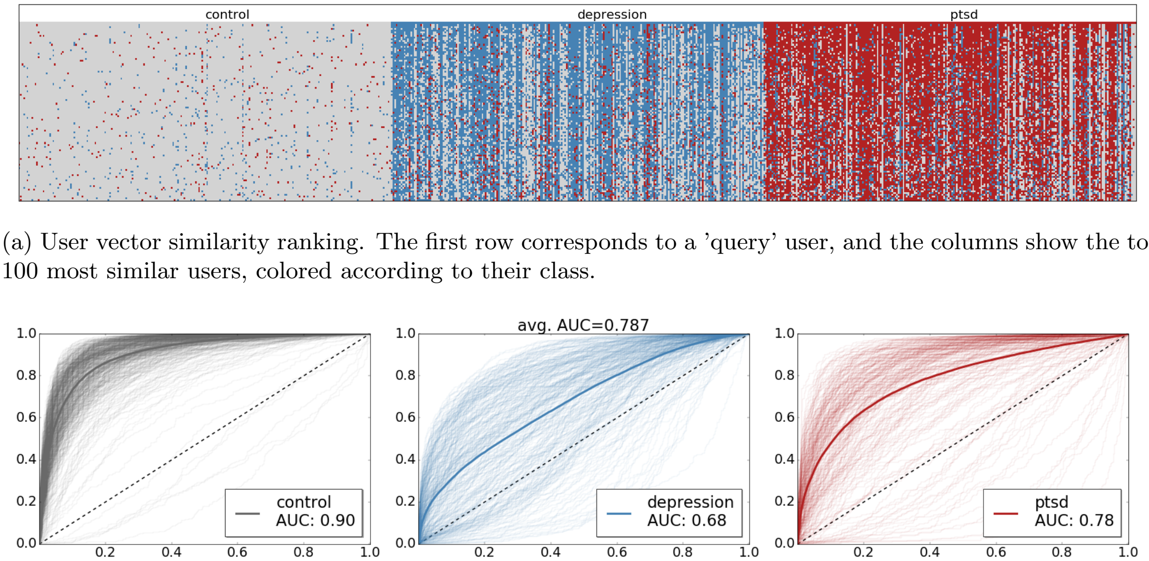
[7] Our implementation is publicly available at https://github.com/samiroid/usr2vec



(a) User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

(b) ROC curves of the induced user similarity rankings, per class. Each line represents the curve of a user and the thicker line shows the average ROC

Figure 1: Measuring homophilic relations with respect to mental conditions with vector distances over the user embedding space induced with the PV-DM model.



(a) User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

(b) ROC curves and AUC scores of the induced user similarity rankings, per class.

Figure 2: Measuring homophilic relations with respect to mental conditions via distances in the embedding subspace induced by transforming PV-DM embeddings using the NLSE model.

and using $P(\mathcal{H}_j|\mathbf{u}_j)$ as the early stopping criteria. SKIP-GRAM and *Paragraph2Vec* vectors were estimated using `Gensim` (Řehůřek and Sojka, 2010). To ensure a fair comparison, we kept the hyper-parameters consistent across all the models, which were set as follows: window size $w = 5$, negative sample size $s = 20$ and vector size $d = 400$.

Finally, we leveraged the user labels to estimate *mental-health specific* embedding subspace projections, via the NLSE model. Following Astudillo et al. (2015), we set the subspace size $s = 10$ and learning rate $\alpha = 0.01$, and induced adapted representations from each of the aforementioned user embeddings. In what follows, we will refer to these new representation as USER2VEC$_{\text{sub}}$, PV-DM$_{\text{sub}}$ and PV-DBOW$_{\text{sub}}$.

5.2 Measuring Homophily

To answer our first research question, we evaluated the embeddings in regard to their ability to capture homophilic relations with respect to mental health statuses. If the vectors of users affected with some condition are close, then they must have similar values along some dimensions, which implies that these carry information about said condition. Thus, we considered each user in the corpus in turn as a 'query' with which to retrieve similar users. We calculated cosine similarities between the query user embedding and vectors corresponding to all other users, inducing a similarity-based ranking. Intuitively, we would hope to see that individuals in the same mental health categories as the query user are comparatively similar to one another — e.g., a depressed person should be most similar to other depressed persons. The induced rankings were evaluated with respect to the Receiver Operating Characteristic (ROC) curves and the average Area Under the Curve (AUC).

All models perform significantly better than chance; the rankings induced the USER2VEC, PV-DBOW and PV-DM obtained average AUC scores of 0.56, 0.57 and 0.59, respectively. Figure 1 shows results obtained with the PV-DM vectors. The top plot shows the induced similarity rankings, where each column includes the top $k = 100$ most similar users to the query user (first row), colored according to the respective class. The bottom plot shows the respective ROC curves under the induced rankings. Similar plots for the other models are in the Appendix. These results suggest that learned user embeddings do indeed capture signals relevant to mental health, if only weakly. Nonetheless, we believe this is somewhat surprising, given that these are unsupervised, generic representations that were in no way explicitly trained to capture attributes of mental health.

We repeated the same experiment with the tuned embeddings fit using the NLSE model. The user similarity rankings induced with USER2VEC$_{\text{sub}}$, PV-DM$_{\text{sub}}$ and PV-DBOW$_{\text{sub}}$ obtained an average AUC of 0.69, 0.68 and 0.79, respectively. Figure 2 presents the results obtained with the PV-DM$_{\text{sub}}$ embeddings. These refined representations greatly improve performance overall, and particularly with respect to discriminating controls from unhealthy users, suggesting that the induced subspace captures more fine-grained signal related to mental statuses. To better understand the effect of this transformation on the representation space, we used t-Stochastic Neighborhood Embedding algorithm (Van der Maaten and Hinton, 2008) to embed the 400-dimensional PV-DM and the 10-dimensional PV-DM$_{\text{sub}}$ vectors into a 2-dimensional space. In Figure 3, we show the resulting plots, where each point denotes a user, colored by class. One can see that while the unsupervised embeddings do a reasonable job of clustering similar users, in the subspace induced by the NLSE there is a better separation between users from different cohorts.

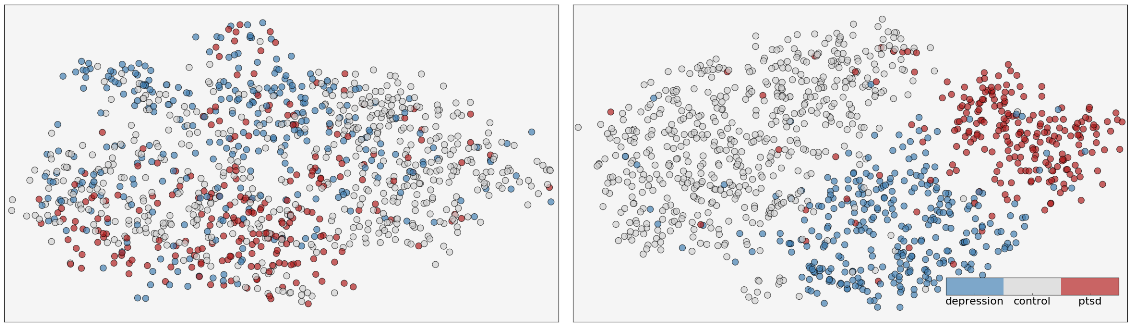


Figure 3: User embeddings projected into two dimensions, and colored according to the respective cohort. The plot shows the PV-DM embeddings on the right-hand side and PV-DM$_{\text{sub}}$ on the left-hand side.

5.3 Predicting Mental Health from Twitter Data

To address our second research question, we measured the predictive performance of the learned user embeddings in downstream mental health analysis applications. To that end, we used the CLPsych shared task data and developed models to predict users' mental health statuses, given their posting histories. Concretely, we trained Multinomial Logistic Regression classifiers leveraging the aforementioned user embeddings as features and compared their performance against baselines using textual features based on:

1. BOW: bag-of-words vectors with binary weights, $\mathbf{x} \in \{0,1\}^{|\mathcal{V}|}$;
2. BOE: bag-of-embeddings. Leveraging the SKIP-GRAM embeddings we built vectors, $\mathbf{x} = \sum_w \mathbf{E}_{[w]}$, where $\mathbf{E}_{[w]} \in \mathbb{R}^d$ is the embedding of word w;
3. LDA: bag-of-topics. We induced $t = 100$ topics using Latent Dirichlet Allocation (Blei et al., 2003), to build vectors $\mathbf{x} \in \{0,1\}^t$ indicating the topics present in user's posts;
4. BWC: bag-of-word-clusters. We induced $k = 1000$ Brown et al. (1992) word clusters, to build vectors $\mathbf{x} \in \{0,1\}^k$ mapping words in a user's posts to their respective clusters;

We also evaluated models that combine user vectors with text features (U2V+BOW/E).

Experiments were conducted using 10-fold cross-validation. For each split, the training partition was divided into 80%/20% train/validation sub-splits to facilitate hyperparameter selection and early-stopping. We used the same partitions for all models. We performed nested grid-search to choose the best ℓ_2 regularization coefficient, over the range $c = \{0.001, 0.01, 0.5, 1, 10, 100\}$, for the LR models; and the optimal subspace size $s = \{10, 15, 20, 25\}$ and learning rate $\alpha = \{0.01, 0.1, 0.5, 1\}$, for the NLSE model.

The classifiers were mainly evaluated with respect to the macro average F_1. We also report results in terms of *binary* F_1, where we only average the scores for the `depression` and `ptsd` classes. This allows us to better ascertain the ability of the models to discriminate between mentally afflicted patients, which are less prevalent than the controls, but are the cases that we mostly care about. The main classification results are shown in Figure 4. The first thing to note is that the BOW is a very strong baseline, essentially outperforming all the other linear classifiers based on textual features and generic user embeddings. One reason is that users affected with mental illnesses, often talk about their conditions and the BOW model can easily pick-up on such clues. Regarding the user embeddings, we found that, despite being similar, the PV-DBOW performed much worse than the USER2VEC, showing

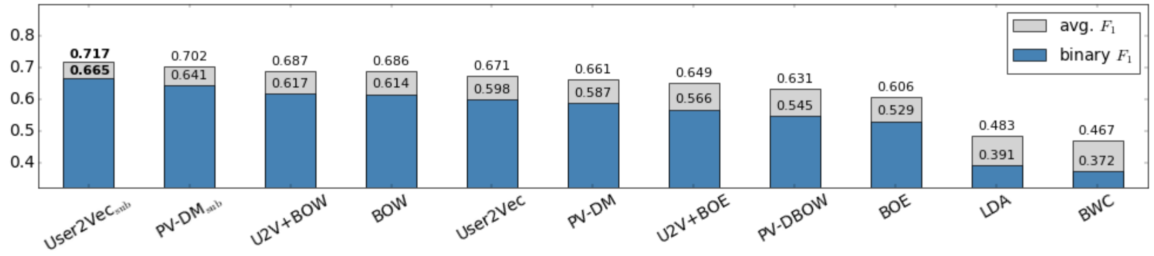


Figure 4: Predictive performance of different models.

that better embeddings can be obtained by trying to predict **all** the words in users posts, and leveraging pre-trained word vectors. On the other hand, the PV-DM model has a performance comparable to that of the USER2VEC.

As discussed above, generic embeddings are sub-optimal for downstream tasks, and our results are in line with this observation. By inducing task-specific representations via subspace projection, we were able to outperform all the other baselines by a fair margin. Note also that the NLSE approach is particularly better at discriminating the minority classes, i.e., patients with `depression` and `ptsd`, as evidenced by the greater improvements in *binary* F_1, when compared to the other baselines.

6. Conclusions

In this paper, we investigated if embeddings induced from user posting histories capture relevant signals for clinical applications. In particular, we compared different user embedding methods, with respect to their ability to capture homophilic relations between users, and their performance as features in downstream mental health prediction models. The evaluation conducted over a dataset comprising of users diagnosed with depression and PTSD, and demographically matched controls, showed that these representations can indeed capture mental health related signals. Interestingly, embeddings induced without knowledge of user labels capture similarities with respect to mental condition. This is in agreement with prior results from the field of psychology, establishing connections between word usage and mental status (Pennebaker et al., 2001). Furthermore, we have shown that these embeddings can be tailored — with a small amount of task-specific labeled data — to capture more granular information, thus improving the quality of downstream models and applications.

Ultimately, this work is a step toward more accurate inference concerning the mental health status of social media users, in turn enabling more accurate epidemiological real-time population-wide monitoring of mental health. Such accurate monitoring – which is currently impossible – may provide empirical support for increased resource allocation to programs dedicated to preventing and alleviating mental health issues. Moving forward, user embeddings may provide a pivotal piece to allow clinical psychologists to take full advantage of digital phenotyping data. In particular, learned user embeddings may provide a representation at a sweet spot between instantaneous (proximal) state and lifelong (distal) state, which is critical to understanding psychological phenomena and risk of crisis.

Acknowledgments

This work was supported in part by the Fundação para a Ciência e a Tecnologia, Portugal (FCT), through funds with reference UID/CEC/50021/2013 and Ph.D. scholarship SFRH/BD/89020/2012.

References

Silvio Amir, Byron C Wallace, Hao Lyu, and Paula Carvalho Mário J Silva. Modelling context with user embeddings for sarcasm detection in social media. In *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*, page 167, 2016.

Ramón Astudillo, Silvio Amir, Wang Ling, Mario Silva, and Isabel Trancoso. Learning word representations from scarce and noisy data with embedding subspaces. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1074–1084, Beijing, China, July 2015.

David Bamman and Noah A Smith. Contextualized sarcasm detection on twitter. In *Proceedings of the 9th International Conference on Web and Social Media*, pages 574–77. AAAI Menlo Park, CA, 2015.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *The Journal of Machine Learning Research*, 3:1137–1155, 2003.

Yoshua Bengio, Ian J. Goodfellow, and Aaron Courville. Deep learning. Book in preparation for MIT Press, 2015. URL `http://www.iro.umontreal.ca/~bengioy/dlbook`.

Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. *EACL 2017*, page 94, 2017.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

Peter F. Brown, Peter V. Desouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4), 1992.

CDC. Current depression among adults—united states, 2006 and 2008. *MMWR. Morbidity and mortality weekly report*, 59(38):1229, 2010.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537, 2011.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in Twitter. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014a.

Glen Coppersmith, Craig Harman, and Mark Dredze. Measuring post traumatic stress disorder in Twitter. In *Proceedings of the 8th International AAAI Conference on Weblogs and Social Media (ICWSM)*, 2014b.

Glen Coppersmith, Mark Dredze, Craig Harman, and Kristy Hollingshead. From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado, USA, June 2015a. North American Chapter of the Association for Computational Linguistics.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. CLPsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the Shared Task for the NAACL Workshop on Computational Linguistics and Clinical Psychology*, 2015b.

Glen Coppersmith, Kim Ngo, Ryan Leary, and Tony Wood. Exploratory data analysis of social media prior to a suicide attempt. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, San Diego, California, USA, June 2016. North American Chapter of the Association for Computational Linguistics.

Munmun De Choudhury, Emre Kiciman, Mark Dredze, Glen Coppersmith, and Mrinal 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.

Yoav Goldberg. A primer on neural network models for natural language processing. *Journal of Artificial Intelligence Research*, 57:345–420, 2016.

Geoffrey E Hinton. Learning distributed representations of concepts. In *Proceedings of the eighth annual conference of the cognitive science society*, volume 1, page 12. Amherst, MA, 1986.

Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. *arXiv preprint arXiv:1405.4053*, 2014.

Deirdre McCaughey, Catherine Baumgardner, Andrew Gaudes, Dominique LaRochelle, Kayla Jiaxin Wu, and Tejal Raichura. Best practices in social media: Utilizing a value matrix to assess social media's impact on health care. *Social Science Computer Review*, 32(5):575–589, 2014.

Jude Mikal, Samantha Hurst, and Mike Conway. Ethical issues in using twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics*, 17 (1):1, 2016.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

Margaret Mitchell, Glen Coppersmith, and Kristy Hollingshead, editors. *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. North American Association for Computational Linguistics, Denver, Colorado, USA, June 2015.

Michael J Paul and Mark Dredze. You are what you tweet: Analyzing twitter for public health. *Icwsm*, 20:265–272, 2011.

ES Paykel, A Tylee, A Wright, RG Priest, et al. The defeat depression campaign: psychiatry in the public arena. *The American journal of psychiatry*, 154(6):59, 1997.

Ted Pedersen. Screening twitter users for depression and ptsd with lexical decision lists. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 46–53, 2015.

James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

Daniel Preotiuc-Pietro, Maarten Sap, H Andrew Schwartz, and LH Ungar. Mental illness detection at the world well-being project for the clpsych 2015 shared task. *NAACL HLT 2015*, page 40, 2015.

Ashwin Rajadesingan, Reza Zafarani, and Huan Liu. Sarcasm detection on twitter: A behavioral modeling approach. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 97–106. ACM, 2015.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. `http://is.muni.cz/publication/884893/en`.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107, 2015.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Cognitive modeling*, 5(3):1, 1988.

H. Andrew Schwartz, Johannes Eichstaedt, Margaret L. Kern, Gregory Park, Maarten Sap, David Stillwell, Michal Kosinski, and Lyle Ungar. Towards assessing changes in degree of depression through Facebook. In *Proceedings of the ACL Workshop on Computational Linguistics and Clinical Psychology*, 2014.

Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 959–962. ACM, 2015.

Noah A Smith and Jason Eisner. Contrastive estimation: Training log-linear models on unlabeled data. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*, pages 354–362. Association for Computational Linguistics, 2005.

Duyu Tang, Bing Qin, and Ting Liu. Learning semantic representations of users and products for document level sentiment classification. In *ACL (1)*, pages 1014–1023, 2015.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(2579-2605):85, 2008.

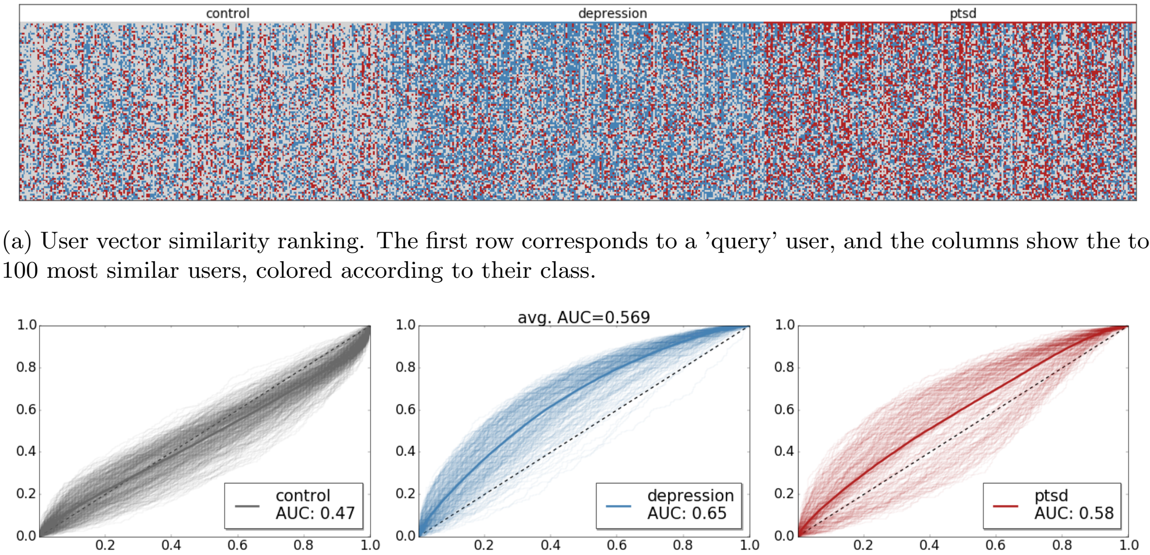
Harvey A Whiteford, Louisa Degenhardt, Jürgen Rehm, Amanda J Baxter, Alize J Ferrari, Holly E Erskine, Fiona J Charlson, Rosana E Norman, Abraham D Flaxman, Nicole Johns, et al. Global burden of disease attributable to mental and substance use disorders: findings from the global burden of disease study 2010. *The Lancet*, 382(9904):1575–1586, 2013.

Yi Yang, Ming-Wei Chang, and Jacob Eisenstein. Toward socially-infused information extraction: Embedding authors, mentions, and entities. *arXiv preprint arXiv:1609.08084*, 2016.

Yang Yu, Xiaojun Wan, and Xinjie Zhou. User embedding for scholarly microblog recommendation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 449–453, 2016.

Ye Zhang and Byron Wallace. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. *arXiv preprint arXiv:1510.03820*, 2015.

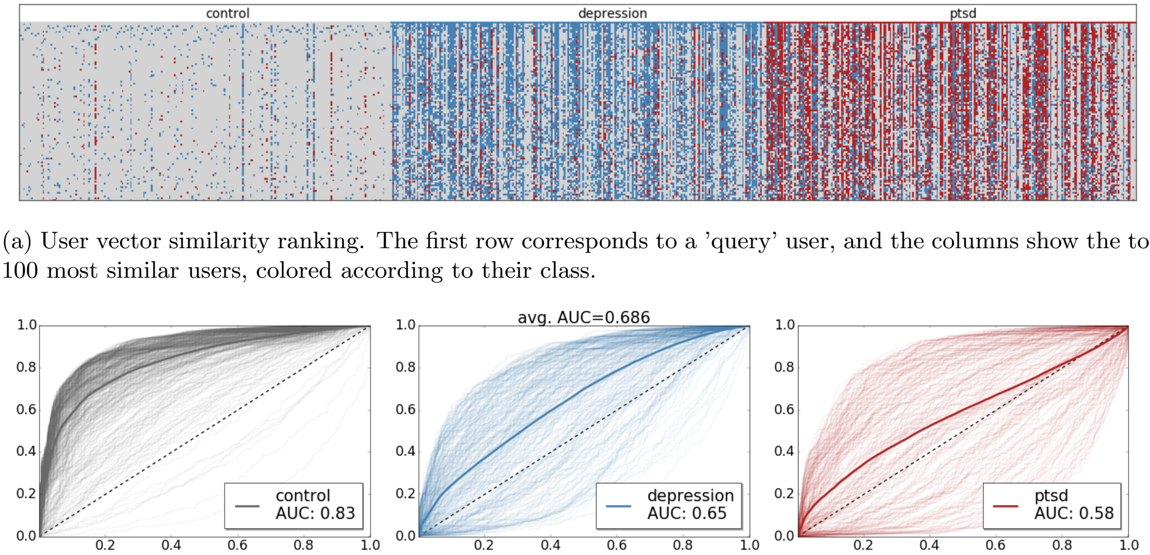
Appendix A. Measuring Homophily (continued)



(a) User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

(b) ROC curves of the induced user similarity rankings, per class. Each line represents the curve of a user and the thicker line shows the average ROC

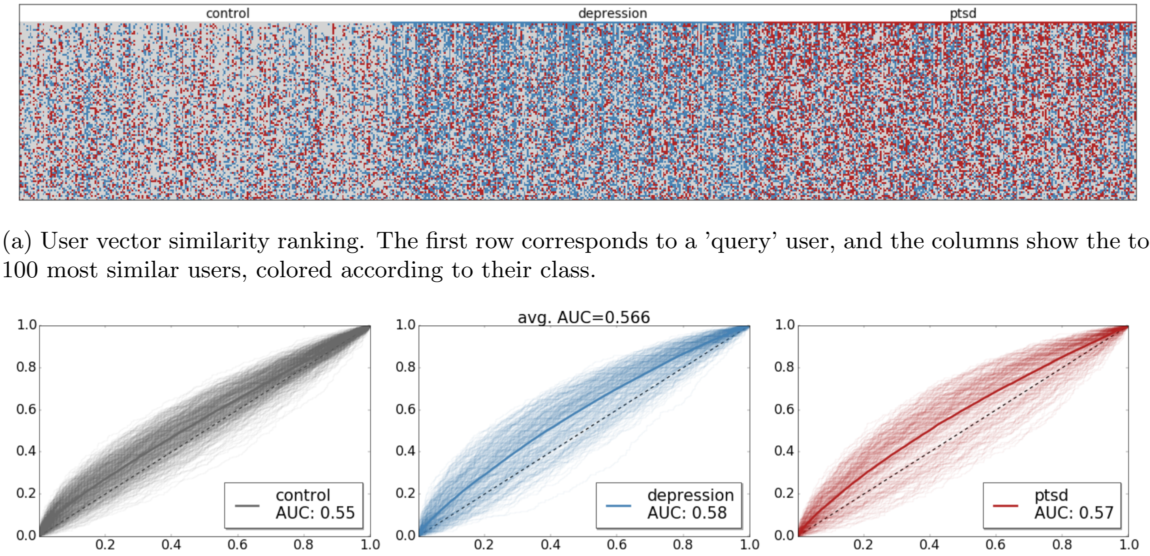
Figure 5: Measuring homophilic relations with respect to mental conditions with vector distances over the user embedding space induced with the USER2VEC model.



(a) User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

(b) ROC curves of the induced user similarity rankings, per class. Each line represents the curve of a user and the thicker line shows the average ROC

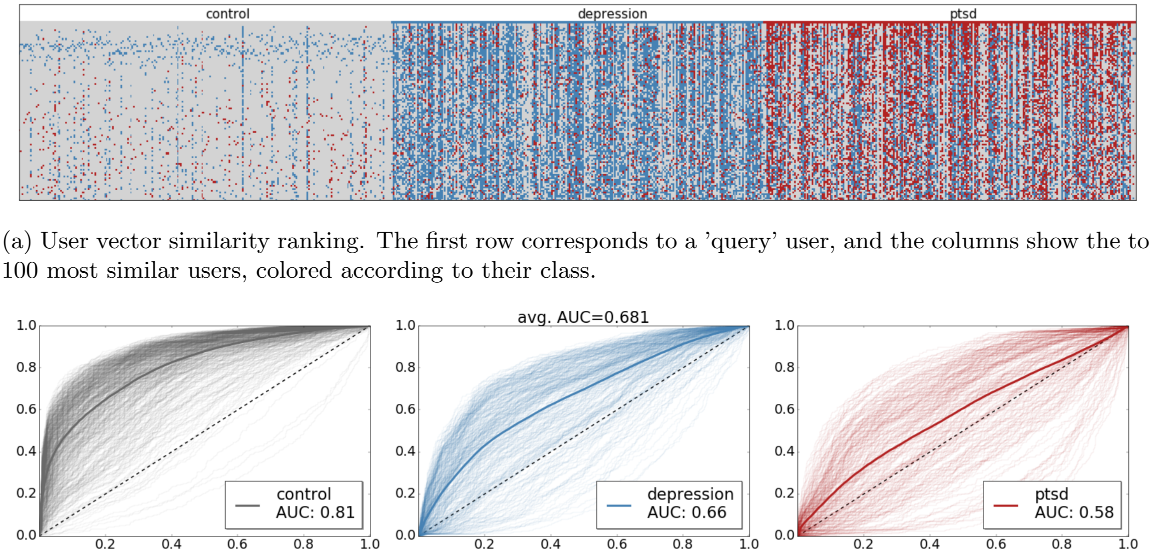
Figure 6: Measuring homophilic relations with respect to mental conditions with vector distances over the embedding subspace induced by transforming USER2VEC embeddings with the NLSE model.



(a) User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

(b) ROC curves of the induced user similarity rankings, per class. Each line represents the curve of a user and the thicker line shows the average ROC

Figure 7: Measuring homophilic relations with respect to mental conditions with vector distances over the user embedding space induced with the PV-DBOW model.



(a) User vector similarity ranking. The first row corresponds to a 'query' user, and the columns show the top 100 most similar users, colored according to their class.

(b) ROC curves of the induced user similarity rankings, per class. Each line represents the curve of a user and the thicker line shows the average ROC

Figure 8: Measuring homophilic relations with respect to mental conditions with vector distances over the embedding subspace induced by transforming PV-DBOW embeddings with the NLSE model.