Published in final edited form as:
J Biomed Inform. 2017 November ; 75 Suppl: S129–S137. doi:10.1016/j.jbi.2017.06.014.

Psychiatric symptom recognition without labeled data using distributional representations of phrases and on-line knowledge

Yaoyun Zhang, Ph.D.[1], **Olivia Zhang**[2], **Yonghui Wu, Ph.D.**[1], **Hee-Jin Lee, Ph.D.**[1], **Jun Xu, Ph.D.**[1], **Hua Xu, Ph.D.**[1], and **Kirk Roberts, Ph.D.**[1]

[1]School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX 77030 USA

[2]St. John's School, Houston, TX 77019 USA

Abstract

Objective—Mental health is becoming an increasingly important topic in healthcare. Psychiatric symptoms, which consist of subjective descriptions of the patient's experience, as well as the nature and severity of mental disorders, are critical to support the phenotypic classification for personalized prevention, diagnosis, and intervention of mental disorders. However, few automated approaches have been proposed to extract psychiatric symptoms from clinical text, mainly due to (a) the lack of annotated corpora, which are time-consuming and costly to build, and (b) the inherent linguistic difficulties that symptoms present as they are not well-defined clinical concepts like diseases. The goal of this study is to investigate techniques for recognizing psychiatric symptoms in clinical text without labeled data. Instead, external knowledge in the form of publicly available "seed" lists of symptoms is leveraged using unsupervised distributional representations.

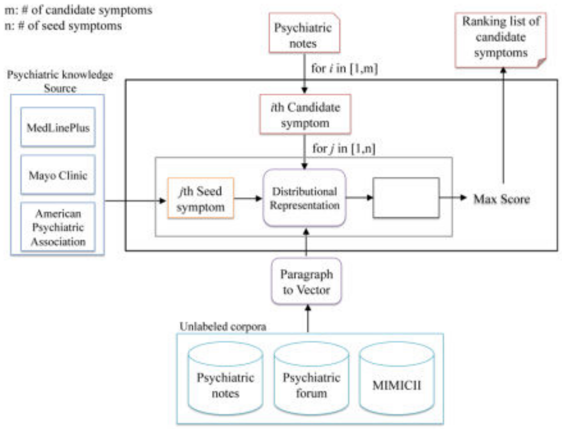
Materials and Methods—First, psychiatric symptoms are collected from three online repositories of healthcare knowledge for consumers—MedlinePlus, Mayo Clinic, and the American Psychiatric Association—for use as seed terms. Candidate symptoms in psychiatric notes are automatically extracted using phrasal syntax patterns. In particular, the 2016 CEGS N-GRID challenge data serves as the psychiatric note corpus. Second, three corpora—psychiatric notes, psychiatric forum data, and MIMIC II—are adopted to generate distributional representations with `paragraph2vec`. Finally, semantic similarity between the distributional representations of the seed symptoms and candidate symptoms is calculated to assess the relevance of a phrase. Experiments were performed on a set of psychiatric notes from the CEGS N-GRID 2016 Challenge.

Results & Conclusion—Our method demonstrates good performance at extracting symptoms from an unseen corpus, including symptoms with no word overlap with the provided seed terms.

Co-corresponding authors: Hua Xu, Professor, Director, Center for Computational Biomedicine, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, Phone: 713-500-3924, hua.xu@uth.tmc.edu. Kirk Roberts, Assistant Professor, Center for Computational Biomedicine, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, 7000 Fannin St, Suite 600, Houston, TX 77030, Phone: 713-500-3653, kirk.roberts@uth.tmc.edu.

Semantic similarity based on the distributional representation outperformed baseline methods. Our experiment yielded two interesting results. First, distributional representations built from social media data outperformed those built from clinical data. And second, the distributional representation model built from sentences resulted in better representations of phrases than the model built from phrase alone.

Graphical Abstract



Keywords

psychiatric symptom recognition; unsupervised learning; distributional representation; online knowledge

INTRODUCTION

Mental health includes people's emotional, psychological, and social well-being, which is becoming an increasingly important topic in healthcare.[1] Due to the lack of objective tests of many mental disorders, psychiatric practice often requires detailed, interpersonal communications and observations of patients, which is often recorded in narrative text in electronic health record (EHR) systems.[2] Psychiatric symptoms, a key piece of information in such narratives, often consist of subjective descriptions with details of the patient's experience. These are individualized descriptions that convey the nature, severity, and impact of the symptom, often in the patient's own lay terms (Figure 1). Such information is critical to support phenotypic classification for personalized prevention, diagnosis, and intervention of mental disorders.[3] Therefore, to enable quantitative analysis of symptomatic manifestations, it is important to develop automated approaches to extract psychiatric symptoms from clinical text.

However, the diverse and personalized expressions of psychiatric symptoms make it difficult to use traditional natural language processing (NLP) techniques to automatically extract such information from text. Instead of a single word or simple noun phrase, psychiatric symptoms have tremendous syntactic and semantic variability.[3] Existing clinical terminologies such as in Unified Medical Language System (UMLS), SNOMED-CT, and ICD-9 code have low coverage of such complex expressions,[3] making it infeasible to use dictionary-matching methods for psychiatric symptom recognition. Moreover, conventional supervised learning-based models widely employed for clinical concept recognition are trained from manually-curated corpora. Generating such corpora is an expensive and time-consuming process, and the results are often not generalizable to other clinical texts with different structures (making supervised approaches even more difficult for psychiatric texts, which lack any universal organization). Furthermore, symptoms in clinical notes of different types of psychiatric disorders (e.g., depression vs. personality disorders) and of different populations and environments (e.g., children, military veterans, college students) may have their own sub-languages. The diversity and sparity of psychiatric symptoms would thus require a vastly larger labeled corpus compared to concepts like diseases and medications.

Therefore, NLP techniques using unsupervised or minimally supervised methods bear a great potential for psychiatric symptom recognition, especially if existing psychiatric knowledge could be leveraged to alleviate extra manual work. Fortunately, there are many mental health symptom lists readily available in online healthcare knowledge repositories (e.g., MedlinePlus, Mayo Clinic, American Psychiatric Association). Such information is provided for consumers and usually expressed in lay languages. This is often similar to the way psychiatric symptoms are stated in clinical notes, and could thus prove useful in automatic symptom recognition. Further, symptom lists are easy to manage and customize to different sub-populations and psychiatric disorder types, resulting in a system that requires little human supervision. Based on the assumption that candidate symptoms with a high semantic similarity with known psychiatric symptoms are more likely to be positive symptoms, the known psychiatric symptoms could be used as “seeds” to identify similar symptoms from psychiatric notes. To further address the diversity and sparity problems of psychiatric symptoms, distributional representations of flexible-length text (e.g., phrase embeddings) instead of individual words (word embeddings) can be generated from unlabeled corpora.[4] Distributional semantic models have been successfully applied in various semantic similarity tasks due to their ability to generalize and overcome data sparsity issues.[5,6]

Therefore, in this study, we propose an unsupervised framework for psychiatric symptom recognition from unlabeled clinical notes that combines distributional representation of phrases with “seed” symptoms taken from online knowledge sources. Distributional representations are first constructed using the `paragraph2vec` (paragraph to vector) model[4] in an unsupervised manner on a large unlabeled corpus. Both the candidate and seed symptoms are then represented as fixed-length vectors inferred from the models.

Next, their semantic similarity is calculated using the cosine similarity between their vector representations. From these similarities, symptoms can be classified using some similarity threshold, or the similarities can be integrated into a traditional supervised system. *This paper concentrates entirely on the process of extracting a high-quality candidate list of symptoms.* The second step - classifying individual mentions in their context - is left to future work. Evaluation, therefore, focuses on the quality of ranking a list of context-free symptoms. Our experimental evaluation demonstrates that our proposed method is promising, greatly outperforming baseline methods.

BACKGROUND

NLP for psychiatric notes

Recently emerging research activities have used natural language processing (NLP) techniques to unlock information in psychiatric text in EHRs for various applications. For example, Pestian et al. used NLP features and semi-supervised machine learning methods to discriminate between the conversation of suicidal and non-suicidal individuals.[7] Patel et al. used NLP techniques to identify cannabis use that was documented in free text clinical records.[8] Further, Rumshisky et al. used features generated from the Latent Dirichlet Allocation (LDA) model to enhance the accuracy of predicting early psychiatric readmission.[9] McCoy et al. also used NLP features extracted from discharge summaries and regression models to predict the risk of suicides.[10]

To the best of our knowledge, the only study of psychiatric symptom recognition using NLP-based methods is by Gorrell et al. for negative symptom recognition of schizophrenia.[2] Similar to our work, they also point out that it is time-consuming and expensive to annotate a psychiatric corpus of sufficient size for supervised learning methods. In their approach, they applied active learning to mitigate this problem.[2] However, domain experts still need to be heavily involved, and it is difficult to scale the supervised framework to symptoms of other types and other mental disorders. In contrast, seed symptoms from external psychiatric knowledge sources and unsupervised distributional representations of short text are employed in this study, making our framework labor-alleviated, flexible and scalable to large-scale unlabeled corpus.

Symptom NLP

The semantic type of ‘sign or symptom’ is part of the UMLS semantic group of disorders{Lindberg, 1993 #64}, which are essential concepts in biomedical domain. Many automated clinical concept annotation tools such as MetaMap[11], MedLEE[12], cTAKES[12,13], and CLAMP[14] are capable of recognizing disorders from clinical text. Moreover, many biomedical shared tasks are devoted to disorder recognition, including i2b2 2010[15], ShARe/CLEF eHealth task 2013[16], SemEval 2014 task 7[17], and SemEval 2015 task 14[18]. However, the vase majority of disorders in these tasks are not symptoms, and only a few existing works focus specifically on symptom recognition. Matheny et al.[19] developed rule-based algorithms using keywords and SNOMED-CT concepts for infectious disease symptoms using clinical narratives. Cater and Matthew[3] annotated subjective symptom expressions in clinical notes from the Department of Veterans Affairs. Based on their comparisons, among the 543 subjective expressions, only two of them were coded using ICD-9-CM, and only 45.3% instances of subjective expressions were restated in semantically related clinical terms. They also found that it is necessary to develop NLP techniques to extract the diverse symptom expressions unobtainable by other automated methods. Roberts et al.[20] recognize coronary artery disease (CAD) risk factors, demonstrating that CAD symptoms are among the most difficult risk factors to recognize due to the diversity of phrasing. They use a lexicon-based model combined with a binary classifier, yet this method will fail to recognize semantically identical symptoms that have minor lexical differences to those in the pre-built lexicon.

Semantic similarity

One of the key components of our work is semantic similarity calculation between short texts. Semantic similarity of short texts (such as phrases, sentences, or paragraphs) has been widely applied in many tasks such as paraphrase recognition, textual entailment, information retrieval, and question answering.[21] Many approaches have been suggested based on measurements of different linguistic levels and their combinations, ranging from lexical matching, handcrafted patterns, syntactic parse trees, knowledge-based methods using external sources of structured semantic knowledge, and corpus-based methods such as distributional semantics. However, lexical features, such as string matching using edit distance, do not capture semantic similarity beyond a trivial level. Hand-crafted patterns on the other hand, are not flexible and scalable to unseen data.[21] Furthermore, approaches depending on full parse trees are restricted to syntactically well-formed texts, typically of one sentence in length, which are not suitable for psychiatric symptoms or clinical text.

Knowledge-based approaches utilize knowledge sources such as dictionaries, taxonomies, and semantic networks, and include path-finding measures and intrinsic information content (IC) measures[22,23]. On the other hand, the corpus-based approaches utilize the distribution and co-occurrence of terms or concepts within a corpus to compute similarity[24], such as vector-space-model (VSM), pointwise mutual information (PMI), latent semantic analysis (LSA) and neural network based word embedding[25]. However, given that most of the knowledge-based and corpus-based approaches are mainly designed for word-to-word (or concept-to-concept) semantic similarity, multiple methods are proposed to calculate the semantic similarity of two short text segments (such as sentences, tweets etc.) by summing up the maximum similarities between words from different text segments, respectively[21,23]. Word-to-word similarity is also represented as various features to build supervised learning models using large-scale annotation corpora for paraphrases and machine translation. One challenge of using word-to-word similarity as the proxy for short-text similarity is that the word ordering information is missing, which plays an important role in forming the semantic structure of short text. To address this problem, Le and Mikolov[4] propose Paragraph Vector (`paragraph2vec`), an unsupervised algorithm that learns fixed-length distributed representations from variable-length pieces of texts, such as sentences, paragraphs, and documents. `paragraph2vec` directly generates paragraph embeddings by expanding the widely used neural network based model of word embeddings (`word2Vec`). Experimental evaluations on sentiment analysis and information retrieval tasks demonstrated that `paragraph2vec` outperformed the conventional models of bag-of-words significantly.

Currently, knowledge-based and corpus-based approaches are the two major methods for semantic similarity of concepts in the biomedical domain[26]. However, existing standard corpora of semantic similarity in the biomedical domain are mainly focused on individual concepts or terms [23], few works are devoted to phrase-level semantic similarity yet. Moreover, as mentioned previously, lexicons of clinical concepts cannot cover the subjective symptom expressions in a sufficient and clinical meaningful way, making it infeasible to use well-structured clinical taxonomies for psychiatric symptom detection. To overcome this limitation, this study employs a corpus-based approach to detect psychiatric symptoms from clinical notes. Specifically, the `paragraph2vec` model is used to generate distributed representations of phrases and calculate the semantic similarity between candidate symptoms and seed symptoms collected from online medical knowledge portals.

Resource evaluation for distributional representation

Due to the fact that there is no publicly available, large corpus of psychiatric notes, external resources are necessary for building the distributional representations. To evaluate the effects of external resources, Roberts[27] assessed the performance of two standard clinical NLP tasks (the i2b2 2010 concept and assertion tasks) using word embedding based features generated from six different corpora, including i2b2, MMIC, MEDLINE, WebMD, Wikipedia and Gigaword. The main findings from this study are that combinations of corpora are generally found to work best, and that the single-best corpus is generally task-dependent.[27] Another work by Pakhomov et al.[25] constructed neural network representations of single-word clinical terms found in a publicly available benchmark dataset manually labeled for semantic similarity and relatedness between pairs of disorders, symptoms and drugs. Similarity and relatedness measures computed from text corpora in three domains (Clinical Notes, PubMed Central articles and Wikipedia) were compared using the benchmark as reference. Although it is found that measures computed from full text of biomedical articles in PubMed Central are on par with measures computed from clinical reports, the comparison also demonstrate that measures from Wikipedia are worse than sources from the biomedical domain.

Similar with the work of Pakhomov et al., this study also evaluated and compared different resources (i.e., psychiatric notes, psychiatric forum text, MIMIC II) for their effects of building distributional representations. However, this study differs from their work in three-respects: (1) distributional representations of short text are constructed, instead of words; (2) the evaluation task is semantic similarity between pairs of psychiatric symptoms that include multi-word phrases, in contrast to single-word clinical concepts; (3) The experimental results from this study demonstrate that social media data – psychiatric forum text – make positive contributions for the underlined task, while the clinical corpus decreased the performance. One possible reason could be that the effects of different corpora are task-dependent, as also demonstrated in previous studies.[27]

MATERIALS AND METHODS

Study design

As illustrated in Figure 2, we present a framework for psychiatric symptom recognition that combines unsupervised distributed representation learning with seed symptoms collected from publicly available psychiatric knowledge repositories. Distributional representations of variable-length text are first constructed using the `paragraph2vec` (paragraph to vector) model on a large unlabeled corpus.[4] Both the candidate and known symptoms are then represented as fixed-length vectors inferred from the models. Next, their semantic similarity is calculated using the cosine similarity between their vector representations. From these similarities, highly possible psychiatric symptoms can be selected.

Dataset

Unlabeled psychiatric corpus—The psychiatric notes provided by the CEGS N-GRID 2016 challenge organizers are used for experimental evaluation in this study[*]. This is the first corpus of mental health records released to the NLP research community, which contains about 1,000 initial psychiatric evaluation records. An initial psychiatric evaluation record was produced by a psychiatrist in order to elicit psychiatric signs and symptoms, disorders, and other medical conditions in order to decide the course of treatment.

Knowledge sources for seed symptoms—The psychiatric symptoms employed as the knowledge source for unsupervised symptom recognition in this study are extracted from three major online repositories of health care information for consumers: MedlinePlus, the Mayo Clinic, and the American Psychiatric Association. First, webpages were retrieved by searching the query “mental disorders” from (1) the health topics and linked Medical Encyclopedia provided in MedlinePlus[†]; (2) the “diseases-conditions” category provided by the Mayo Clinic[‡]; (3) and American Psychiatric Association[§]. The semi-structured webpages were downloaded and parsed to obtain the listed symptoms under the “Symptom” subtitle. Since some metal disorders could also be symptoms of the other disorders (e.g., depression is a symptom of suicide), the mental disorder names themselves are also incorporated as symptoms in this study. In total, 876 psychiatric symptoms are collected from these three sources.

Distributed representation of phrases

The *paragraph2vec* algorithm— `paragraph2vec` is an unsupervised algorithm that learns fixed-length distributed vector representations (embeddings) from variable-length pieces of texts, such as sentences, paragraphs, and documents.[4] `paragraph2vec` is an expansion of word2vec, which constructs distributed vector representations of words (word embeddings)[28]. `paragraph2vec` generates the embeddings of variable-length pieces of text together with word embeddings. The embeddings are produced based on a language modeling task of predicting a word (W_t) given its context (word context: W_{t-k}, …, W_{t+k}, and paragraph context: d_j), and the object is to maximize the average prediction probability as shown in Eq. (1). To construct the paragraph vector, contexts of fixed-length are sampled from a sliding window over the paragraph, and the length of contexts could be tuned to adjust for different tasks. In this way, the paragraph vector represents the missing information from the current context and can act as a memory of the topic of the paragraph. The task of word prediction is illustrated in Figure 3. The paragraph vectors and word vectors are trained together using stochastic gradient descent and the gradient is obtained via backpropagation. At every step of stochastic gradient descent, a fixed-length context is sampled from a random paragraph, which is used to compute the error gradient from the network in Figure 3 and use the gradient to update the parameters in the model. At prediction time, an inference step is conducted to compute the paragraph vector for a new

[*] https://www.i2b2.org/NLP/RDoCforPsychiatry/
[†] https://medlineplus.gov/mentalhealthandbehavior.html
[‡] http://www.mayoclinic.org/diseases-conditions
[§] https://www.psychiatry.org/patients-families

paragraph, by doing gradient descent on the paragraph vector with fixed parameters in the rest of the model.

$$\frac{1}{\mathrm{T}}\sum_{\mathrm{t=k}}^{\mathrm{T-k}}\log p\left(\mathrm{w_t}|\mathrm{w_{t-k}},\ldots,\mathrm{w_{t+k}},\mathrm{d_j}\right) \quad (1)$$

$$\mathrm{p}\left(\mathrm{w_t}|\mathrm{w_{t-k}},\ldots,\mathrm{w_{t+k}},\mathrm{d_j}\right)=\mathrm{e}^{\mathrm{y_{wt}}}/\sum_{\mathrm{i}}\mathrm{e}^{\mathrm{y_i}} \quad (2)$$

$$\mathrm{y=b+Uh(w_{t-k},\ldots,w_{t+k},d_j;W,D)} \quad (3)$$

Granularity of variable-length text—In `paragraph2vec`, fixed length vectors of real values (embeddings) are generated for variable-length text. In this study, we tried to compare the performance of using both sentences and phrases as “paragraphs”, in order to examine the effects of documents of different granularities to generate the semantic representations for short text.

We experimented with two embedding models. In the first model, each phrase acted as the “paragraph” (i.e., each paragraph given to the model was a phrase of at most a few words). In the second model, each sentence acted as the “paragraph”. Note that for both models, we query the model with just the phrase (the inference step). We never actually obtain a sentence vector. The intuition is that sentences, being longer, provide more semantic context. Thus the phrase vector obtained from the sentence model could, under this hypothesis, be more semantically consistent than a phrase vector obtained from the phrase model. An illustration of the training and inference process of the `paragraph2vec` algorithm is shown in Figure 4.

Resources of unlabeled corpus—Three corpora are utilized to construct the paragraph2vec representations: (1) the psychiatric notes provided by the CEGS N-GRID 2016 challenge organizers[1], (2) psychiatric forum data collected from WebMD[**], and (3) MIMIC II intensive care records.[29]

Candidate symptom scoring

Three types of phrases – verb phrases (VP), noun phrases (NP) and adjective phrases (ADJP) present in psychiatric notes are considered as candidate symptoms, denoted as $C=\{c_i\}$. They are extracted using the shallow syntactic information, which is more reliable for clinical text than full syntactic parses. The psychiatric symptoms collected from online knowledge repositories are considered as the seed symptoms, denoted as $S=\{s_j\}$. After representing both c_i and s_j as vectors through the embedding model generated by

[**] http://www.webmd.com/

paragraph2vec, their cosine similarity sim_{ij} is calculated. The highest similarity score yielded by each c_j and the most similar seed symptom is kept as its final similarity score $fsim_i$ (i.e., $fsim_i = \max(sim_{ij})$, j ∈ [1, |S|]). A ranked list of candidate symptoms is generated using $fsim_i$, for the convenience of later evaluation of this method.

Evaluation

For evaluation, 130 clinical notes were randomly selected from the training dataset of the CEGS N-GRID 2016 challenge, and psychiatric symptoms were manually annotated from the narrative text in the “chief complaints”, “formulation” and “patient history” sections of the clinical notes as the gold standard. Examples of symptoms in each section are listed in Figure 4. We developed an annotation guideline and recruited two annotators, who manually annotated all the psychiatric symptoms mentions in each note by following the guideline. First, 20 psychiatric notes were annotated by the two annotators, and the kappa value between them was 0.70. A domain expert manually reviewed these 20 notes and resolved the disagreements between the two annotators. Problems present in this initial annotation were noted in the guideline, based on which the second round of annotation was conducted. In total, 3, 742 symptoms were annotated. The limited number of annotated psychiatric symptoms is insufficient for training a supervised symptom (for perspective, the i2b2 2010 data[14] contains over 72 thousand concepts), but it should be sufficient to evaluate an unsupervised method that incorporates external knowledge.

As for parameter configuration of the paragraph2vec model, the sliding window size was set to four, based on a pilot study which evaluated the window sizes in the range of three to eight on a small set of candidate symptoms. In addition, the other parameters of the model include: (1) the distributed memory algorithm was adopted to train the model; (2) the dimensionality of embedding vectors was 50 when the model was trained only on the corpus of psychiatric notes and 300 when the corpora of forum and MIMICII were also incorporated; (3) the initial learning rate (i.e., alpha) was 0.025; (4) and all the words with total frequency lower than 5 (i.e., min_count) were ignored.

Candidate symptoms are ranked by their maximum similarity. Intuitively, a better ranking implies better semantic similarity, as well as higher performance for downstream methods using our approach. The performance of the candidate symptom list ranking is evaluated using *nDCG* (normalized discounted cumulative gain) as formalized in Eq. (4)[30], where the unique candidate symptoms are weighted by their partial overlap likelihood with the gold-standard symptoms (e.g., if “depression” overlaps a symptom annotation 34 of 46 times in the corpus, its “gold” score is 0.739). *nDCG* is a widely employed evaluation criterion in Information Retrieval (IR)[30]. The intuition of using *nDCG* is that it outputs measurements in a normalized [0,1] scale, and that it leverages real-valued weights to evaluate the partially-matched phrases.

$$nDCG_p = DCG_p / IDCG_p, \quad (4)$$

where $DCG_p = (\sum_{i=1}^{p} 2^{rel_i} - 1)/log_2(i+1)$ and $IDCG_p = (\sum_{i=1}^{|REL|} 2^{rel_i} - 1)/log_2(i+1)$

p: the chosen position in the ranking list. The complete list of ranked candidate symptoms are used in this study for *nDCG* calculation.

rel_i: percentage of the *i*th candidate symptom present within the manually annotated symptoms overall occurrences.

REL|: the list of relevant candidates up to position *p*.

In addition, another evaluation criterion commonly used in IR – precision of the top K ranked candidates (P@K, K= 5, 10, 100, 200, 1000) - is also adopted for evaluation. Instead of assigning weights to partially matched candidates as in *nDCG*, we only consider exactly matched candidates as positive symptoms for the P@K.

As explained in the introduction section, the purpose of this paper is to extract high-quality candidate list of symptoms, based on semantic similarity with seed symptoms. Therefore, we use the *nDCG* and precision@k as evaluation metric, to measure the quality of ranking a list of context-free symptoms, instead of F-measure for typical named entity recognition systems.

The following experiments are conducted to assess our approaches:

1. sentences versus phrases as the "paragraphs" for constructing embeddings models,
2. different corpora combinations for building the embeddings,
3. comparison with a weak baseline (random ordering), and relatively strong baselines (dictionary-lookup based on the list of seed symptoms and cosine similarity with TF-IDF vectors).

RESULTS

Coverage of phrase-based candidate symptoms over manual annotations

As mentioned in the Materials and Methods Section, three types of phrases - NP, VP and ADJP- are selected as candidate symptoms. This was designed to maximize recall: most symptoms could be in such a phrase, but the vast majority of phrases are not symptoms. Therefore, the coverage of the candidate symptoms over the manually-annotated symptoms is evaluated using recall. Moreover, since the string boundaries of candidates and manual annotations may not be matched exactly, both the recall of exact matches and partial matches are reported. In total, there were 14,943 candidates and 3,742 manually annotated symptoms. As illustrated in Table 1, the percentage of exact matches and partial matches (excluding extract) are 42.36% and 54.06%, respectively. The overall recall of 98.71% indicates that using phrases as candidates has a sufficient coverage of symptoms in the dataset.

Moreover, we further looked into pairs of matched candidates and symptoms. Table 2 illustrates several examples of exactly matched and partially matched candidate-symptom pairs. For exact match (42.36%) and the case of partial match that the symptom is a substring of the candidate phrase (26.11%), the candidate contains the complete information of the symptom. In terms of the case that the candidate is a substring of the symptom (27.95%) or others (2.30%), it can be seen from Table 2 that the modifiers of symptoms with more specific information (e.g., “self” in the symptom “self - injurious behaviors”) are not present in the candidate phrases. However, key information of symptoms is still partially maintained in the candidates.

Table 3 lists the experimental performance of our method. Overall, our method outperformed the random ordering list baseline (0.696) and the TF-IDF-based cosine similarity baseline (0.745). Interestingly, social media data, i.e., psychiatric forum text, makes positive contributions to symptom ranking upon the psychiatric notes (0.741 vs. 0.784; 0.734 vs. 0.736), while the use of the much larger MIMIC II data severely reduced performance (0.784 vs. 0.768; 0.736 vs. 0.728). Further, the distributed representations using sentence embeddings outperformed phrase embeddings. For example, using the combined corpus of psychiatric notes and forum text, the sentence embedings achieved a *nDCG* of 0.784, while the phrase embeddings gained an *nDCG* of 0.736. Again, the vector representation is obtained from phrase - not the sentence that contains the phrase. The sentence embeddings model only used sentences to build the model. By analogy to supervised symptoms, the model was trained on sentences, but tested on phrases.

Table 4 lists the experimental performance of P@K of using the dictionary-lookup baseline method, the TF-IDF baseline method, and our proposed approaches. Overall, the performance follows a similar trend as that of *nDCG*. For the embedding based approaches, the P@5 and P@10 are consistent to be 100.0% and 80.0%, respectively, while using sentence embeddings produced by the combined corpus of psychiatric notes and forum text achieved the optimal P@100, 200 and 1000.

To observe the differences between the ranked candidate psychiatric symptoms generated by our approach and TF-IDF, top-ranked candidates (ranks 1–10 and 101–110) are listed in Table 5. The top ranked candidates generated by our approach contain more diverse and accurate psychiatric symptoms. In comparison, the top ranked candidates generated by the TF.IDF based semantic similarity contain more psychiatric disorders (despite that they are also considered as symptoms in our current standards) and are relatively noisy with trivial words.

High ranked candidates with no lexical overlap with the most semantically similar seed symptoms

Given that there are 14,943 candidate psychiatric symptoms, in contrast to the much smaller set of seed collection with 876 symptoms currently, it is inevitable that a majority part of candidates do not have lexical overlap with seed symptoms. Therefore, we examined the high ranked candidates of no lexical overlap with the most semantically similar seed symptoms to examine if positive symptom phrases could be identified by our approaches. Table 6 lists some examples yielded from the sentence embeddings generated using the combined corpus of psychiatric notes and forum text, which obtains the optimal performance (ref. Table 3). As illustrated in Table 6, the seed symptoms are capable of identifying positive symptoms from the candidates without any lexical overlap, such as “binged”, “her decreased appetite”, etc.

DISCUSSION

This study investigated techniques for recognizing psychiatric symptoms in clinical text without labeled data, by using seed lists of symptoms available from public knowledge sources and unsupervised distributional representations. Experimental performance demonstrated that our proposed method using semantic similarity based on distributional representation outperformed baseline methods, and yielded good performance at extracting symptoms from an unseen corpus, including symptoms with no word overlap with the provided seed terms.

When looking into the false positive candidates of high semantic similarity scores with the seeds, we found that one type of error is caused by matching a general phrase with a seed symptom, especially when the phrase is a substring of the seed such as the candidate “feel” matched with the seed symptom “feel alone” or the candidate “generalized” matched with the seed symptom “generalized anxiety disorder”. Another type of error is present between candidates of certain semantic relations with the seed symptoms, including treatments of psychiatric disorders such as drugs (e.g., “zoloft 150 mg”), common disorders that frequently co-occur with psychiatric disorders (e.g., “his brain hemmorage”), adjectives modifying the psychiatric symptoms (e.g., “escalating”), and stressors having cause-effect relations with the symptoms (e.g., “the incident”, “family obligations”), etc. Thus, filtering to remove general phrases or of certain semantic types may help to improve the performance.

In comparing the different unlabeled corpora for generating the distributional representations of phrases, one interesting finding is that social media data, i.e., psychiatric forum text, contributed the most to symptom ranking, while the use of the much larger MIMIC II data severely reduced the performance. (Table 3) One potential reason for this could be that the sub-language used in psychiatric forums is more similar with that of free text in clinical notes in terms of psychiatric symptoms. Symptoms are often expressed in the patient’s own words (i.e., consumer language) which should be similar to online forum language. In contrast, the MIMIC II corpus mainly contains clinical notes from ICUs, which is not suitable for this task despite of its huge volume and successful use cases in clinical concept recognition such as disorders.

In terms of the different granularities of document representation in `paragraph2vec`, experimental results demonstrate that distributed representations using sentence embeddings outperformed phrase embeddings. This is unexpected because we use phrases as candidate symptoms. One potential reason is that embedding vectors obtained from sentences encode more context information. In contrast, phrases contain limited context information to maintain implicit semantic similarities, especially for semantic similarity calculation between pairs of strings without any lexical overlap.

Limitation and Future Work

Given the novelty of using phrase embeddings for identifying symptoms of clinical notes, we acknowledge an important limitation of this work. Namely, few experiments were done selecting the best embedding model(s) from a given corpus. The evaluation section states the parameters used for the embedding model, but other parameters could easily have resulted in better embedding vectors. Further, the randomness involved in training the embedding model means that re-training the same model on the same data may have resulted in better (or worse) embeddings. An ensemble similarity measure that averages out these differences may lead to more robust results. We do not expect, and we certainly do not hope, that changes to the embedding creation process would lead to significant differences in the scores seen in Table 3, but future work is necessary to investigate this possibility.

Another limitation of our current study is the insufficient amount of seed symptoms collected for unsupervised psychiatric symptom recognition. Other knowledge resources such as DSM[31] and WebMD will be incorporated in the future. In addition, the effects of extra social media corpora such as Twitter will be investigated next to build distributional representation models. Furthermore, based on the above error analysis, the scope of candidate symptoms could be filtered first by lexicon or semantic type constrains to enhance the precision. Finally, additional studies to combine our current framework with supervised learning methods will also be conducted.

CONCLUSION

Psychiatric symptom recognition from clinical notes is an important task for computational applications concerning psychiatric disorders. This study proposed an unsupervised learning framework for psychiatric symptom recognition. Experimental evaluation indicates that our proposed method was promising. The proposed unsupervised learning framework could also be generalizable to other tasks of medical concept recognition.

Acknowledgments

Funding: This work was supported by the National Institutes of Health [grant numbers 5R01LM010681-05, 1R01GM102282-01A1, 4R00LM012104-02]. The CEGS N-GRID challenge was supported by NIH grants P50MH106933, 4R13LM011411.

References

1. Proctor EK, Landsverk J, Aarons G, Chambers D, Glisson C, Mittman B. Implementation research in mental health services: an emerging science with conceptual, methodological, and training challenges. Administration and Policy in Mental Health and Mental Health Services Research. 2009; 36(1):24–34. [PubMed: 19104929]
2. Gorrell G, Jackson R, Roberts A, Stewart R. Finding negative symptoms of schizophrenia in patient records. Proc NLP Med Biol Work (NLPMedBio), Recent Adv Nat Lang Process (RANLP). 2013:9–17.
3. Carter M, Matthew Samore M. “Sitting on Pins and Needles”: Characterization of Symptom Descriptions in Clinical Notes. 2013
4. Le, QV., Mikolov, T. ICML 2014. 2014. Distributed Representations of Sentences and Documents; p. 1188-96.

5. Guo, J., Che, W., Wang, H., Liu, T. Revisiting embedding features for simple semi-supervised learning. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP); 2014; 2014. p. 110-20.
6. Dai AM, Olah C, Le QV. Document embedding with paragraph vectors. 2015 arXiv preprint arXiv: 150707998.
7. Pestian JP, Grupp-Phelan J, Bretonnel Cohen K, et al. A controlled trial using natural language processing to examine the language of suicidal adolescents in the emergency department. Suicide and life-threatening behavior. 2015
8. Patel R, Wilson R, Jackson R, et al. Cannabis use and treatment resistance in first episode psychosis: a natural language processing study. The Lancet. 2015; 385:S79.
9. Rumshisky A, Ghassemi M, Naumann T, et al. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. Translational Psychiatry. 2016; 6(10):e921. [PubMed: 27754482]
10. McCoy TH, Castro VM, Roberson AM, Snapper LA, Perlis RH. Improving prediction of suicide and accidental death after discharge from general hospitals with natural language processing. Jama psychiatry. 2016; 73(10):1064–71. [PubMed: 27626235]
11. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. Journal of the American Medical Informatics Association. 2010; 17(3):229–36. [PubMed: 20442139]
12. Savova GK, Masanz JJ, Ogren PV, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. Journal of the American Medical Informatics Association. 2010; 17(5):507–13. %@ 1527-974X. [PubMed: 20819853]
13. Friedman C, Alderson PO, Austin JHM, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association. 1994; 1(2):161–74. %@ 1527-974X. [PubMed: 7719797]
14. Kogan, Y., Collier, N., Pakhomov, S., Krauthammer, M. Towards semantic role labeling & IE in the medical literature. AMIA Annual Symposium proceedings/AMIA Symposium AMIA Symposium; 2005; p. 410-4.
15. Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. Journal of the American Medical Informatics Association. 2011; 18(5): 552–6. [PubMed: 21685143]
16. Suominen, H., Salanterä, S., Velupillai, S., et al. Overview of the ShARe/CLEF eHealth evaluation lab 2013. International Conference of the Cross-Language Evaluation Forum for European Languages; 2013; Springer; 2013. p. 212-31.
17. Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., Savova, G. Semeval-2014 task 7: Analysis of clinical text. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014); 2014; 2014. p. 54-62.
18. Elhadad, N., Pradhan, S., Chapman, W., Manandhar, S., Savova, G. SemEval-2015 task 14: Analysis of clinical text. Proc of Workshop on Semantic Evaluation Association for Computational Linguistics; 2015; 2015; p. 303-10.
19. Matheny ME, FitzHenry F, Speroff T, et al. Detection of infectious symptoms from VA emergency department and primary care clinical documentation. International journal of medical informatics. 2012; 81(3):143–56. [PubMed: 22244191]
20. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. Journal of Biomedical Informatics. 2015; 58:S111–S9. [PubMed: 26122527]
21. Islam A, Inkpen D. Semantic similarity of short texts. Recent Advances in Natural Language Processing V. 2009; 309:227–36.
22. McInnes, BT., Pedersen, T., Liu, Y., Melton, GB., Pakhomov, SV. U-path: An undirected path-based measure of semantic similarity. AMIA Annual Symposium Proceedings; 2014: American Medical Informatics Association; 2014; p. 882
23. McInnes, B., Liu, Y., Pedersen, T., Melton, G., Pakhomov, S. Umls:: similarity: Measuring the relatedness and similarity of biomedical concepts. Association for Computational Linguistics; 2013.

24. Islam A, Inkpen D. Semantic text similarity using corpus-based word similarity and string similarity. ACM Transactions on Knowledge Discovery from Data (TKDD). 2008; 2(2):10.

25. Pakhomov SV, Finley G, McEwan R, Wang Y, Melton GB. Corpus domain effects on distributional semantic modeling of medical terms. Bioinformatics. 2016; 32(23):3635–44. [PubMed: 27531100]

26. Lee MC, Chang JW, Hsieh TC. A grammar-based semantic similarity algorithm for natural language sentences. The Scientific World Journal. 2014

27. Roberts K. Assessing the Corpus Size vs. Similarity Trade-off for Word Embeddings in Clinical NLP. ClinicalNLP 2016. 2016:54.

28. Collobert, R., Weston, J. A unified architecture for natural language processing: Deep neural networks with multitask learning. The 25th international conference on Machine learning; 2008; ACM; 2008. p. 160-7.

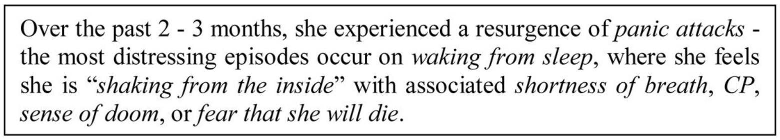
29. Saeed M, Villarroel M, Reisner AT, et al. Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database. Critical care medicine. 2011; 39(5): 952. [PubMed: 21283005]

30. Kanoulas, E., Aslam, JA. Empirical justification of the gain and discount function for nDCG. Proceedings of the 18th ACM conference on Information and knowledge management; 2009; ACM; 2009. p. 611-20.

31. Greenberg, G. The book of woe: The DSM and the unmaking of psychiatry. Penguin: 2013.

Highlights

- One of the initial studies to extract symptoms from psychiatric notes
- An unsupervised learning framework for psychiatric symptom recognition
- Leverage online consumer information and large-scale unlabeled corpora for semantic similarity of psychiatric symptoms



Over the past 2 - 3 months, she experienced a resurgence of *panic attacks* - the most distressing episodes occur on *waking from sleep*, where she feels she is "*shaking from the inside*" with associated *shortness of breath*, *CP*, *sense of doom*, or *fear that she will die*.

Figure 1.

An example paragraph from psychiatric notes with symptoms. The psychiatric symptoms are highlighted in italic.

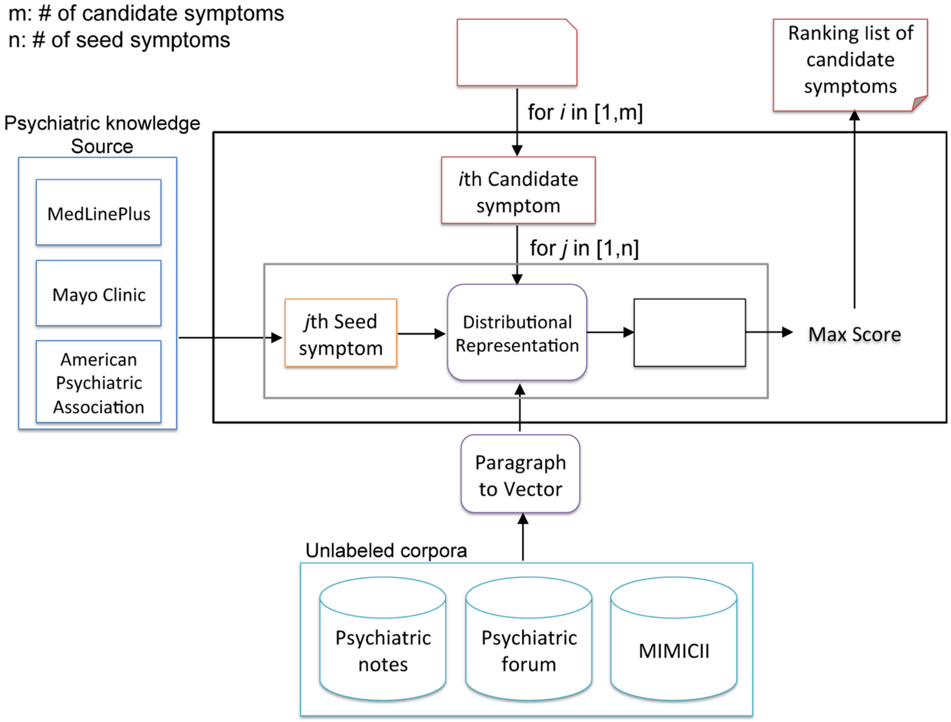


Figure 2.

Study design of psychiatric symptom recognition system.

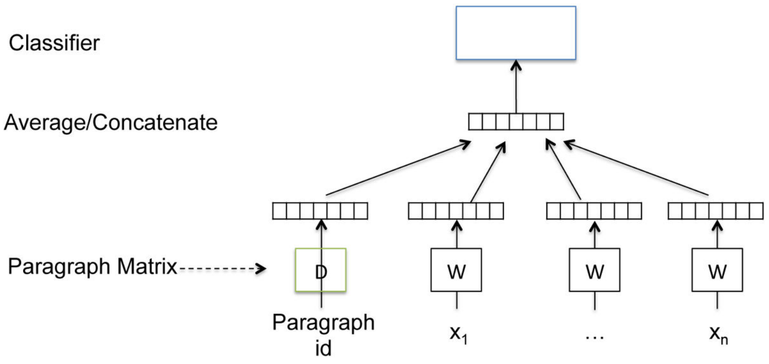


Figure 3.

A framework for learning paragraph vector. In addition to word vectors W as in the word2vec model, a paragraph token is mapped to a vector via matrix D. In this model, the concatenation or average of this vector with a context of n words is used to predict the (n +1)th word.

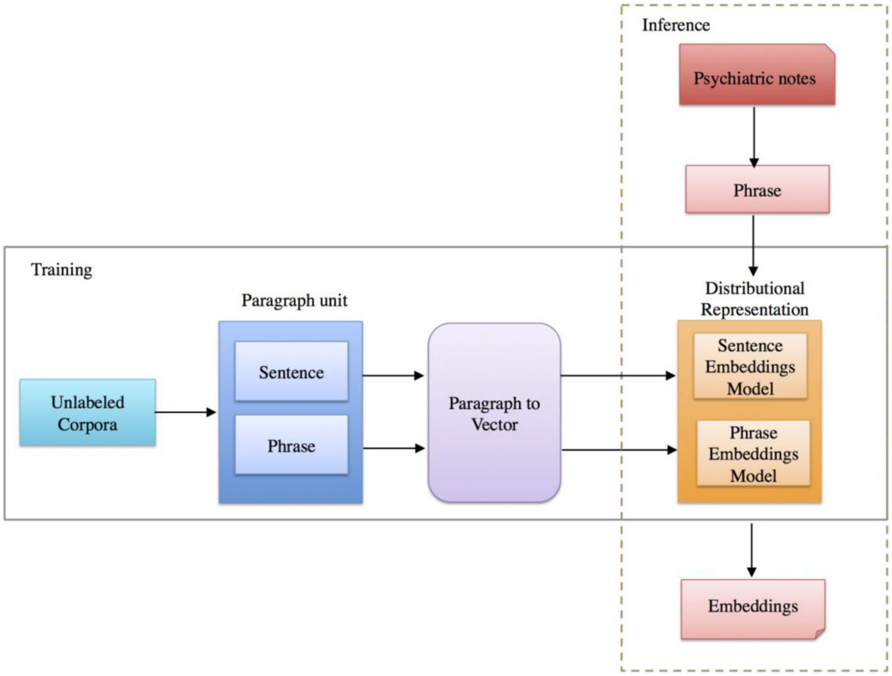
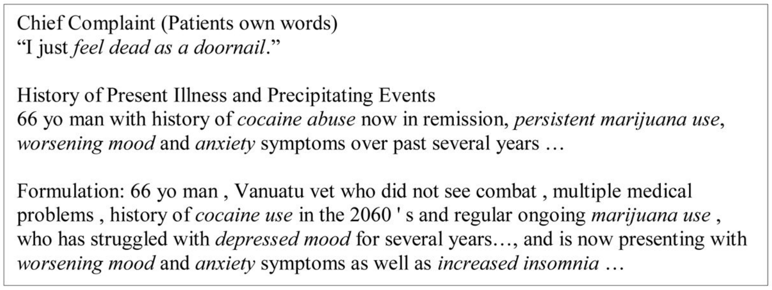


Figure 4.

An illustration of the training and inference processes of the paragraph2vec model. During the training process, sentences are used to generate the sentence embeddings model, and phrases are used to generate the phrase embeddings model, respectively. During the inference process, embeddings of phrases in psychiatric notes are generated based on the sentence and phrase embeddings models, respectively.



Chief Complaint (Patients own words)
“I just *feel dead as a doornail*.”

History of Present Illness and Precipitating Events
66 yo man with history of *cocaine abuse* now in remission, *persistent marijuana use*, *worsening mood* and *anxiety* symptoms over past several years …

Formulation: 66 yo man , Vanuatu vet who did not see combat , multiple medical problems , history of *cocaine use* in the 2060 ' s and regular ongoing *marijuana use* , who has struggled with *depressed mood* for several years…, and is now presenting with *worsening mood* and *anxiety* symptoms as well as *increased insomnia* …

Figure 5.

Examples of psychiatric symptom annotation in different sections of psychiatric notes. The psychiatric symptoms are highlighted in italic.

Table 1

Coverage of 14,943 candidate symptoms over 3,742 manual annotations. The set of candidate symptoms consist of adjective, verbal and noun phrases in psychiatric notes (%).

Category	**Number**	**Recall**
Exact match	**1,585**	**42.36**
Partial match	**2,109**	**54.06**
Phrase ⊃ Symptom	977	26.11
Phrase ⊂ Symptom	1,046	27.95
Others	86	2.30
Exact & Partial match	3,694	**98.71**

Table 2

Examples of the exact partial matches between candidate symptoms and gold standard symptoms

Exact match	
Symptom	Phrase
an underlying depressive disorder	an underlying depressive disorder
generalized anxiety	generalized anxiety
panic	panic
prior inpatient psychiatric hospitalizations	prior inpatient psychiatric hospitalizations
binge eating	binge eating
Phrase ⊃ Symptom	
Symptom	Phrase
depression	depression during periods
substance use	Discussed substance use
anxiety	may help reduced anxiety
his worried feelings	help manage his worried feelings
minimal anxiety	minimal anxiety and anger
Phrase ⊂ Symptom	
Symptom	Phrase
his substance use issues	issues
self - injurious behaviors	injurious behaviors
difficulties completing long or more mentally taxing tasks	difficulties

Table 3

Experimental performance of *nDCG* for candidate psychiatric symptoms ranking by using sentence and phrase embeddings based sematic similarity. The employed corpora - psychiatric notes, psychiatric forum postings, and MIMICII - for embedding generation are integrated incrementally.

Method		***nDCG***
Baseline	Random	0.696
	Dictionary	0.734
	TF-IDF	0.745
Sentence embeddings model	notes	0.741
	+forum	**0.784**
	+mimic	0.768
Phrase embeddings model	notes	0.734
	+forum	0.736
	+mimic	0.728

Table 4

Experimental performance of P@K for candidate psychiatric symptoms ranking by using sentence and phrase embeddings based sematic similarity. The employed corpora - psychiatric notes, psychiatric forum postings, and MIMIC II - for embedding generation are integrated incrementally. (%)

Method		**P@5**	**P@10**	**P@100**	**P@200**	**P@1000**
Baseline	Dictionary	**100.0**	**100.0**	45.0	26.0	13.0
	TF-IDF	80.0	80.0	52.0	35.5	17.6
Sentence embedding	notes	**100.0**	80.0	49.0	32.55	17.4
	+forum	**100.0**	80.0	**61.0**	**45.5**	**26.3**
	+mimic	**100.0**	80.0	53.0	37.5	18.0
Phrase embedding	notes	**100.0**	80.0	44.0	27.0	12.9
	+forum	**100.0**	80.0	51.0	33.5	20.0
	+mimic	**100.0**	80.0	52.0	31.5	16.1

Table 5

Comparison between top (ranks 1–10 and 101–110) candidate psychiatric symptoms generated by using our approach and TF-IDF for semantic similarity calculation.

Rank	Sentence embedding +forum	TF-IDF
1	bipolar disorder	obsessive compulsive disorder
2	poor concentration	bipolar disorder
3	intrusive memories	a head
4	postpartum depression	bipolar 1 disorder
5	dry mouth	is easily distracted
6	self harm	a major depressive episode
7	agoraphobia	poor concentration
8	chills	fatigue
9	difficulty breathing	dry mouth
10	hot flashes	intrusive memories
101	ptsd and depression	an
102	anxiety and pain	to
103	mania	suicide attempts
104	persistent headache	teen
105	alcohol abuse	reported avoidance
106	tired	his shoulder
107	dying	pain control
108	hopelessness	a distressing pain
109	fatigue and/or headaches	significant avoidance
110	chronic fatigue and pain	disorder

Table 6

Examples of high ranked candidate symptoms, which have no lexical overlap with the most semantically similar seed symptoms. Positive candidate symptoms are highlighted in bold.

Candidate	Most similar seed	Score
the incident	hopelessness	0.879
psychotic depression	eating disorders	0.801
affect instability	suicide	0.783
tenex (1 mg	Episodes of violence	0.774
binged	Agitation or excitability	0.770
shallow breathing	muscle tension	0.769
disturbed sleep	Gender dysphoria	0.768
fat	Autism spectrum disorder	0.756
his brain hemmorage	Thoughts of death or suicide	0.750
chemotherapy	Cocaine intoxication	0.748
increased depression and anxiety	Chest pain	0.745
physiologically hypersensitive	Feeling inadequate, inferior or unattractive	0.741
family obligations	Ongoing feelings of emptiness	0.741
frequent sleep disturbances	muscle tension	0.740
zoloft 150 mg	Excessive irritability, aggressive behavior	0.731
uncomfortable	Feeling alone	0.721
hopeless	Loss of energy, fatigue	0.718
did not work	Histrionic personality disorder	0.717
hallucinations	Unstable and intense relationships	0.717
selective mutism	Has problems playing or working quietly	0.708
her decreased appetite	muscle tension	0.704
escalating	Changing normal routine, including eating or sleeping patterns	0.704