Detecting psychological distress in adults through transcriptions of clinical interviews

Joana Correia[1,2,3], Isabel Trancoso[2,3], and Bhiksha Raj[1]

[1] Language Technologies Institute, Carnegie Mellon University, USA,
[2] Instituto Superior Técnico, Universidade de Lisboa, Portugal,
[3] Spoken Language Systems Laboratory, INESC-ID, Portugal

Abstract. Automatic detection of psychological distress, namely post-traumatic stress disorder (PTSD), depression, and anxiety, is a valuable tool to decrease time, and budget constraints of medical diagnosis. In this work, we propose two supervised approaches, using global vectors (GloVe) for word representation, to detect the presence of psychological distress in adults, based on the analysis of transcriptions of psychological interviews conducted by a health care specialist. Each approach is meant to be used in a specific scenario: online, in which the analysis is performed on a per-turn basis and the feedback from the system can be provided nearly live; and offline, in which the whole interview is analysed at once and the feedback from the system is provided after the end of the interview. The online system achieves a performance of 66.7% accuracy in the best case, while the offline system achieves a performance of 100% accuracy in detecting the three types of distress. Furthermore, we re-evaluate the performance of the offline system using corrupted transcriptions, and confirm its robustness by observing a minimal degradation of the performance.

1 Introduction

Affective disorders such as post-traumatic stress disorder (PTSD), depression, and anxiety are common disorders that require specialized health care professionals to diagnose and assess them. However, the availability of such professionals might not always exist or even be affordable. An alternative is to use computerized system that perform automatic detection of psychological distress. Though these systems could not fully replace a health care specialist, they could be used as a valuable screening tool or to aid in diagnosis, thus reducing the time and budget requirements to perform a diagnosis.

Over the last few decades there has been significant interest in studying the descriptors of psychological distress. Some verbal descriptors of psychological distress include aggregate dialogue-level features, and features that represent subject behavior during specific moments in the dialogues show significant differences in subjects with depression and PTSD, compared to non-distressed subjects [1]. Furthermore, it has been showed that the prevalence of certain words such a "I" is correlated with the presence of depression, as well as the prevalence

of some function words [2]. Even in social media platforms, some forms of psychological distress, such as depression, can be detected through rule-based data mining approaches and analysis of the use of emotions, interaction with others and other behaviours [3] [4]. However, the descriptors of psychological distress can vary across cultures and languages [5].

In this work, we focus on the analysing the word content and verbal descriptors of psychological distress. We hypothesise that such cues are different for subjects suffering from psychological distressed (PTSD, depression and anxiety), compared to non-distressed subjects. To test this hypothesis, we propose two new methods that analyse a corpus of transcriptions of clinical interviews to adults that might suffer from psychological distress.

Our main contribution is the introduction of two new systems that take advantage of global vectors (GloVes) [6] for word representation to perform a diagnosis regarding the presence of psychological distress in a subject. The first proposed system performs a per-turn analysis of the interview, which allows its use in an online setting, providing near live feedback to the health care specialist. The second system analyses the interview as a whole, and attributes a "healthy" or "distressed" connotation to the interview. Since GloVe models are capable of mapping words into a space where the distance between words reflects their semantic similarity (closer words are more similar), we can take advantage of this property to deal with new words in the test data, by looking to its neighbours in the model, and inferring its meaning. Furthermore, simple arithmetic operations with GloVes, such as summing, averaging, etc. yield a meaningful result in the same space.

The rest of the document is organized as follows: in Section 2 we introduce the corpus we used in this work; in Section 3 we introduce the proposed systems; in Section 4 we present our experiments and results, and finally, in Section 5, we draw the main conclusions and propose possible future work.

2 Corpus

The Distress Analysis Interview Corpus (DAIC) [7] consists of video recordings of semi-structured clinical interviews. In this work, we use a subset of DAIC, consisting of interviews conducted by a health care specialist, face-to-face with the subject. The study sample is a group of adults and is biased towards participants who have been diagnosed with PTSD, depression, or anxiety at some point in their lives.

The subset of the corpus being used contains recordings of interviews from 65 subjects, with each interview spanning from 30 to 60 minutes. In total, this translates to roughly 25 hours of recordings and 300 000 words uttered by the patients. Each subject turn can range from a single word to several sentences, and on average, a turn has 22 words. We discard the contributions of the interviewer and focus only on the patient side of the conversation.

The presence, and severity of psychological distress, in this case, PTSD, depression or anxiety, in the subjects was accessed using standard clinical screening measures:

- Post-traumatic stress disorder checklist-civilian (PCL-C): a widely used 5-point scale self-report measure that evaluates 17 PTSD criteria [8]
- State/Trait Anxiety Inventory (STAI): a self-report questionnaire used to help differentiate anxiety from depression [9]
- Patient Health Questionnaire-Depression 9 (PHQ-9): a ten-item self-report measure based directly on the diagnostic criteria for major depressive disorder in the DSM-IV [10]

The incidence of each form of distress in the subjects being studied is reported in Table 1. We note that each form of distress is not exclusive in a subject, for example, a subject can suffer from depression and PTSD simultanesouly.

Table 1. Incidence of each form of distress in the subjects of the DAIC subset

	Form of distress		
	PTSD	*Depression*	*Anxiety*
Healthy	43	38	27
Distressed	22	27	38
Total	65	65	65

Our subset of the DAIC includes high quality audio, video and depth sensor recordings of all the interactions, as well as manual transcriptions and annotations.

3 Detection of psychological distress with GloVe

There are two settings in which automatic detection of psychological distress might be useful: the first one is online, for example, during an interview with health specialist. It should be useful to have a system providing live feedback to the health care specialist of the analysis of the subject in terms of the presence of psychological distress. It would be an additional tool to help the health care specialist steer the interview. However, given the online nature of the system,it would be limited to analysing short segments of the interview at a time, for instance, the most recent turn from the subject. The second one is offline, for example, after the clinical interview is finished. In this scenario, the health care specialist could refer back to the system and obtain feedback from the analysis of the whole interview. In this scenario, the system would have access to the whole interview, which would allow a more in depth analysis, at the sacrifice of live feedback.

In this work, we introduce two systems, which take advantage of the GloVe for word representation framework, that implement an online system and an

offline system that detect psychological distress from the analysis of psychological interviews. In the case of the online system, it analyses one patient turn at a time to provide a diagnosis regarding the presence of cues related to psychological distress in that turn. The offline system analyses the whole interview at once to provide a diagnosis regarding the presence of psychological distress in the subject that was interviewed.

The GloVe framework is described in more derail in 3.1 two systems are described in further detail in Sections 3.2 and 3.3.

3.1 GloVe for word representation

GloVe is a word representation model know for its ability to map words into a meaningful space where the distance between words is related to their semantic similarity [6].

It is a global log-bilinear regression model used to learn vector space representations of words [6]. The model can be estimated as follows: Given a very large corpus, collect word co-occurrence statistics in a form of word co-occurrence matrix X. Each element X_{ij} of such matrix represents measure of how often word i appears in context of word j. The context of a word is defined as a predefined number of words before and after it. For each word pair, the following soft constraint is defined:

$$w_i' w_j + b_i + b_j = log(X_{i,j}) \tag{1}$$

where w_i is the vector of the main word, w_j is the vector of the next word, b_i and b_j are scalar biases for the main and context words, respectively. The modelis obtained by minimizing the cost function:

$$J = \sum_{i=1}^{V} \sum_{j=1}^{V} f(X_{i,j})(w_i' w_j + b_i + b_j - log(X_{i,j})) \tag{2}$$

where $f(.)$ is a weighting function used to prevent learning only the most common pairs:

$$f(X_{ij}) = \begin{cases} (\frac{X_{ij}}{x_{max}})^a & if X_{ij} < x_{max} \\ 1 & otherwise \end{cases}$$

3.2 Online detection of psychological distress with GloVe

To achieve a system that performs a per-turn analysis of the interview, we use the GloVe framework to compute a vector for each turn. The resulting vectors are used as the feature vector to train a binary classifier that detects the presence of a given form of psychological distress.

In more detail, assuming we have a GloVe model already trained according to Section 3.1, given a turn from an interview, consisting of n words, we retrieve the GloVes for each word to obtain $\mathbf{w} = w_1, w_2, ..., w_n$, where each $w_i \in \mathbb{R}^D$, where D is the dimensionality of the GloVe model. Since the distance between

GloVes is a measure of the semantic similarity of their corresponding words, we can compute a meaningful turn vector, t, as the average of the GloVes for the words of the turn:

$$\mathbf{t} = \frac{\sum_{i=1}^{n} w_i}{n} \tag{3}$$

The set of turn vectors from the train corpus are used to train a support vector machine (SVM) with with radial basis function (RBF) kernel.

3.3 Offline detection of psychological distress with GloVe

This work proposes a second system that uses GloVe for word representation, that takes in account the transcription of the whole interview at once to perform the diagnosis. The proposed system relies on determining the tendency of word being more correlated with healthy or distressed people, which we call "connotation" of a word.

Given a word in the test corpus, its connotation is computed by finding the words in the training corpus that are most semantically similar to the test word, and then computing whether these neighbour words are more correlated with healthy or distressed people. This approach is particularly useful when dealing with new words in the test data that were not present in the train data, since we can infer their connotation from their neighbours regardless if they were seen during previously.

Given a train corpus of N words and a vocabulary size of V, we compute the connotation for each word, $i = 1, ..., V$, in the vocabulary as follows:

$$f_H(i) = \frac{\#w_H(i)}{N_H} \tag{4}$$

$$f_D(i) = \frac{\#w_D(i)}{N_D} \tag{5}$$

$$\Delta f(i) = f_H(i) - f_D(i) \tag{6}$$

$$c_{train}(i) = sign(\Delta f(i)) \tag{7}$$

where $f_H(i)$ and $f_D(i)$ are the relative frequencies of the word i in the healthy and distressed sub-corpora, respectively; $\#w_H(i)$ and $\#w_D(i)$ are the counts of the word i in the healthy and distressed sub-corpora, respectively; N_H and N_D are the total counts of words in the healthy and distressed sub-corpora, respectively; $\Delta f(i)$ is the relative frequency for word i; and $c_{train}(i)$ is the binary label that corresponds the "healthy" or "distressed" connotation to the word i in the training corpus.

At test time, the connotation of a test word t is computed assuming a GloVe model, as described in Section 3.1, is available. For a test word t, present in the

GloVe model, we retrieve its GloVe, $\mathbf{w}_t \in \mathbb{R}^D$ and compute the similarity to all the GloVes for the words in the train vocabulary $\mathbf{w}_v$, for $v = 1, ..., V$:

$$s(\mathbf{w}_t, \mathbf{w}_v) = \left|\mathbf{w}_t.\mathbf{w}'_v\right| \tag{8}$$

Computing the similarity allows us to find the top K most similar train words, which can include the test word itself, in case it was present in the training data. Each of the top K words has a connotation associated to is, as computed in Eq. 7.

To find the connotation of a test word t, $c_{test}(t)$, we compute the average connotation of its K closest neighbours:

$$c_{test}(t) = \frac{\sum_{k=1}^{K} c_{train}(k)}{K} \tag{9}$$

For $K = 1$, the model corresponds to a Naive Bayes model with flat distribution, and as K increases the distributed smooths out.

The connotation for a whole interview is obtained by averaging the connotation of each word in it. Finally, the interview connotation scores are used to train an SVM with linear kernel, which in a 1D problem translates to finding the optimal scalar decision threshold.

4 Experiments and results

In all the experiments performed in this work we used the DAIC subset described in 2 to train and validate the performance of the models proposed in Sections 3.2 and 3.3. The corpus was split into training and validation, by randomly assigning 55 interviews for training and the remaining 10 for validation. There is no overlap between speakers in training and validation sets. All the experiments are performed three times, to evaluate the presence of each form of distress: PTSD, depression and anxiety.

We also used a pre-trained GloVe for word representation model of dimensionality 50, trained from a twitter corpus of 2 billion tweets, 27 billion tokens, and a vocabulary of 1.2 million words.

First, we evaluate the online system described in Section 3.2. The pre-trained GloVe model was used to obtain the turn representation vectors of all the turns in the corpus, computed as the mean of the word representation vectors for each turn. The vectors from the training set of the corpus were used to train an SVM with RBF kernel. The model was validated with the remaining turn representation vectors.

The performance of the online system using GloVe, for each form of distress are summarized in Table 4. The performance is reported in accuracy.

The performances obtained by this approach was generally poor across all forms of psychological distress. To some extent, a parallel can be drawn between the performance of the systems and of a human: it is hard to assess whether someone sounds depressed or suffering from PTSD from just a few words, without

Table 2. Performance in accuracy of the online system used to detect PTSD, depression, and anxiety

Approach	**Form of distress [acc.]**		
	PTSD	*Depression*	*Anxiety*
online GloVe	0.567	0.533	0.667

any temporal dependency in the model. Nevertheless, the results show that the easiest form of distress to capture is anxiety.

In the case of the offline system proposed in Section 3.3, the training interviews were used to build the relative frequency difference table. Each word from the validation interviews was used to find the 20 most similar words to compute the connotation of each test word. Using the 20 neighbours allows for a smother model, compared to a smaller number. Should the number of neighbours used to compute the connotation of a test word be 1, and the model would correspond to a Naive Bayes with flat distribution.

The classifier was trained in a leave on out fashion. Since there were only 10 interviews left for testing only 9 interview scores were used to train the classifier at a time. Due to time constraints and the heavy nature of the approach, it was impossible perform cross validation of the model. The main consequence of this compromise is the increase of variance in the results.

The performance for the offline approach was measured in accuracy and is presented in Table 4.

Table 3. Performance in accuracy of the online systems for detecting PTSD, depression, and anxiety

Approach	**Form of distress [acc.]**		
	PTSD	*Depression*	*Anxiety*
offline GloVe	1.000	1.000	1.000

We can see that the proposed model achieves a perfect classification score for all labels. However, it is important to note that since there were only 10 validation interviews there is a significant variance in the reported results. Nevertheless, the system performs remarkably well, showing the added value of taking in account longer periods of information at a time. Given the exceptional performance of this method, we refrained from benchmarking it against other state-of-the-art methods, since their performance would always be comparable or worse. Establishing a parallel with humans, this would resemble how a health care specialist pays attention to the interview as a whole to perform a diagnosis.

The conditions present in this dataset - handmade transcriptions - are not always available, for time and budget constraints. To replace those, we can use a sub-par alternative, such as an automatic speech recognition (ASR) system. In an attempt to simulate these conditions, we corrupt our handmade transcriptions

with some noise, corresponding to mistakes from the ASR. In our approach that can be achieved by replacing a given percentage of GloVes for the words of the validation interviews by random vectors of the same dimensionality.

We experimented with simulations of increasingly worse ASRs, introducing a word corruption rate of 20%, 40% and 60%. The performance of the offline approach using the corrupted corpora is reported in accuracy in Table 4. From it, we can see that even in extremely noisy scenarios, this system is robust to noise, being able to maintain the perfect performance in almost all settings.

Table 4. Performance in accuracy of the offline system with different levels of corruption of the transcription for PTSD, depression, and anxiety

GloVe corruption rate	**Form of distress [acc.]**		
	PTSD	*Depression*	*Anxiety*
20%	1.000	1.000	1.000
40%	1.000	1.000	1.000
60%	0.800	1.000	1.000

Finally, we provide some interesting examples of words with healthy and distressed connotation, by analysing a subset of the sorted relative frequency table. In Figure 1 we can see that healthy people tend to talk more about casual subjects like school, prom; about relationships and feelings, and to laugh more than distressed people. Conversely, distressed people talk tend to talk more about traumatic events or topics, such as alcoholism, prison, and drugs, as well as to be generally more uninterested and bored by the conversations, saying "blah" and "uh" much more than healthy people. All these findings correlate well with the general perception from a human, of what is the content of distressed discourse.

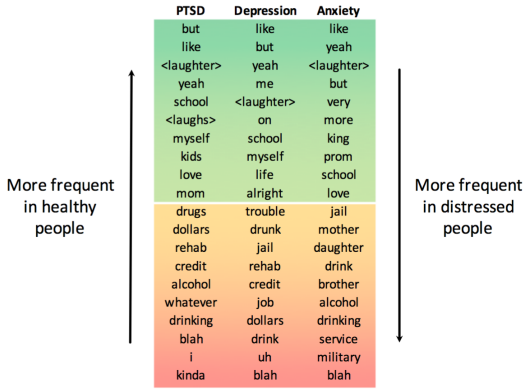

Fig. 1. Examples of words with large relative frequency difference for each label

5 Conclusions and Future work

In this work, we performed a study on the detection of three forms of psychological distress: PTSD, depression and anxiety in adults, though the analysis of transcriptions of clinical interviews. We considered two approaches to analyse the transcriptions, both using GloVe for word representation models.

In the first approach, we propose a system that can be used online, and that performs a per-turn analysis of the interview. This approach was not successful in assessing the presence of psychological distress, which meant that the systems was incapable of learning to detect meaningful cues that describe psychological distress in the short-term. Nevertheless, it showed that the easiest form of distress to capture with little information was anxiety, which might be an indicator that there are short-term descriptors of anxiety.

For the second approach, we propose a system meant to be used offline, that analyses the whole interview at once. Naturally, this approach used more information from the subject in order to perform a diagnosis, than the first one. Taking in account more information provided a significant improvement in performance. This approach was able to capture perfectly the presence of all forms of distress, obtaining an accuracy of 100% for the detection of the three forms of distress, which is a very significant result: the proposed method cannot be outperformed by any other in the conditions of this experiment. Even after corrupting the transcriptions to simulate machine made transcriptions with a variable rate of errors, the performance of the system remained the same in most of the scenarios, thus confirming the robustness of the approach. The performance of this model can be attributed to the proposed "connotation" system introduced in this work.

Finally, there are still a number of issues were left to be addressed to fully solve this research problem. As an examples, it would be useful to use a model with temporal dependencies at a turn-level analysis of the interview, to better capture the short-term cues that describe PTSD, and depression, thus maybe overcoming the poor performance our system achieved.

References

1. D. DeVault, K. Georgila, R. Artstein, F. Morbini, D. Traum, S. Scherer, A. Rizzo, and L.-P. Morency, "Verbal indicators of psychological distress in interactive dialogue with a virtual human," in *Proceedings of SIGDIAL*, 2013.
2. D. Watson and J. W. Pennebaker, "Health complaints, stress, and distress: exploring the central role of negative affectivity.," *Psychological review*, vol. 96, no. 2, p. 234, 1989.
3. X. Wang, C. Zhang, Y. Ji, L. Sun, L. Wu, and Z. Bao, "A depression detection model based on sentiment analysis in micro-blog social network," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pp. 201–213, Springer, 2013.
4. C. J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

5. N. Ramirez-Esparza, C. K. Chung, E. Kacewicz, and J. W. Pennebaker, "The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches.," in *ICWSM*, 2008.
6. J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation.," in *EMNLP*, vol. 14, pp. 1532–1543, 2014.
7. S. Scherer, G. Lucas, J. Gratch, A. Rizzo, and L.-P. Morency, "Self-reported symptoms of depression and ptsd are associated with reduced vowel space in screening interviews,"
8. E. B. Blanchard, J. Jones-Alexander, T. C. Buckley, and C. A. Forneris, "Psychometric properties of the ptsd checklist (pcl)," *Behaviour research and therapy*, vol. 34, no. 8, pp. 669–673, 1996.
9. C. D. Spielberger, "Stai manual for the state-trait anxiety inventory," *Self-Evaluation Questionnaire*, pp. 1–24, 1970.
10. K. Kroenke and R. L. Spitzer, "The phq-9: a new depression diagnostic and severity measure," *Psychiatric annals*, vol. 32, no. 9, pp. 509–515, 2002.