Depression and Anxiety Prediction Using Deep Language Models and Transfer Learning

Tomasz Rutowski
Ellipsis Health Inc
San Francisco, USA
tomek@ellipsishealth.com

Elizabeth Shriberg
Ellipsis Health Inc
San Francisco, USA
liz@ellipsishealth.com

Amir Harati
Ellipsis Health Inc
San Francisco, USA
amir@ellipsishealth.com

Yang Lu
Ellipsis Health Inc
San Francisco, USA
yang@ellipsishealth.com

Piotr Chlebek
Ellipsis Health Inc
San Francisco, USA
piotr@ellipsishealth.com

Ricardo Oliveira
Ellipsis Health Inc
San Francisco, USA
ricardo@ellipsishealth.com

Abstract— **Digital screening and monitoring applications can aid providers in the management of behavioral health conditions. We explore deep language models for detecting depression, anxiety, and their comorbidity using input from conversational speech. Speech data comprise 16k spoken interactions labeled for both depression and anxiety. We find that results for binary classification range from 0.86 to 0.79 AUC, depending on condition and comorbidity. Best performance occurs for comorbid cases. We show that this result is not attributable to data skew. Finally, we find evidence suggesting that underlying word sequence cues may be more salient for depression than for anxiety.**

Keywords—mental health, depression, anxiety, natural language processing, deep learning, transfer learning, digital health

I. INTRODUCTION

Depression and anxiety are prevalent behavioral health conditions, with significant impact on society [1] and with high rates of co-occurrence [2][3]. Therapies are available for both conditions [4], but there is a pressing need for screening, so that patients can be identified early [5][6][7]. A growing line of international studies has shown that machine-based analysis of spoken language offers promise for the remote detection and management of behavioral health conditions [8][9]. Spoken language is natural, engaging, and requires only a microphone.

Depression prediction in speech technology has been studied in particular, with good results for both language models [10][11] and acoustic models [8][9][12][13][14][15]. Advances in machine learning have been supported by common evaluations for depression, including [16][17]. There has been less work on anxiety in the technology community, most likely because there has been less availability of labeled data.

We are interested in machine learning based approaches for both depression and anxiety when information about their comorbidity is available. To this end, we introduce a corpus jointly labeled for depression and anxiety, comprising roughly 11,000 different speakers and 16,000 joint labeled sessions.

Although the corpus is currently proprietary, it is used because we found no large data public collections that contain both depression and anxiety labels [18] for spoken language data. Current data collections are either small or self-reported, with challenges involving the induction of labels based on social media statements [19][20][21][22][23].

We first describe the corpus, including the co-occurrence of depression and anxiety as a function of the level of severity for each condition. We then propose a transfer-learning based deep NLP system. The approach is used to compare performance for each condition individually, and for the conditions jointly. Finally, we define a measure of model variability. Results on this measure suggest that word-based cues may be fundamentally more salient for one condition than the other.

II. METHOD

A. Data

We use a corpus collected by Ellipsis Health. To facilitate comparative research we have initiated discussions with the Linguistic Data Consortium on a future release of portions of this corpus to the larger speech community [24]. An earlier, smaller version of the data was used to study depression in [25]. Corpus statistics are given in Table I. The data set contains roughly 16,000 sessions from over 11,000 unique speakers. Train and test partitions used for our experiments contain no overlapping speakers. There are more sessions than speakers since some users used the system multiple times; all repeat users are in the training set.

The data is comprised of American English spontaneous speech, with users allowed to talk freely in response to questions within a session. Users range in age from 18 to over 65, with a mean of roughly 30. They interacted with a software application that presented questions on different topics, such as "work" or "home life". Responses average about 175 words. Users responded to 4-6 (mean 4.52) different questions per session, resulting in session durations that averaged just under six minutes each.

During each session, the user completed a PHQ-8 [26] (PHQ-9 after the suicidality question was removed), and a GAD-7. These serve as the depression and anxiety gold standard labels for speech within a session. (Since sessions are only a few minutes, it is assumed users do not change state within them.) For both PHQ-8 and GAD-7 scores were mapped for binary classification. Following [27], scores at or above 10 mapped to presence of the condition; scores below 10 were mapped to absence of the condition.

	Total	Train-	Train+	Test-	Test+
PHQ-8					
Responses	72,369	41,558	16,277	11,395	3,139
Sessions	15,950	9,266	3,606	2,425	653
GAD-7					
Responses	72,369	42,662	15,173	11,539	2,995
Sessions	15,950	9,538	3,334	2,460	618

B. NLP System

Learning word representations has been well known for many years. One of the most relevant approaches was based on building the representation in a vector space that enabled to group similar words together. Initially, in our research we used word embedding [28][29] and SVM classification algorithm. Another direction was to use a transfer learning approach [30] the entire deep learning model was used for classification purposes. The main NLP model is trained from scratch on multiple, primarily publicly available, data sources such as Wikipedia [31]. Various methods can be used for tokenization purposes [32][33]. We use the spaCy tokenization library; each word is represented by a unique ID. Our word dictionary contains over 20,000 individual tokens. This study is primarily based on LSTM topology inspired by the AWD-LSTM architecture [34] of the core language model and ULMFiT [35] work for model fine tuning. The approach in [34] is based on the following contributions: DropConnect for hidden to hidden layers which differs from the drop out approach by deactivating certain weights rather than activation mechanisms. Another beneficial technique, back-propagation through time; dynamically changes the sequence length for the forward and backward passes. Embedding dropout: where occurrences of certain words are gone during the training stage. Furthermore, the approach in [35] primarily leverages the likes of transfer learning: the language model is trained on a large data corpus none related to the downstream task.

Our language model is first trained on the open data, and subsequently retrained for the downstream task. To transcribe the data in our collection, we used a third party ASR service (WER calculated on the subset of our data is within the range of 20-30 depending on quality of files). Finally, additional layers are added on top of the retrained model for classification or regression purposes. The key mechanisms within this implementation encompass discriminative fine tuning, where the different layers of the network use a different learning rate. This rate is the slanted triangular learning rate, which is where the rate of change is dependent on the stage of the training process. The remaining main features cover gradual unfreezing, back-propagation through time to handle longer language dependencies and concatenated pooling of multiple time steps from the recurrent network.

III. LABEL DISTRIBUTION AND COMORBIDITY

We first examined the distribution of depression, anxiety, and their co-occurrence in our data. As noted in the Method section, labels for each behavioral health condition come from self-report surveys and are quantized using a threshold of 10, after [16], into binary classes (dep+, dep- and anx+, anx-).

Table II provides information on the joint distribution of labels. Initial analysis encompasses overlapping labels and the correlation between dep+ and anx+ vs dep- and anx-. There are around 25% positive cases either for PHQ or GAD but only 18.5% of the overall data is positive for both labels (validation 14%). In total, roughly 15% (~2,000) of the sessions are single-condition, meaning either the user was positively diagnosed as depressed or anxious, but not both depressed and anxious. For the test subset the proportions are roughly the same; joint-positive plus joint-negative samples represent 88%, vs. 85% in the full data set. In Table II, values listed after the "/" refer to the number of sessions in the test partition only.

	anx+	anx-
dep+	**2,964 (18.5%)** / 455 (14.7%)	**1,295 (8.1%)** / 198 (6.4%)
dep-	**988 (6.1%)** / 163 (5.2%)	**10,703 (67.1%)** / 2,262 (73.4%)

As shown, most of the speakers had either neither condition, or both conditions. While our data set is relatively large compared with past speech corpora for these conditions, the amount of test data for the speaker having only one condition, is below 200 test sessions. For this reason we do not use 4-way classification in the present study; such work awaits future additional test data.

Fig. 1 represents a normalized histogram (in line plot format) for each survey (gold standard) value. The PHQ and GAD histograms are nearly identical, except for label values at 0, where there is a 5% discrepancy (800 sessions). Fig. 1 is for the entire data set (train and test). We observe a similar pattern in the test set only.
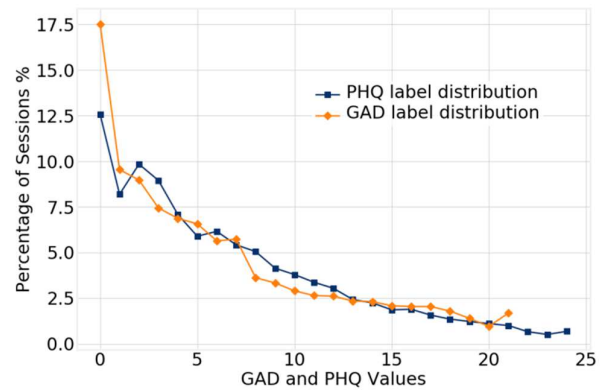

Fig. 1 Distribution of label values in corpus

The overall correlation between GAD and PHQ in our train and test set is high, at 0.80 [36]. Another way of looking at the label distribution is the matrix of each label score relative to the other, as shown in Fig. 2. Note the difference in score ranges. Each survey instrument question offers 4 possible scores (0, 1, 2, and 3). GAD-7 scores thus range from 0 to 20; PHQ-8 scores range from 0

PHQ GAD	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
0	1488	507	354	203	95	53	36	17	8	18	6	0	3	1	1	2	0	1	0	0	0	0	0	0	1
1	246	319	329	253	144	74	55	37	20	9	17	11	3	0	0	2	1	0	1	0	0	0	0	1	0
2	127	210	300	248	161	124	91	63	26	31	17	6	6	9	3	3	4	0	0	1	0	0	0	0	0
3	62	101	232	217	140	111	127	54	41	27	31	15	11	5	1	4	3	1	0	1	1	1	0	0	0
4	40	66	147	166	140	120	125	71	58	38	34	22	26	13	10	7	3	3	2	1	0	4	0	0	0
5	22	53	75	136	153	106	121	117	71	52	45	29	21	14	11	5	7	5	3	0	1	0	0	0	1
6	10	21	61	73	103	94	110	101	87	64	39	44	34	20	8	7	8	3	4	3	3	3	0	0	0
7	8	17	34	45	73	78	93	110	159	71	53	54	36	29	12	16	10	4	5	2	4	1	0	0	0
8	2	1	14	22	42	56	57	56	57	65	55	36	35	24	17	10	8	8	5	3	2	1	1	0	0
9	1	2	7	20	23	25	39	58	51	58	59	36	36	34	24	16	11	13	9	2	4	1	2	0	0
10	0	3	5	14	15	20	31	36	45	49	47	39	33	31	32	27	10	9	4	3	5	3	0	0	2
11	0	2	1	6	11	23	19	41	35	36	39	39	35	28	29	13	20	11	14	7	5	5	0	2	0
12	0	1	3	5	4	19	22	22	28	34	38	36	28	28	29	27	31	16	17	13	6	2	8	0	0
13	0	2	1	4	8	6	14	17	35	31	31	42	39	21	31	20	18	17	10	8	5	10	1	2	1
14	0	1	2	3	5	11	13	17	20	18	24	36	24	29	33	33	36	19	16	11	5	7	2	1	2
15	0	0	1	6	5	7	14	12	19	17	16	23	26	26	25	20	28	23	17	16	9	10	9	2	0
16	0	1	0	1	3	7	4	11	17	13	18	18	28	20	29	23	22	30	23	17	17	15	3	3	3
17	0	0	2	0	3	2	3	10	10	6	19	18	20	19	23	19	32	31	20	21	21	17	13	10	6
18	0	0	0	3	1	3	4	5	6	16	6	13	10	10	17	26	18	24	20	28	28	23	8	8	9
19	0	0	0	2	0	1	2	4	5	3	5	9	14	11	13	9	14	14	21	18	24	14	16	12	10
20	0	0	0	0	1	0	2	3	3	2	2	3	9	7	4	1	6	10	14	10	16	24	15	14	6
21	0	0	1	0	0	0	1	2	4	2	3	9	8	7	5	6	11	9	11	30	19	19	27	26	68

Fig. 2 Co-occurrence of PHQ-8 and GAD-7 by severity level, in the corpus.

to 24. Within each scale, higher values indicate higher condition severity. Fig. 2 represents the counts of sessions for a given combination of two specific label values. As shown, the majority of sessions occur near the diagonal, consistent with high correlation of two mental health conditions [37][2]. Interestingly, there is a greater variation of PHQ labels for each GAD label rather than vice versa. i.e., there is more variability in rows than in columns. This may reflect the fact that anxiety tends to be a precondition for depression more than vice versa [2].

IV. DEEP NLP PERFORMANCE

A. Model accuracy by condition

Before looking at the binary classification results, we analyzed model accuracy as a function of raw PHQ and GAD scores prior to their mapping for binary classification. As shown Fig. 3, there is a large variation in accuracy depending on score value. Values near the ends of the score range have accuracies near 0.90, whereas those near the threshold of 10, which are expected to be more confusable, are closer to 0.50. The noise at higher scores reflects an increase in data sparsity.
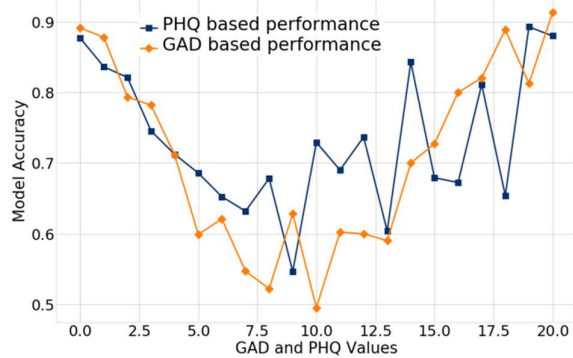

Fig. 3 Model accuracy by condition severity

The model has been trained for classification using binary classes. For both GAD and PHQ values at 10 or higher are mapped to positive for the condition. Our model achieves an AUC score for depression of 0.828. The result for anxiety is slightly lower, at 0.792. One possible reason for the difference is simply less attention so far in our work to model tuning for anxiety than depression. A second, independent possibility is discussed in *Section C*. As described in that section, there is some evidence for a fundamental difference in language sequence cues for prediction of anxiety, as compared with similar types of cues for depression. *Section C* discusses this possibility, as well as a countering effect of the condition priors, that may help explain these performance differences.

TABLE III
MODEL PERFORMANCE BY CONDITION

	AUC	Specificity (at EER)	Sensitivity (at EER)
PHQ	0.828	0.755	0.755
GAD	0.792	0.722	0.721

Given the large literature on depression detection in common corpora noted in Section I, we also seek to provide comparison information on our performance. Although we did not have access to the common data, we provide the following results to allow informal, indirect comparison. Leading, recent results on the AVEC common task are described in [16]. The AVEC data and Ellipsis data differ, but both use PHQ-8 scores as labels. We report metrics found in [16], using regression. The test results in [16] for RMSE, an error metric, are 5.51 versus 4.21 for our study. We believe we are thus in the range of competitive performance.

B. Comorbidity effects

We discovered that model performance is best when the speaker has either both conditions present, or both conditions absent. AUC for discriminating joint-presence and joint-absence sessions increases to 0.861 and 0.841, for PHQ and GAD respectively. We then asked whether this effect could be simply due to a change in session priors. The priors for joint- presence and joint-absence data change from around 0.20 to 0.16 for the positive class. After rebalancing the data (changing priors from 0.167 to 0.211), we found that this is not the case. The improved result remained—and even increased after rebalancing, to 0.863 and 0.849 respectively. This finding suggests that class discrimination is better for joint modeling of depression and anxiety than for individual modeling of either condition.
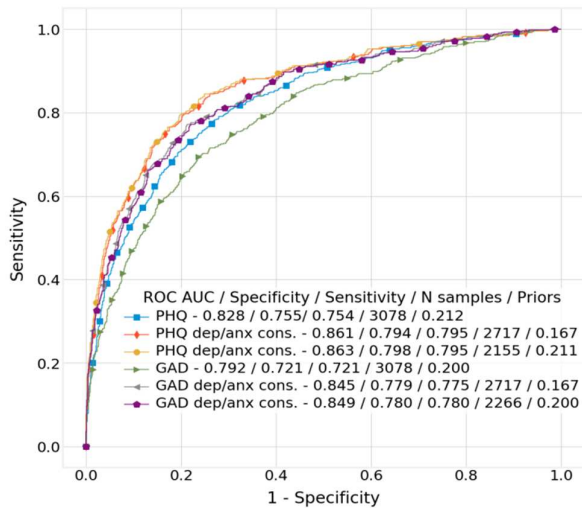

Fig. 4 AUC and EER-based sensitivity and specificity results. The legend is ordered by ascending AUC performance within each condition

C. Within-session model variability by class

Given the nature of the models used (see Section 2.2), we assume that certain word sequences and their dependencies are used as cues for the models, to separate positive from negative cases for each condition. To investigate, we gated the word sequence in a forward direction, to estimate the amount of prediction information available at a given time during a test session. For example, in the session of 800 words, the same number of cumulative gated samples was generated by adding one word at a time starting from the first token; in this case, 800 predictions were returned. For our total of 3078 test sessions, roughly 2.4M predictions were generated. Based on these predictions, we calculated a value we call Within-Session Model Variability. This process was performed separately for each condition. Note that the test set is exactly the same word sequence data (from the same speech), for both models.

Table IV provides results for this measure of variability for the depression model. As shown, variability within a session is highest for +,+ (both conditions), lowest for -,- (neither condition), and in between for the mixed cases. This suggests that a model tuned for maximum AUC on binary depression classification, is using word sequence cues associated with higher variability on this measure within a session.

TABLE IV
WITHIN-SESSION (W-S) MODEL VARIABILITY BY CROSSED CONDITION, FOR THE PHQ (DEPRESSION) MODEL

W-S Model Variability	dep+	dep-
anx+	0.090	0.088
anx-	0.084	0.077

The model trained for anxiety shows a similar pattern; see Table V. Here, however, (1) overall variability is lower than that for depression, and (2) variability for the -,- case is much lower than expected given the other three values. Because the same test data is used for both tables, and the NLP model methods are the same, this suggests that the word sequence cues for anxiety may be fundamentally weaker or less prevalent, than those for depression.

TABLE V
WITHIN-SESSION (W-S) MODEL VARIABILITY BY CROSSED CONDITION, FOR THE GAD (ANXIETY) MODEL

W-S Model Variability	dep+	dep-
anx+	0.077	0.065
anx-	0.061	0.048

A question, then, is why anxiety performance in AUC is only slightly below that for depression, as shown in Fig. 4 and earlier. The answer may lie in the difference in severity distributions. Fig. 1, Fig. 2 and Fig. 3 together, indicate that in our data, there is a preponderance of labels with a score of 0 on anxiety. Because this value is far from the binary classification threshold, the overall AUC performance for anxiety is boosted, compared with that for depression.

V. SUMMARY AND CONCLUSIONS

We have described a large corpus of spontaneous speech, labeled jointly for depression and anxiety. The data show a high level of comorbidity of the two conditions, as well as a positive data skew for severity within each condition. We examine performance of deep language models that make use of transfer learning for model pre-training. We achieve a classification performance of 0.83 AUC for depression, and 0.79 for anxiety; these results increase to AUC=0.86 and 0.85 respectively, for sessions with either both conditions or neither condition. After adjusting for data skew, these results persist, suggesting that discrimination is better for joint modeling. In addition, we examine within-session model variability. Because our data are perfectly matched over conditions, results suggest that word sequence cues for depression may be fundamentally stronger than those for anxiety. Taken together these results reveal the importance of considering comorbidity of behavioral health conditions for speech-technology-based screening and monitoring applications.

REFERENCES

[1] R. C. Kessler, M. Petukhova, N. A. Sampson, A. M. Zaslavsky, and H. Wittchen, "Twelve-month and lifetime prevalence and lifetime morbid risk of anxiety and mood disorders in the United States," *International Journal of Methods in Psychiatric Research* 21, 3, 169–184, 2012.

[2] N. Jacobson and M. Newman. "Anxiety and depression as bidirectional risk factors for one another: a meta-analysis of longitudinal studies." *Psychol Bull,* 143, pp. 1155-1200. 2017.

[3] A. Bittner, R.D. Goodwin, H.-U. Wittchen, "What characteristics of primary anxiety disorders predict subsequent major depressive disorder?" *Journal Clinical Psychiatry*. 2004;65:618–626

[4] D. Westen, and K. Morrison, "A multidimensional meta-analysis of treatments for depression, panic, and generalized anxiety disorder: An empirical examination of the status of empirically supported therapies." *Journal of Consulting and Clinical Psychology*, 69(6), 875–899. 2001.

[5] S. El-Den, T.F. Chen, Y.L. Gan, E. Wong and CL. O'Reilly. "The psychometric properties of depression screening tools in primary healthcare settings: A systematic review." *Journal of Affect Disorders*; 225:50.,3-22. 2018.

[6] A. Halfin, "Depression: The benefits of early and appropriate treatment," *American J. of Managed Care*, :S92–S97, 2007.

[7] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie and A. and T. Campbell, "Next-Generation Psychiatric Assessment: Using Smartphone Sensors to Monitor Behavior and Mental Health," *Psychiatr. Rehabil. J.*, vol. 38, no. 3, pp. 218–226, 2015.

[8] M. Ghai, S. Lal, S. Duggal and S. Manik. "Emotion recognition on speech signals using machine learning," *Proceedings of International Conference on Big Data Analytics and Computational Intelligence* (ICBDAC) (2017), pp. 34-39

[9] A. S. Cohen and B. Elvevag, "Automated computerized analysis of speech in psychiatric disorders," *Current Opinion in Psychiatry*, 27(3):203–209. 2014.

[10] A.H. Orabi, P. Buddhitha, M.H. Orabi and D. Inkpen "Deep Learning for Depression Detection of Twitter Users." *In proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 88-97). 2018.

[11] A. Yates, A. Cohan and N. Goharian "Depression and self - harm risk assessment in online forums," https://aclweb.org/anthology/D17-1322, 2017.

[12] R. P. Lippmann, "Speech recognition by machines and humans," *Speech Communication*, vol. 22, no. 1, pp. 1–16, 1997.

[13] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou. "Adieu features' end-to-end speech emotion recognition using a deep convolutional recurrent network," *In Proc. of ICASSP, Shanghai, China, March, IEEE* pages 5200—5204, 2016.

[14] J. Yuan, M. Liberman, and C. Cieri. "Towards an integrated understanding of speaking rate in conversation." in *INTERSPEECH 2006 - Ninth international conference on spoken language processing, Pittsburgh, Pennsylvania*, 2006

[15] A. Nilsonne, J. Sundberg, S. Ternstrom and A. Askenfelt. "Measuring the rate of change of voice fundamental frequency in fluent speech during mental depression," *The Journal of the Acoustical Society of America* 1988; 83: 716–728, 1988.

[16] Z. Zhao, Z. Bao, Z. Zhang, N. Cummins, H. Wang and B. Schuller, "Hierarchical Attention Transfer Networks for Depression Assessment from Speech," *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020*, pp. 7159-7163

[17] S. J. Sinclair, C. J. Siefert, J. M. Slavin-Mulford, M. B. Stein, M. Renna and M. A. Blais, "Psychometric Evaluation and Normative Data for the Depression, Anxiety, and Stress Scales-21 (DASS-21) in a Nonclinical Sample of U.S. Adults", *Evaluation & the Health Professions*, 35(3), 259–279. 2012

[18] A. Salekin, J. W. Eberle, J. J. Glenn, B. A. Teachman and J. A. Stankovic, "A weakly supervised learning framework for detecting social anxiety and depression," *ACM Interactive, Mobile, Wearable, and Ubiquitous Technologies,,* vol. 2, no. Article 1, 2018.

[19] F. Ringeval, B. Schuller, M. Valstar, N. Cummins, R. Cowie, L. Tavabi, M. Schmitt, S. Alisamir, S. Amiriparian, E.-M. Messner, et al. "Avec 2019 workshop and challenge: state-of-mind, detecting depression with AI, and cross-cultural affect recognition." *In Proc. 9th International on Audio/Visual Emotion Challenge and Workshop,* pages 3–12, 2019.

[20] M. De Choudhury M, M. Gamon, S. Counts, E. Horvitz. "Predicting depression via social media." *Proc. 7th Int. AAAI Conf. Weblogs Soc. Media, Boston, MA,* July 8–July 10, 2013

[21] G. A. Coppersmith, C. T. Harman, and M. Dredze. "Measuring post-traumatic stress disorder in Twitter," *In Proceedings of the Int. AAAI Conference on Weblogs and Social Media*, 2014

[22] J. Hanwen Shen and F. Rudzicz. "Detecting anxiety through reddit," In Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology – From Linguistic Signal to Clinical Reality. Association for Computational Linguistics, pages 58–65. 2017.

[23] I. Pirina and C. Coltekin, "Identifying depression on reddit: The effect of training data," In Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task, pages 9–12, 2018.

[24] Personal communication with Mark Liberman, Linguistic Data Consortium. 2019.

[25] T. Rutowski, A. Harati, Y. Lu and E. Shriberg, "Optimizing Speech-Input Length for Speaker-Independent Depression Classification." *Proc. Interspeech 2019*, 2019

[26] L. Manea, S. Gilbody, and D. McMillan, "Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): A meta-analysis." CMAJ: Canadian Medical Association journal. 184. E191-6. 10.1503/cmaj.110829. 2011.

[27] K. Kroenke, T.W. Strine, R. Spitzer, J.B.W. Williams, J.T. Berry and A.H. Mokdad, "The PHQ-8 as a measure of current depression in the general population." *Journal of Affective Disorders, Volume 114*, Issues 1–3, 2009.

[28] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean. "Distributed representations of words and phrases and their compositionality." In Advances in neural information processing systems, pages 3111–3119. 2013

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

[30] L. Mou, Z. Meng, R. Yan, G. Li and Y. Xu, "How transferable are Neural Networks in NLP Applications?" Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. 2016.

[31] S. Merity, C. Xiong, J. Bradbury and R. Socher "Pointer sentinel mixture models," *CoRR,* abs/1609.07843, http://arxiv.org/abs/1609.07843. 2016.

[32] R. Sennrich, B. Haddow, and A. Birch. "Neural machine translation of rare words with subword units." arXiv preprint arXiv:1508.07909, 2015.

[33] T. Kudo. "Subword regularization: Improving neural network translation models with multiple subword candidates." ACL 2018.

[34] S. Merity, N. Shirish Keskar, and R. Socher, "Regularizing and optimizing lstm language models." *abs/1708.02182*. 2017.

[35] J. Howard and S. Ruder. "Universal language model fine-tuning for text classification." arXiv preprint arXiv:1801.06146. 2018.

[36] K. Kroenke, J. Wu, Z. Yu, M. J. Bair, J. Kean, T. Stump and P. O. Monahan, "Patient health questionnaire anxiety and depression scale: initial validation in three clinical trials," *Psychosomatic Medicine 78,* 716–727. 2016

[37] N. Jacobson, M. Newman. "Avoidance mediates the relationship between anxiety and depression over a decade later." *Journal of Anxiety Disorders*. 2014;28(5):437–445. 2014.