Passive Diagnosis incorporating the PHQ-4 for Depression and Anxiety

Fionn Delahunty[1,2] **Robert Johansson**[3,4] **Mihael Arcan**[2]
[1] University of Gothenburg | Chalmers University of Technology, Sweden
[2] Insight Centre for Data Analytics, Data Science Institute, National University of Ireland Galway, Ireland
[3] Department of Psychology, Stockholm University, Sweden
[4] Department of Computer and Information Science, Linköping University, Sweden
{Fionn.Delahunty,Mihael.Arcan}@Insight-centre.org
Robert.Johansson@liu.se

Abstract

Depression and anxiety are the two most prevalent mental health disorders worldwide, impacting the lives of millions of people each year. In this work, we develop and evaluate a multilabel, multidimensional deep neural network designed to predict PHQ-4 scores based on individuals written text. Our system outperforms random baseline metrics and provides a novel approach to how we can predict psychometric scores from written text. Additionally, we explore how this architecture can be applied to analyse social media data.

1 Introduction

According to the World Health Organization (WHO), major depressive disorder[1] is the largest cause of disability worldwide (World Health Organization, 2018), with a lifetime prevalence rate between 15% and 17% (Ebmeier et al., 2006). Depression is highly co-morbid with several other mental disorders, the most prevalent of which is a generalized anxiety disorder.[2] Almost 50% of individuals diagnosed with depression will also be diagnosed with anxiety (Johansson et al., 2013).

As a result, many clinicians will investigate for the presence of both disorders at the time of diagnosis. To do so, psychometric questionnaires are often employed as a quick and reliable initial assessment tool, the most common of which is the Patient Health Questionnaire (PHQ). The PHQ-4 is a short form questionnaire design to access the presence or absence of the core symptoms in depression and anxiety (Löwe et al., 2010). The questionnaire has demonstrated both high validity and reliability across several languages and cultures (Kroenke et al., 2010).

Despite the usefulness of these questionnaires, there is still a reliance on individuals actively seeking a diagnosis from a medical professional before they can be applied. Research has shown that those suffering depression and anxiety often are unaware their symptoms are due to a medical disorder and attribute them to poor mood or external factors (Barney et al., 2006; Latalova et al., 2014). This presents a unique challenge in the medical community, in how to inform and encourage individuals to come forward for diagnosis.

Delahunty et al. (2018) have proposed the concept of **passive diagnosis**, also known as high-performance medicine (Topol, 2019). This term refers to the ability for machine learning algorithms to constantly monitor an individuals health and inform the individual if certain changes are evidence of a possible disorder in the future. This is in comparison to the traditional concept of active diagnosis where an individual suffering certain symptoms would actively seek out a medical diagnosis.

Examples of applications in this domain include *DeepCare*, which is an end-to-end application designed to diagnose a wide range of disorders (Pham et al., 2016). Such systems allow clinicians to either prevent a disorder occurring or provide early intervention to minimise its effects.

2 Related work

While exploring the effects of expressive writing on PTSD[3] treatment, (Pennebaker et al., 2003) established that the way in which individuals wrote was often indicative of their mental state, specifically their use of function words (Prendinger and Ishizuka, 2005). Examples of this included higher counts of the personal pronouns and negative

[1] Hereafter referred to as simply depression.
[2] Hereafter referred to as simply anxiety.
[3] Post-Traumatic Stress Disorder

words in depressed individuals' writing, which is attributed to a manifestation of Beck's cognitive model and Pyczsinski and Greenberg's self-focus model of depression (Rude et al., 2004).

Over recent years this work has been combined with the fields of natural language processing and machine learning to develop classifiers algorithms which can predict if an individual is likely to be diagnosed with a certain disorder. Work has focused on bipolar disorder (Huang et al., 2017), depression (De Choudhury et al., 2013) and anorexia (Ramiandrisoa and Benamara, 2018). For the last number of years, the *CLEF* conference has hosted a workshop on early risk prediction of mental disorders based on social media data (Losada and Crestani, 2016), resulting in almost 50 publications in this area.

However, much of the existing work suffers from the limitation of viewing these disorders as binary occurrences, whether a disorder is present or not. Although this approach makes sense given the nature of machine learning classifiers, from the perspective of medical professionals, however, individuals can rarely be placed into binary classes. Different combinations of symptoms can dramatically affect the diagnosis (American Psychiatric Association, 2013).

Previous work in 2018 was the first to view these disorders on a symptomatic level (Delahunty et al., 2018). In this paper, we expand the previous work by including anxiety and making use of the PHQ, which compared to the Beck's depression inventory is a non-commercial psychometric questionnaire (Kung et al., 2013).

The PHQ-4 assess the severity of the two primary symptoms for depression and anxiety respectfully, anhedonia, depressed mood, excessive anxiety and uncontrollable worry (American Psychiatric Association, 2013). An individual is asked to rate the occurrence of each symptom over the last two weeks on a four-point scale from "Not at all" to "Nearly every day". The aim of our work was to develop a machine learning algorithm that given textual data could predict an outcome value for each of the four questions on the PHQ-4. Unlike previous work, e.g. Delahunty et al. (2018), we did not employ separate algorithms for the four symptoms, but considered that all four symptoms are intrinsically interconnected. Within the machine learning literate, multilabel and multi-class approaches have been shown to outperform individual separate classifiers (Schmidhuber, 2015).

Previous work in this domain has often employed extracted data from social media sites as training data (Losada and Crestani, 2016). In many cases, this limits the application of the work because it is impossible if the individuals in the training data actually had clinical diagnosis (De Choudhury and De, 2014). To overcome this limitation, our work employs a dataset collected in an in-person medical setting where clinical diagnosis are performed by trained professionals. We aim to explore if training on non-social media data will allow for accurate evaluation on social media data.

3 System Description

3.1 Data

Our initial dataset is the DAIC-WOZ, which is composed of transcribed clinical interviews collected through a Wizard-of-Oz approach for 142 patients (Gratch et al., 2014). The topic of the interviews are general conversations and were all collected within the United States. For each patient, a transcript of their interview is provided along with PHQ-8 scores, where bot statements were removed leaving only patient statements. PHQ-8 scores can be mapped to PHQ-2 scores, and GAD-2 scores were inferred from data provided by Johansson et al. (2013). The final dataset was composed of 23,726 text statements.

To evaluate our system on social media data, we employed the *Reddit* depression dataset (Losada and Crestani, 2016). This dataset we gained access to contained *Reddit* posts for 253 users (of which 161 are attributed as to be suffering depression). Diagnosis is binary (depressed or not depressed) depending on if users post on certain depression sub-forums.

3.2 Feature extraction

Three methods of feature extraction were employed.

Text representation was employed using the Universal Sentence Encoder (USE), specifically developed for longer than word representations. The model is trained using a deep learning transformer neural network architecture on a variety of datasets (Cer et al., 2018). Each of our patient statements was passed into their pretrained model and a statement level representation vector of shape 512 was returned.

LIWC is a psycholinguistic dictionary containing 94 psychological trait dimensions and over 2,000 words related to these dimensions (Pennebaker et al., 2001). A percentage count of the number of words in the text related to each dimension is computed. To identify an optimal subset of the number of relevant dimensions, we reviewed all proceedings from the CLEF *eRisk* workshop 2017 and 2018 (Losada and Crestani, 2016). For each proceeding that employed LIWC, the list of dimensions included was taken. An intersection of these lists was then taken to create a subset of 22 relevant dimensions, which resulted in the following features being included in our model: *word count, analytical thinking, authentic, emotional tone, function words, pronoun, personal pronouns, 1st person singular, 1st person plural, 2nd person, 3rd person singular, articles, auxiliary verbs, conjunctions, negations, regular verbs, negative emotions, social words, cognitive processes, past focus, present focus, future focus*.

Psychometric similarity Recent work has seen success in comparing word embeddings in terms of semantic similarity (Mihalcea et al., 2006; Li et al., 2003), where the distance between embeddings in x^N-dimensional space is considered equal to their likeness in terms of the semantic content. Since USE creates sentence level embeddings, this allows us the ability to compare sentences in terms of similarity. We employed this approach by comparing the semantic similarity of patient statements with responses from psychometric questionnaires. The principle was that if a patient statement reflected the same content of a psychometric test it should have a higher similarity score compared with a random statement.

Four questionnaires were identified by choosing cognitive theories relevant to the aetiology of each of the four PHQ-4 symptoms. Details regarding the theories are included in Table 1. The concatenation of questions across all four questionnaires amounted to 104 questions. For each patient statement, a 512 embedding dimension was computed with the USE pre-trained model, along with this, embeddings for each of the 104 patient questions were computed. The inner dot product for each statement and question was computed and returned as a feature. The inner dot product measures how close two vectors are in the Euclidean space of the trained model, closer vectors implies more similar semantic similarity.

The resulting dataset was composed of 638 features. All features were scaled by removing the mean and scaling to unit variance within the bounds of -1 and 1.

3.3 Our approach

To model the interconnectivity of the four PHQ-4 symptoms, we employed a deep neural network (DNN) architecture. Unlike simpler algorithms, such as classical regression, which uses a single function, ($Y \approx f(X, \beta)$),[4] DNNs employ a large number of "neurons", each of which is fitted with an independent function with a set of weights and an activation function (Schmidhuber, 2015). Current work demonstrates that this architecture models the internal representation better than separate classifiers (Schmidhuber, 2015).

For each patient statement, the neural network needs to be able to output an ordinal value score for each question. This requires that the network outputs both multilabel (four symptoms) and multivalue (ordinal score). This architecture is regarded as multi-dimensional or multi-targeted classification, where the output is assigned both a set of labels $y = (y_0, \ldots, y_d)$, and for each label y an ordinal value in the 0 to d (Read et al., 2014). These methods are still in early development are mostly untested outside of theoretical proposals.

Our proposed method to address this problem is a two-step approach. Firstly, we apply a multilabel learning approach to constantly predict a *Sigmoid* score for each of the four symptoms. This is achieved by using a binary cross entropy loss function that can model the interconnectivity of the labels (Trotzek et al., 2018; Nam et al., 2014; Zhang and Zhou, 2014; Mencıa and Fürnkranz, 2008) and a *Sigmoid* function on the final layer (Trotzek et al., 2018). Secondly, following that, we set manual threshold values to refine this score into ordinal values for interpretability.

For the final output per symptom, we set the value to 0, if the outcome of the *Sigmoid* function is less than 0.25, 1 if the *Sigmoid* score is between 0.25 and 0.50, 2 between 0.50 and 0.75 and 3, if the *Sigmoid* score is larger than 0.75.

To compare our approach against a simpler model architecture, and determine if a DNN architecture is appropriate, we also trained a random forest classifier which is equally able to model

[4] Y is dependent variable, X is independent variable & β is unknown parameter.

PHQ-4 Symptom	Theory	Assessment tool
Feeling nervous or anxious	Intolerance of uncertainty (Clark et al., 1994)	Intolerance of Uncertainty Scale
Uncontrollable worry	Positive belief about worry (Clark et al., 1994)	Penn State Worry Questionnaire
Anhedonia	Avoidance behaviour (Clark et al., 1994)	Cognitive-Behavioural Avoidance Scale
Depressed mood	Negative triad (Beck, 1991)	Beck's depression inventory

Table 1: Summary of aetiology theories and assessment tools.

multilabel outputs (Gharroudi et al., 2014).

3.4 Hyperparameter Tuning

Hyperparameter tuning was achieved using a genetic algorithm approach. This approach takes its basis from the biological concept of evolution (Friedrichs and Igel, 2005). A broad set of hyperparameters are chosen (details in the appendix), the algorithm creates a generation by choosing a random subset of these and trains a population of 20 network networks with different random hyperparameters. Each network is evaluated on a metric, in this case, the minimization of the *Hamming loss* criteria (Zhang and Zhou, 2014). The five best networks based on this metric are chosen, along with five random ones to allow some variability in the population. Another generation is created with random hyperparameters chosen from within the subset of the last generation, while we repeat this process for a total of ten generations.

The final optimal hyperparameters, based on the minimized *Hamming loss*, were six dense layers with dimensions of 1024, 768, 256, 128, 64 and 4 in that order. Each layer contained a *relu* activation function, except for the final layer, which contained *Sigmoid*. Binary cross entropy was applied to compute the loss function and *adagrad* function as the optimizer.

The following hyperparameters were employed for the RFC, number of trees in the forest = 10, split criterion = gini, no max depth of trees, minimum samples to split a tree = 2, minimum leaf sample = 1.

4 Results

Multilabel Our evaluation was first performed on the multilabel aspect of the network. PHQ scores were reduced to a binary class (0 for 0, 1 for 1,2,3) and *Sigmoid* outputs were binarized on a cutoff point of 0.5. *Hamming loss* was the chosen metric for evaluation (Zhang and Zhou, 2014), which computes the distance between predicted and true values. A ten-fold cross-validation resulted in a score of 0.388, 95% [0.3870, 0.3905]. To compare this against a random baseline, where a set of prediction scores are computed using a random number generator, a *Hamming loss* of 0.49, 95% [0.481, 0.519], is achieved.

Question	Accuracy	Sensitivity	Specificity
1	0.25	0.96	0.66
2	0.58	0.79	0.84
3	0.39	0.87	0.91
4	0.32	0.94	0.71

Table 2: Sensitivity and specificity scores for each question as predicted by the model

	Precision	Recall	F1-score
Depressed	0.16	0.17	0.16
Non-Depressed	0.59	0.56	0.57

Table 3: Classification scores from the *eRisk* data

Multidimensional Using the cutoffs mentioned above, *Sigmoid* scores were transformed into ordinal values. Since the *Hamming loss* is unsuited to this evaluation, a more suitable metric is the *Example Accuracy*, which consists of comparing if the prediction of each individual is completely correct (all values match) or incorrect and taking the mean value across all predictions (Read et al., 2014). The result across ten-fold cross-validation is 0.221, 95% [0.201, 0.243]. In comparison to ten-fold cross-validation of the RFC which resulted in a score of 0.087, 95% [0.086, 0.085].

$$EX.\ ACCURACY = \frac{1}{N}\sum_{i=1}^{N} I(\hat{y}^{(i)}, y^{(i)}) \quad (1)$$

Sensitivity, Specificity are both common evaluation metrics employed in medical literature and are important in considering the real-life implications of true positives and false negatives. Results from the trained neural network per question are presented in Table 2.

Social media evaluation To perform this, we evaluated our network on the *Reddit* dataset compiled by the authors (Losada and Crestani, 2016). We considered a score above 3 on the PHQ-2 (latter two questions on the PHQ-4) to be indicative of a user suffering depression. Results are presented in Table 3. Accuracy score was 43%, which was 18% below the majority class baseline.

5 Conclusion and Future Work

Exploring new methods to diagnose and treat mental health disorders has become a priority in many countries. Passive diagnosis has the potential to allow for early treatment and diagnosis to become standard practice in society. In the course of this work, we have developed a method to apply this concept to the PHQ-4 to screen for depression and anxiety.

Our approach is the first publication to explore how multilabel neural networks can predict depression and anxiety. We have developed a Multidimensional classification architecture to model the interconnectivity of the symptoms combined with a hardcoded threshold value to output ordinal scores. For multilabel evaluation, our model scores considerably better than the random baseline. While for multidimensional classification our system outperforms a simpler RFC by 14%. When evaluating on social media data from Losada and Crestani (2016), the models fail to match the majority class baseline.

In almost all questions on the PHQ-4, we demonstrate high sensitivity for predicting the disorder. Specificity is slightly lower in many cases, however, for early-stage diagnostics, this is often an acceptable outcome since it often better to ensure false negatives do not occur.

This demonstrates the non-trivial nature of training on one domain of data and evaluating on another. Two out of three of our feature sets, psychometric similarity and text representation employed the pre-trained USE model, which was also trained on non-social media style data. Future work will need to explore the ability to create models that are less semantically domain specific and better able to generalize across writing styles. The concept of transfer learning has seen success in this area (Glorot et al., 2011).

Our approach is incomparable to the proceedings in the *eRisk* workshop who focus on the temporal aspect of the prediction. Data is released in chunks over time and accuracy is penalized as the length of time from the beginning increases.

In final conclusion, our work has demonstrated that neural networks offer a potential new route for the area of passive diagnosis and prediction of depression and anxiety. Future work is required to ensure the generalizability of the approach, however.

Acknowledgments

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (Insight).

References

. American Psychiatric Association. 2013. *Diagnostic and Statistical Manual of Mental Disorders, 5th Edition: DSM-5*. American Psychiatric Pub.

Lisa J Barney, Kathleen M Griffiths, Anthony F Jorm, and Helen Christensen. 2006. Stigma about depression and its impact on help-seeking intentions. *Australian & New Zealand Journal of Psychiatry*, 40(1):51–54.

Aaron T. Beck. 1991. Cognitive therapy: A 30-year retrospective. *American Psychologist*, 46(4):368–375.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

David A. Clark, Robert A. Steer, and Aaron T. Beck. 1994. Common and Specific Dimensions of Self-Reported Anxiety and Depression: Implications for the Cognitive and Tripartite Models. *Journal of Abnormal Psychology*, 103(4):645–654.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13*, pages 47–56.

Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. *Proceedings of the Eight International AAAI Conference on Weblogs and Social Media*, pages 71–80.

Fionn Delahunty, Ian D. Wood, and Mihael Arcan. 2018. First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot. In *CEUR Workshop Proceedings*, volume 2259, pages 327–338.

Klaus P. Ebmeier, Claire Donaghey, and J. Douglas Steele. 2006. Recent developments and current controversies in depression. *Lancet*, 367(9505):153–167.

Frauke Friedrichs and Christian Igel. 2005. Evolutionary tuning of multiple svm parameters. *Neurocomputing*, 64:107–117.

Ouadie Gharroudi, Haytham Elghazel, and Alex Aussem. 2014. A comparison of multi-label feature selection methods using the random forest paradigm. In *Canadian conference on artificial intelligence*, pages 95–106. Springer.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520.

Jonathan Gratch, Ron Artstein, Gale M Lucas, Giota Stratou, Stefan Scherer, Angela Nazarian, Rachel Wood, Jill Boberg, David DeVault, Stacy Marsella, et al. 2014. The distress analysis interview corpus of human and computer interviews. In *LREC*, pages 3123–3128. Citeseer.

Yen-Hao Huang, Lin-Hung Wei, and Yi-Shin Chen. 2017. Detection of the Prodromal Phase of Bipolar Disorder from Psychological and Phonological Aspects in Social Media.

Robert Johansson, Per Carlbring, Å sa Heedman, Björn Paxling, and Gerhard Andersson. 2013. Depression, anxiety and their comorbidity in the swedish general population: point prevalence and the effect on health-related quality of life. *PeerJ*, 1:e98.

Kurt Kroenke, Robert L Spitzer, Janet BW Williams, and Bernd Löwe. 2010. The patient health questionnaire somatic, anxiety, and depressive symptom scales: a systematic review. *General hospital psychiatry*, 32(4):345–359.

Simon Kung, Renato D Alarcon, Mark D Williams, Kathleen A Poppe, Mary Jo Moore, and Mark A Frye. 2013. Comparing the beck depression inventory-ii (bdi-ii) and patient health questionnaire (phq-9) depression measures in an integrated mood disorders practice. *Journal of affective disorders*, 145(3):341–343.

Klara Latalova, Dana Kamaradova, and Jan Prasko. 2014. Perspectives on perceived stigma and self-stigma in adult male patients with depression. *Neuropsychiatric disease and treatment*, 10:1399.

Yuhua Li, Zuhair A Bandar, and David McLean. 2003. An approach for measuring semantic similarity between words using multiple information sources. *IEEE Transactions on knowledge and data engineering*, 15(4):871–882.

David E Losada and Fabio Crestani. 2016. A Test Collection for Research on Depression and Language Use CLEF 2016, Évora (Portugal). *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 28–29.

Bernd Löwe, Inka Wahl, Matthias Rose, Carsten Spitzer, Heide Glaesmer, Katja Wingenfeld, Antonius Schneider, and Elmar Brähler. 2010. A 4-item measure of depression and anxiety: Validation and standardization of the Patient Health Questionnaire-4 (PHQ-4) in the general population. *Journal of Affective Disorders*, 122(1-2):86–95.

Eneldo Loza Mencıa and Johannes Fürnkranz. 2008. Pairwise learning of multilabel classifications with perceptrons. In *IEEE International Joint Conference on Neural Networks*.

Rada Mihalcea, Courtney Corley, Carlo Strapparava, et al. 2006. Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6, pages 775–780.

Jinseok Nam, Jungi Kim, Eneldo Loza Mencía, Iryna Gurevych, and Johannes Fürnkranz. 2014. Large-scale multi-label text classification - Revisiting neural networks. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8725 LNAI(PART 2):437–452.

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

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577.

Trang Pham, Truyen Tran, Dinh Phung, and Svetha Venkatesh. 2016. Deepcare: A deep dynamic memory model for predictive medicine. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 30–41. Springer.

Helmut Prendinger and Mitsuru Ishizuka. 2005. The empathic companion: A character-based interface that addresses users'affective states. *Applied Artificial Intelligence*, 19(3-4):267–285.

Faneva Ramiandrisoa and Farah Benamara. 2018. IRIT at e-Risk 2018. In *E-Risk workshop*, pages 367–377.

Jesse Read, Concha Bielza, and Pedro Larrañaga. 2014. Multi-dimensional classification with super-classes. *IEEE Transactions on knowledge and data engineering*, 26(7):1720–1733.

Stephanie S. Rude, Eva Maria Gortner, and James W. Pennebaker. 2004. Language use of depressed and depression-vulnerable college students. *Cognition and Emotion*, 18(8):1121–1133.

Jürgen Schmidhuber. 2015. Deep learning in neural networks: An overview. *Neural networks*, 61:85–117.

Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56.

Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. 2018. Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences.

. World Health Organization. 2018. Depression Fact Sheet. Technical report.

Min-Ling Zhang and Zhi-Hua Zhou. 2014. A review on multi-label learning algorithms. *IEEE transactions on knowledge and data engineering*, 26(8):1819–1837.

A Appendices

A.1 Hyperparameters

The full pool of possible hyperparameter fed into the genetic algorithm is as follows; possible neurons (64, 128, 256, 768, 2014), possible layers (1, 2, 3, 4, 5, 6, 7), possible activation functions (relu, elu, tanh, sigmoid, hard sigmoid, softplus, linear), possible optimizers (rmsprop, adam, sgd, adagrad, adadelta, adamax, nadam)