First insights on a passive major depressive disorder prediction system with incorporated conversational chatbot

First Insights on a Passive Major Depressive Disorder Prediction System with Incorporated Conversational Chatbot

Fionn Delahunty[1 [0000−0002−2185−924X]], Ian D. Wood[3 [0000−0002−6094−0358]] , and Mihael Arcan[2 [0000−0002−3116−621X]]

[1] Computer Science and Engineering, Chalmers University of Technology & University of Gothenburg, Sweden
Fionnd@student.chalmers.se
[2] Insight Centre for Data Analytics, Data Science Institute, National University of Ireland, Galway
mihael.arcan@insight-centre.org
[3] Hong Kong University of Science and Technology
drevicko@gmail.com

Abstract. Almost 50% of cases of major depressive disorder go undiagnosed. In this paper, we propose a passive diagnostic system that combines the areas of clinical psychology, machine learning and conversational dialogue systems. We have trained a dialogue system, powered by sequence-to-sequence neural networks that can have a real-time conversation with individuals. In tandem, we have developed specific machine learning classifiers that monitor the conversation and predict the presence or absence of certain crucial depression symptoms. This would facilitate real-time instant crisis support for those suffering from depression. Our evaluation metrics have suggested this could be a positive future direction of research in both developing more human like chatbots and identifying depression in written text. We hope this work may additionally have practical implications in the area of crisis support services for mental health organisations.

Keywords: Depression · Social Media · Conversational Chatbot

1 Introduction

In 1952, the American Psychiatric Association published the first diagnostic and statistical manual of mental disorders [3]. The publication, now in its fifth edition [4], is still considered as one of the most important resources within clinical psychology. This edition includes 297 separate mental disorders, grouped into 31 different categories based on shared etiologies and symptoms.

One of these categories is depressive disorders, which includes nine separate disorders an individual can suffer from. These disorders are primarily characterised by the presence of *sad, empty* or *irritable* mood [4]. Out of these nine disorders, major depressive disorder (MDD) is the most commonly occurring.

Each year, 300 million individuals worldwide will suffer a major depressive episode lasting a minimum of two weeks [27]. Furthermore, less than 50% will be

correctly diagnosed and offered appropriate treatment. Of those left untreated, MDD can lead to suicide, an estimated 800,000 people a year lose their life to suicide [27]. One particular issue which contributes to the low diagnostic and treatment rates is how the etiology of depression can directly interfere with an individual seeking treatment.

Since many mental disorders are not characterised by clear changes in external or physical appearance, detection and diagnosis become more challenging. In the case of depressive disorders, individuals are often not aware that their symptoms are due to a medical disorder and often attribute them to poor mood or external factors [6]. To make this increasingly more complex, MDD often also negatively affects an individual's social interactions. This can hamper an individual to actually seek professional support or talk about their experiences. This presents a unique challenge in the medical community, in how to identify and support individuals to come forward for diagnosis.

In this work, we propose the concept of **passive diagnosis**, a term for a new field of research seen over the last two or three years. This work is not exclusive to the research of mental health disorders, but concerns itself with using machine learning techniques for predicting a potential future medical disorder. In comparison to the traditional concept of active diagnosis, where an individual suffering certain symptoms would actively seek out a medical diagnosis, this process can now be facilitated by adding a passive element. Unlike a medical professional who has limited time and resources, it is feasible to have machine learning algorithms constantly passively observe an individual's health. Once these algorithms detect certain changes in an individual's health that might be indicative of a disorder, the algorithm can inform the individual and an appropriate human professional for further investigation.

An example of this application is *DeepCare*, developed by a research team in Google [33]. This end to end application is designed to diagnose a wide range of disorders, but it has also inspired more specific approaches as well. The authors of [40] have developed a model which predicted suicide attempts up to two years in the future by reviewing historical medical records. This field of work allows medical professionals to actively provide interventions to those at high risk before the disorder even sets in. Our work can be considered to follow a similar trend, where we propose a passive diagnostic approach to MDD.

2 Related Work

2.1 Depression Detection

As far back as 1901, Sigmund Freud proposed that language could give us an insight into certain mental illnesses [30]. This idea received little attention until the start of the 21st century where links with the field of natural language processing (NLP) and computer science allowed for a more in-depth study of the topic [9].

This combination of NLP and psychology led to a number of publications investigating how different mental illnesses such as bipolar disorder [13], MDD [38] and anorexia [34] can manifest through an individual's specific use of certain language characteristics. Some of the earliest work by [36] noted the higher use of the word "I" in depressed students' written notes. This observation has been

noted in several more recent studies [34, 38] and Pennebaker explains how the observation is consistent with the etiology of MDD. Extrapolations of this, higher counts of personal pronouns and lower counts of future temporal words have also been seen [32].

This work has inspired the machine learning community to investigate if there was sufficient basis for classifiers to distinguish between individuals suffering certain mental disorders and those not [13, 34]. This work led to the formation of the eRisk workshop at the conference and labs evaluation forum in 2017 and 2018. Which has inspired almost 30 different teams to explore how various machine learning classifiers perform when applied to the prediction of MDD [1, 22, 34].

Although this body of work has a solid machine learning background and high evaluation scores, from our perspective its practical medical application is limited. All the approaches look at MDD as a binary outcome variable, predicting at time x, if an individual positively diagnoses for MDD or not.

We understand how this approach makes sense given that many elementary machine learning classifiers perform best when predicting a simple binary outcome. From the perspective of a medical professional, however, we can rarely place individuals into binary classes, there is always some level of ambiguity regarding the diagnosis of a disease. MDD is a complex disorder, where the presence or absence of certain symptoms can have dramatically different effects on the diagnosis [4]. For many professionals, MDD will be viewed as a spectrum, with individuals falling from low risk to high risk [6]. This limitation has also been noted by [15] whose proposed solution was recording the specific mention of certain symptoms in online text.

Following consultation with medical professionals, we decided to overcome this limitation by building a host of separate classifiers that work on a symptomatic level. The DSM-V lists nine different symptoms that can be present during the occurrence of MDD. We propose that five of these symptoms can be reasonably detected to some degree through an online human-computer interaction. These targeted symptoms are *depressed mood* most of the day, *weight change* not attributed to dieting, *sleep change* characterised by insomnia or hypersomnia, *inappropriate guilt*, and *suicidal ideation*. We developed five separate classifiers and allow medical professionals to make an overall diagnosis based on the results of the classifiers and their own domain expertise.

2.2 Conversation Chatbots

Recent approaches to building conversational chatbot systems are dominated by the usage of neural networks. [39] present an approach for conversational modelling, which uses a sequence-to-sequence neural model. Their model predicts the next sentence given the previous sentences for an IT help-desk domain, as well as for an open-domain trained on a subtitles dataset. For an open-domain dialogue generation, [19] propose an adversarial training approach utilising reinforcement learning to produce sequences that are indistinguishable from human-generated dialogue utterances.

A heuristic that guides the development of neural baseline systems for the extractive conversational chatbot task is described in [42]. Their system, called FastQA, demonstrates good performance, due to the awareness of question words

while processing the context. [35] demonstrate an approach to non-factoid answer generation with a separate component, which is based on bidirectional LSTMs to determine the importance of segments in the input.

The increased use of social media as a communication tool between customers and brands has allowed for the development of these systems to handle real-time inquiries [44]. In addition, increasingly, mental health support services are incorporating real-time communication tools such as texting and social media messaging as methods for individuals to talk to a counselor [12]. Some work has already explored the possibility of building conversational chatbots that emulate a counselor, this work makes use of both audio, visual and text-based interactions [43].

3 Experimental Settings

In this section, we give an overview of data collection techniques employed and the feature extraction methods used for each of our five classifiers. Additionally, we outline the process employed for our conversational system.

3.1 Sequence-to-Sequence Neural Network Toolkit

To train the conversational system, we use the **OpenNMT** toolkit [17], which is a generic deep learning framework mainly specialising in sequence-to-sequence (seq2seq) models covering a variety of tasks such as machine translation, summarisation, speech processing, and question answering. We used the default neural network training parameters, i.e. 2 hidden layers, 500 hidden bidirectional LSTM[4] units, input feeding enabled, batch size of 64, 0.3 dropout probability and a dynamic learning rate decay.

Our data was composed of 13,053,384 million question-answer pairs, 6,237,118 of which were obtained from the subreddit `/r/AskReddit` (see examples in Table 1). Subreddit submissions were considered as questions and the first reply as an answer. The remaining pairs were extracted from the OpenSubtitles dataset [20]. The conversational model was trained for 13 epochs, which was further tuned on a selected extraction from the eRisk corpus (cf. 3.2).

3.2 Depressed Mood Classifier

Data We consider the existing work published in the eRisk proceedings to have been inadvertently focused on the symptom of depressed mood. We were provided with the eRisk task training set created by [21] consisting of comments and submissions from 486 Reddit users, of which 83 users were labelled as suffering depression. The original eRisk task was focused on early detection of MDD, whereby our conversation system rather was envisioned to have short conversations instead. Therefore, we considered each submission or comment as a single data point labelled as depressed or non-depressed. This led to a collection of 307,065 data points, of which 32,993 had the depressed label.

Linguistic based features. Five different types of linguistic features were included, the first of which is the Linguistic Inquiry and Word Count lexicon

[4] LSTM - long short-term memory

Table 1: Examples of Questions-Answer pairs used to train the conversational chatbot.

Question	Answer
Reddit Dataset	
What is the best compliment you can say to a girl?	I love your smile because it always makes me smile.
What's a sign that you're getting older?	Everything hurts
OpenSubtitles Dataset	
I think we probably Have the capacity	To write a hit song ,
To record it in a way	That radio would really Accept it with open arms ,
eRisk Dataset	
Do you ever feel miserable or unhappy?	It shoulds like your engage in a fair amount of rumination...
Ever feel so alone that it physically hurts?	This I can relate to. Im staying alone, in another country without anyone...

(LIWC), which is commonly used in the eRisk task [29]. This lexicon scores 78 different linguistic features related to social, clinical, health and cognitive psychology. Scores are percentages of words in a text that reflect a given emotion, scaled between 1 and 0, where 1 indicates all words in a sentence reflect a given emotion [29]. [10] found that the use of activation and dominance sentiment characteristics to be a strong predictor of MDD in their Twitter dataset [10, 26].

The Warriner lexicon [41] contains a more detailed analysis of *valence*, *activation* and *dominance* scores for words grouped by male, female and overall, providing a total of nine features. The NRC Affect Intensity Lexicon contains four different emotion scores: *anger*, *fear*, *joy*, *sadness* [25], while the SenticNet 5 lexicon provides polarity and intensity in combination with one of *aptitude*, *attention*, *pleasantness* or *sensitivity* [7]. All features above were included, by calculating the mean word score of a post. Drawing on the work of psycholinguistics, we included the additional two features, counts of the personal pronoun "I" and the Flesch Kincaid readability scores [38].

Text embedding All of our classifiers employ the same text embedding approach, which draws on the work of [34] and utilized Doc2Vec [18]. This approach, proposed by Le and Mikolov, builds upon their existing Word2Vec algorithm [24]. [38] found that in the context of Reddit data, using pre-compiled word embeddings had significantly lower performance compared with training the embeddings on their own Reddit data directly. A comparison on our part using fastText [14] found a similar situation.

In summary, the approach consists of mapping each word to a unique multidimensional vector and trying to predict the next word in the sentence. The Doc2Vec approach also maps each paragraph to a unique vector. Both vectors

are then concatenated to predict the next word in a context. A number of different variations of this approach have been proposed [18]. We compared a set of these approaches on error rate based on a logistic regression model trained on the paragraph embeddings.

In our training, the approach with the lowest error rate was a combination of the Distributed Bag of Words (DBOW) and Distributed Memory (DM). This combination has been proposed as an optimal method by [18]. It encompasses the DM method, which is explained above and a DBOW version of Doc2Vec, where the word vector is dropped and instead forces the model to predict words randomly sampled from the paragraph vector by using a sliding window. Parameters for both algorithms includes discounting all words which occurred less than twice in the corpus, 20 epochs and a final vector output of size 100. The two 100-dimensional vector outputs were concatenated to a joint 200-dimensional vector. In all cases, the model is trained on the whole corpus.

3.3 Suicidal Ideation Classifier

Data A limited number of publications have gained access to small collections of suicide notes and performed some basic linguistic analysis on them [31, 32]. Our approach is based around a public subreddit titled `/r/SuicideNotes` (SN). This subreddit describes itself as "*A location to immortalize your final words, or read the last words that others have written down.*" and contains 1210 submissions as of the end of August 2018.

To enhance the validity of this data, we drew upon the work of [37] who identified users with suicidal tendencies by applying a time series approach to users who posted on certain mental health subreddits and then on the `/r/SuicideWatch` subreddit. Our goal was to assess the validity of these posts. To achieve this, we selected each user (738) who had posted a note on SN, and selected all their historical posts using the complete Reddit dataset [23]. We began by removing users who only had a single post ever (throwaway accounts), we then only selected users whose last post ever was on SN. This gave us a list of 112 users who met the following two conditions, i.e., *(i)* history of posting to Reddit and *(ii)* had posted to SN and had never interacted with Reddit again. We made the assumption that these users may have actually carried out the actions of their suicide note.

We performed a brief manual evaluation and for the majority of these users, we saw a noticeable descent towards suicidal actions, characterised by increased posting on subreddits focused on mental health supports. Table 2 shows the last 10 subreddits a suicidal user account made a submission on. On average each user account made 16 submissions to Reddit, while the subreddits with the most submissions from this cohort are `/r/AskReddit, /r/SuicideNotes, /r/Depression` and `/r/SuicideWatch`.

A total of 1,502 Reddit posts labelled as having come from a suicidal user where extracted. We randomly extracted 1,500 more posts from Reddit to use as a control group. To further validate the nature of the posts, we refer to the work of [32] on syntactic features associated with suicide notes. Notably, we would expect to see a higher account of the personal pronoun "I" in suicide notes compared to control posts, which we do see in Figure 1.

Table 2: A sample post history from a suicidal Reddit user.

Date	Subreddit
2016-07-05	Pools
2016-07-12	Minecraft
2016-08-18	Depression
2016-09-21	Longisland
2017-03-07	Selfharm
2017-04-17	Anxiety
2017-04-29	OpenAdoption
2017-06-13	Kittens
2017-07-07	Depression
2017-08-05	Suicidenotes

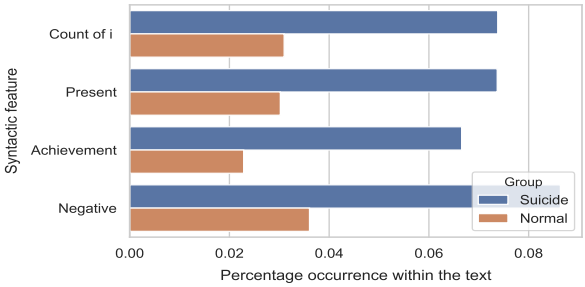

Fig. 1: Differences in syntactic features for suicide note dataset.

Feature creation & text embedding For each post, we extracted the following features: 78 linguistic features related to psychology extracted using the LIWC lexicon. We found the optimal Doc2Vec text embedding approach to be a single DBOW method with a setup of 100 dimensions, a minimum word occurrence count of two and 20 epochs.

3.4 Insomnia / Hypersomnia Classifier

Data Our dataset for this classifier is all posts on the `/r/Insomnia` subreddit, which describes itself as "*Posts and discussion about insomnia and sleep disorders.*". We collected 40,000 posts from `/r/Insomnia` and an additional 40,000 random posts from `/r/AskDocs` a subreddit focusing on general medical related questions. In the results stage, we find an abnormally high degree of accuracy suggesting overfitting of the data. We compensated this by adding 20,000 randomly selected Reddit posts to each class as noise.

Feature creation & text embedding We followed a similar approach to the suicidal ideation classifier, for each post we extracted 78 LIWC linguistic features. Text embedding was Doc2Vec with a DBOW approach providing the lowest error rate. Our total dataset was thus 120,000 posts, each of which had 178 features.

3.5 Weight Change Classifier

Data The DSM-V defines significant weight change (+/-5% of body weight in a month) as a symptom of MDD. For a psychiatrist to make an accurate analysis of this symptom they would require a body weight measurement at time one and time two. Since our conversational system is only designed as an initial classification approach, we are expecting this classifier will give an indication (positive class) that the individual is talking about weight change or in the case of a negative class prediction, the individual has made no mention of aspects related to weight change. Data was collected from the `r/Loseit` subreddit, a community dedicated to weight change. 80,000 posts were collected and labelled as belonging to the positive class and another 80,000 from the `/r/AskDocs` subreddit and labelled as belonging to the negative class. Additional 40,000 posts were randomly allocated between classes to prevent overfitting and simulate noise.

Feature creation & text embedding The exact same process that we applied to our sleep classifier was taken here. A 178 feature space was created with 100 Doc2Vec DBOW approach and 78 LIWC features.

3.6 Excessive or Inappropriate Guilt

Approach We found no existing research published on the syntactic features associated with inappropriate guilt in the English language, additionally we found no specific way to isolate guilt related data on Reddit to use. The developers of Linguistic Inquiry with Word Count (LIWC) suggested that guilt can be recognised in certain cases from a combination of negative emotions and anxiety [29]. Both of these are features extractable with LIWC and their sum can serve as a noisy proxy for guilt in a post. We presented this count directly as an indicator of guilt and performed no further modelling.

4 Methodology

In this section, we give an overview of the algorithms employed in the development of our classifier.

4.1 Depressed Mood Classifier

The primary issue affecting the development of a depressed mood classifier within this context is the uneven class balance seen in the eRisk dataset. With the majority class (non-depressed) occupying 91% of the dataset. Our first approach was a logistic regression attempt with a modified class weight approach proposed by [16]. This approach provided little accuracy above the majority class baseline (2%). Therefore we applied a more forceful approach consisting of a synthetic minority oversampling technique (SMOTE) [8]. This approach consists of oversampling the minority class to create an artificial dataset with an even class balance. We then apply a Random Forest classifier tuned on a grid search method (class weight= 'balanced subsample', bootstrap= 'false', criterion= 'entropy',n_estimators= 9).

4.2 Suicidal Ideation Classifier

This model incorporated a logistical regression classifier. Our choice of this approach was due to its percentage outcome. As [16] suggests, rather than considering binary outcome as two independent events, we can consider the outcome as an unobserved continuous variable. In this case, the propensity of an individual to attempt suicide. We applied an L2 penalty, with balanced class weights and all features underwent standard scaling from -1 to +1.

Threshold values Threshold values allow for distinction of binary classes when working on a continuous scale. The allocation of a threshold value is often considered important in medical literature where there might be a consideration to knowingly over or under predict certain classes. The most naive approach is often maximizing the area under the curve when sensitivity is plotted against 1-specificity. Although this can give the most balanced class allocation, we would consider reducing false negative predictions to be of reasonable importance. To do this, we chose to set a specificity value of 0.95 which allocates us a threshold value of 0.55 sensitivity value of 0.61 and a Youden's index of 0.50 (Figure 2).

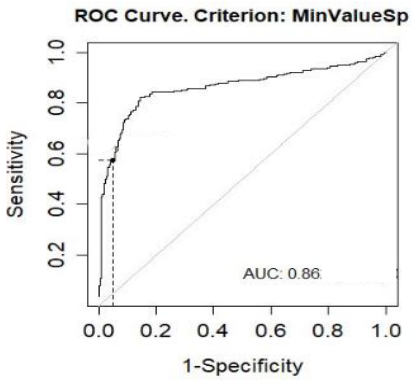

Fig. 2: Plot of sensitivity against 1-specificity for suicidal ideation classifier trained on Reddit data. (AUC = area under the curve)

4.3 Sleep Change & Weight Change Classifiers

Both classifiers employed a logistic regression algorithm with L2 penalization. Independent grid searches were applied but resulted in the same set of optimized parameters. These were intercept scaled to 1, balanced class weight and the SKLearn C parameter which is the inverse of the regularization strength was set to 1 as well [28].

Threshold values & scoring Although false negative predictions for sleep or weight change is not as serious as missing attempts at suicide, we still optimized the threshold value of the logistic regression by setting the specificity value to 0.95. Cutoff values for the sleep change and weight change classifiers are 0.72, and 0.74 respectfully.

4.4 Conversational System Interface Design

A demo interface[5] (Figure 3) was developed which combined the conversational model and classifiers. The interface consisted of a text field where users can write a comment which is sent to the conversational model. The reply to this input is shown on the screen to the user after the twentieth interaction the results of the symptom prediction models are shown.

Fig. 3: Demo interface for the conversational system. Graphs are only displayed after 20 iterations.

5 Evaluation

In this section, we begin by reporting on the metric evaluation scores acquired during the training of the models. The second section reports on the overall evaluation process we employed in the project.

5.1 Classifier Evaluation

Metric scores during training for each of the four classifiers we developed are presented in table 3. In all cases, the scores presented are mean scores following 10 fold cross validation performance on a withheld test set composed of a random 20% sample of the original dataset. The depressed mood classifier however employed the SMOTE balanced subsample approach.

5.2 Overall Project Evaluation Procedure

We recruited seven participants as a convenience sample. All participation was anonymous and voluntary, no demographic details were collected. Initially, participants were instructed to have a short interaction with the conversational system. Beginning with answering the question "How are you? Please consider how you've been feeling over the past two weeks including today.". Participants

[5] `http://server1.nlp.insight-centre.org/marvin/demo.html`

Table 3: Metric scores for all four classifiers when trained on their Reddit own dataset.

Classifier	Class	Precision	Recall	F1-Score
Depressed mood	Presence	0.91	0.95	0.93
	Absence	0.95	0.91	0.93
	Total	0.93	0.93	0.93
Suicidal ideation	Presence	0.84	0.61	0.71
	Absence	0.73	0.90	0.81
	Total	0.78	0.77	0.76
Sleep disorder	Presence	0.69	0.95	0.80
	Absence	0.93	0.62	0.74
	Total	0.82	0.77	0.77
Weight change	Presence	0.71	0.95	0.81
	Absence	0.92	0.61	0.74
	Total	0.82	0.78	0.78

could end the conversation at any stage by exiting out of the conversation, but were asked to try and hold the conversation for at least 20 messages.

The following step of the evaluation was to establish ground truths. Participants were asked to complete the *Beck's Depression Inventory-II* [5], which is a 21 item multiple response questionnaire that ranks participants on a scale of no indication of MDD to an strong indication of MDD. The advantage of the Beck's Inventory is that it is a short and highly standardised instrument that has seen applications across a wide range of research studies [2, 11].

5.3 Overall Project Evaluation Results

We propose two metric evaluation approaches for our project, *(i)* evaluates classifiers individually, *(ii)* overall evaluation. For the individual classifier evaluation, we established ground truths by dividing each question on the Beck's Inventory to a specific symptom it was investigating as per table 4. A score above two on this question was considered a positive score (presence of a symptom), while a score less than two was considered negative (absence of a symptom). We then considered the classifier to have returned a positive prediction if the mean prediction score was above the threshold value. Row two, three and four in table 5 provides the metric scores for each classifier[6].

The second evaluation process, which investigated the overall accuracy was computed by considering a score above 19 on the Beck's Inventory as ground truth presence of an MDD. If two out of the four classifier provided a positive prediction, this was considered an overall positive prediction of depression. Results are presented in the first row of table 5.

6 Conclusions

Over the course of the above text, we initially began by demonstrating why major depressive disorder is worthy of study, and how passive diagnoses is an important

[6] Neither Suicide ideation nor weight change are included as all scores are equal to one.

Table 4: Beck's Inventory questions matched to the symptoms they investigate.

Question number(s)	Type
9	Suicidal Ideation
5	Guilt
18	Weight change
16	Sleep change
1,2,3,4,10,14	Depressed mood

Table 5: Evaluation scores for each symptom classifier. Dataset from the evaluation process. Sample size of seven participants

Classifier	Precision	Recall	F1-Score
Overall	1.00	0.86	0.92
Depressed mood	0.73	0.86	0.79
Sleep change	0.51	0.71	0.60
Guilt	1.00	0.71	0.83

future issue for MDD. Our proposed approach builds on that of the existing machine learning communities contributions to MDD, whereby our methodology is the first to view MDD prediction on a symptomatic level.

In addition to a theoretical proposal, we hope our work may lead to a future practical application. We demonstrate the potential possibility of using these conversational systems to provide instant support to individuals and engage individuals until they can be handed over to a trained operator. The integration of the classifiers allows an operator to instantly see if an individual has displayed certain symptoms.

Within the scope of the work, we note two key limitations that must be addressed in our future works. Initially, the ever-present problem of suitable data presents itself. It can be accepted that our datasets for both the conversational system and classifiers were primarily created out of convenience. [21] explores the advantages and disadvantages with regards to different depression related data collection approaches. In all cases, none of our labelled data actually employs medical diagnoses. Therefore, we can not be completely confident our labelled data is representative of the actual disorder. Future research will need to explore this issue to infer more detailed results.

A final limitation concerns that of the evaluation stage. We accept that a sample size of seven individuals is quite limited in its evaluation scope. Nevertheless, we feel this is sufficient as a proof of concept study, and give direction for future work. For the sleep change classifier, we see low precision and high recall scores, indicating the possibility that the threshold value has been set to high. No one in our sample indemnified as having the presence of suicidal ideation or negative weight change, and respectfully our classifiers did not predict any false positives in these cases.

In conclusion, within the scope of our limited sample size, we are positive regarding the results of four of our classifiers, and suggest a revaluation of the threshold value assigned to one. Ultimately, we hope to see that individuals suffering an MDD episode will no longer suffer alone but rather will have more rapid and easy access to diagnostic services and thus receive timely.

Acknowledgement. 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

[1] Idriss Abdou Malam et al. "IRIT at e-Risk". In: *International Conference of the CLEF Association, CLEF 2017 Labs Working Notes (CLEF 2017)*. Vol. 1866. Dublin, Ireland: CEUR Workshop Proceedings, 2017, pp. 1–7.

[2] Randolph C Arnau et al. "Psychometric evaluation of the Beck Depression Inventory-II with primary care medical patients." In: *Health Psychology* 20.2 (2001), p. 112.

[3] American Psychiatric Association et al. "Mental Hospital Service (1952)". In: *Diagnostic and Statistical Manual (DSM)* (1952).

[4] DSM-5 American Psychiatric Association et al. "Diagnostic and statistical manual of mental disorders". In: *Arlington: American Psychiatric Publishing* (2013).

[5] Aaron Beck. "Beck Depression Inventory". In: (1996).

[6] RH Belmaker and Galila Agam. "Major depressive disorder". In: *New England Journal of Medicine* 358.1 (2008), pp. 55–68.

[7] Erik Cambria et al. "SenticNet 5: Discovering Conceptual Primitives for Sentiment Analysis by Means of Context Embeddings". In: *Aaai* (2018), pp. 1795–1802.

[8] Nitesh V. Chawla et al. "SMOTE: Synthetic minority over-sampling technique". In: *Journal of Artificial Intelligence Research* 16 (2002), pp. 321–357. ISSN: 10769757.

[9] Cindy Chung and James W Pennebaker. "The psychological functions of function words". In: *Social communication* 1 (2007), pp. 343–359.

[10] Munmun De Choudhury, Scott Counts, and Eric Horvitz. "Social media as a measurement tool of depression in populations". In: *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci '13* (2013), pp. 47–56. ISSN: 9781450318891.

[11] David JA Dozois, Keith S Dobson, and Jamie L Ahnberg. "A psychometric evaluation of the Beck Depression Inventory–II." In: *Psychological assessment* 10.2 (1998), p. 83.

[12] William P Evans, Laura Davidson, and Lorie Sicafuse. "Someone to listen: Increasing youth help-seeking behavior through a text-based crisis line for youth". In: *Journal of Community Psychology* 41.4 (2013), pp. 471–487.

[13] Yen-Hao Huang, Lin-Hung Wei, and Yi-Shin Chen. "Detection of the Prodromal Phase of Bipolar Disorder from Psychological and Phonological Aspects in Social Media". In: (2017).

[14] Armand Joulin et al. "Bag of Tricks for Efficient Text Classification". In: *arXiv preprint arXiv:1607.01759* (2016).

[15] Christian Karmen, Robert C Hsiung, and Thomas Wetter. "Screening Internet forum participants for depression symptoms by assembling and enhancing multiple NLP methods". In: *Computer methods and programs in biomedicine* 120.1 (2015), pp. 27–36.

[16] Gary King and Langche Zeng. "Logistic Regression in Rare Events Data". In: *Political Analysis* 9.02 (2001), pp. 137–163. ISSN: 1047-1987.

[17] Guillaume Klein et al. "OpenNMT: Open-Source Toolkit for Neural Machine Translation". In: *CoRR* abs/1701.02810 (2017).

[18] Quoc V. Le and Tomas Mikolov. "Distributed Representations of Sentences and Documents". In: 32 (2014). ISSN: 10495258. URL: http://arxiv.org/abs/1405.4053.

[19] Jiwei Li et al. "Adversarial Learning for Neural Dialogue Generation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*. 2017, pp. 2157–2169.

[20] Pierre Lison and Jörg Tiedemann. "OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles". In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Ed. by Nicoletta Calzolari (Conference Chair) et al. Portorofffdfffd, Slovenia: European Language Resources Association (ELRA), May 2016. ISBN: 978-2-9517408-9-1.

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

[22] David E Losada, Fabio Crestani, and Javier Parapar. "Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations". In: *Working Notes of CLEF 2017-Conference and Labs of the Evaluation Forum*. 2017.

[23] Jason Michael. *Pushshift.io*. URL: https://pushshift.io/ (visited on 08/30/2018).

[24] Tomas Mikolov et al. "Efficient estimation of word representations in vector space". In: *arXiv preprint arXiv:1301.3781* (2013).

[25] Saif M. Mohammad. "Word Affect Intensities". In: (2017).

[26] Finn Årup Nielsen. "A new ANEW: Evaluation of a word list for sentiment analysis in microblogs". In: *CEUR Workshop Proceedings* 718 (2011), pp. 93–98. ISSN: 16130073.

[27] World Health Organization. *Depression Fact Sheet*. 2018. URL: http://www.who.int/news-room/fact-sheets/detail/depression (visited on 08/30/2018).

[28] F. Pedregosa et al. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[29] James W Pennebaker, Roger J Booth, and Martha E Francis. "Linguistic inquiry and word count: LIWC [Computer software]". In: *Austin, TX: liwc. net* (2007).

[30] James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. "Psychological aspects of natural language use: Our words, our selves". In: *Annual review of psychology* 54.1 (2003), pp. 547–577.

[31] John Pestian et al. "Sentiment Analysis of Suicide Notes: A Shared Task". In: *Biomedical Informatics Insights* 5 (2012), p. 3. ISSN: 1178-2226.

[32] John Pestian et al. "Suicide Note Classification Using Natural Language Processing: A Content Analysis." In: *Biomedical informatics insights* 2010.3 (2010), pp. 19–28. ISSN: 1178-2226.

[33] Trang Pham et al. "Deepcare: A deep dynamic memory model for predictive medicine". In: *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer. 2016, pp. 30–41.

[34] Faneva Ramiandrisoa et al. “IRIT at e-Risk 2018”. In: (2018).
[35] Andreas Rücklé and Iryna Gurevych. “Representation Learning for Answer Selection with LSTM-Based Importance Weighting”. In: *IWCS 2017 — 12th International Conference on Computational Semantics — Short papers*. 2017.
[36] Stephanie S. Rude, Eva Maria Gortner, and James W. Pennebaker. “Language use of depressed and depression-vulnerable college students”. In: *Cognition and Emotion* 18.8 (2004), pp. 1121–1133. ISSN: 02699931.
[37] Georgia Tech, Atlanta Ga, and Mark Dredze. “Discovering shifts to suicide ideation from mental health”. In: (2017).
[38] Marcel Trotzek, Sven Koitka, and Christoph M. Friedrich. “Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences”. In: (2018).
[39] Oriol Vinyals and Quoc V. Le. “A Neural Conversational Model.” In: *CoRR* abs/1506.05869 (2015).
[40] Colin G. Walsh, Jessica D. Ribeiro, and Joseph C. Franklin. “Predicting Risk of Suicide Attempts Over Time Through Machine Learning”. In: *Clinical Psychological Science* 5.3 (2017), pp. 457–469. ISSN: 21677034.
[41] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. “Norms of valence, arousal, and dominance for 13,915 English lemmas”. In: *Behavior Research Methods* 45.4 (2013), pp. 1191–1207. ISSN: 1554351X.
[42] Dirk Weissenborn, Georg Wiese, and Laura Seiffe. “Making Neural QA as Simple as Possible but not Simpler”. In: *CoNLL*. 2017.
[43] Genta Indra Winata et al. “Nora the empathetic psychologist”. In: *Proc. Interspeech*. 2017, pp. 3437–3438.
[44] Anbang Xu et al. “A new chatbot for customer service on social media”. In: *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. ACM. 2017, pp. 3506–3510.