Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health

Julia Ive[1], **George Gkotsis**[1], **Rina Dutta**[1],
Robert Stewart[1], **Sumithra Velupillai**[1,2]
King's College London, IoPPN, London, SE5 8AF, UK,[1]
KTH, Sweden[2]
{firstname.lastname}@kcl.ac.uk

Abstract

Mental health problems represent a major public health challenge. Automated analysis of text related to mental health is aimed to help medical decision-making, public health policies and to improve health care. Such analysis may involve text classification. Traditionally, automated classification has been performed mainly using machine learning methods involving costly feature engineering. Recently, the performance of those methods has been dramatically improved by neural methods. However, mainly Convolutional neural networks (CNNs) have been explored. In this paper, we apply a hierarchical Recurrent neural network (RNN) architecture with an attention mechanism on social media data related to mental health. We show that this architecture improves overall classification results as compared to previously reported results on the same data. Benefitting from the attention mechanism, it can also efficiently select text elements crucial for classification decisions, which can also be used for in-depth analysis.

1 Introduction

Mental health problems represent a major public health challenge worldwide, and the accumulation of big data offers the opportunity for improving healthcare processes, interventions, and public health policies (Stewart and Davis, 2016). Recent advances in data science, machine learning and Natural Language Processing (NLP) hold great promise in providing technical solutions for the analysis of large sets of clinically relevant information in Psychiatry (Torous and Baker, 2016). This includes not only routinely collected data such as Electronic Health Records (EHRs), but also patient-generated text or speech. Patient-generated content has been made available by social media, mainly in the form of tweets or forum posts (Névéol and Zweigenbaum, 2017; Gonzalez-Hernandez et al., 2017).

As opposed to e.g. documentation produced by healthcare professionals, social media data captures thoughts, feelings and discourse in people's own voice, and these types of data sources are becoming very important for monitoring a number of public health issues including mental health problems such as drug abuse, alcohol, and depression (De Choudhury et al., 2014; Wongkoblap et al., 2017; Conway and OConnor, 2016; Mikal et al., 2016; Sarker et al., 2016).

In this work, we address the problem of automatically classifying social media posts related to mental health derived from Reddit. Convolutional neural networks (CNNs) applied to this task have shown good performance in previous studies (Gkotsis et al., 2017). However, the performance of recurrent neural networks (RNNs) for the same task remains understudied. RNNs can be particularly beneficial in this case as they are able to model the sequential structure of text. We also attempt to explore the contribution of attention mechanisms to establishing a certain hierarchy in the sequences.

To be more precise, we apply a hierarchical RNN architecture as described in (Yang et al., 2016) to the classification of social media posts related to mental health problems, and seek to answer the following main questions: (a) Is a sequence-based model more beneficial than a CNN model for the accurate classification of social media posts? (b) Which parts of posts are more important for the classification of a post into its mental health topic as defined by the attention mechanism?

Our main contribution in this work is twofold: (1) an attempt to apply an RNN architecture to the text classification task of determining which mental health problem a post is about, which, to our

Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 69–77
New Orleans, Louisiana, June 5, 2018.

knowledge, is the first attempt of its kind. We show that the ability of RNNs to take the sequence of events reflected in the post content can be beneficial for the classification of health-related social media text; (2) we also study the results of applying an attention mechanism to pinpoint the parts of a text that are contributing more to classification decisions. Those results can be useful for an in-depth analysis, to filter out irrelevant content, and to reduce the computational costs for real-life applications. We provide a few examples, and discuss future directions in this area.

2 Related Work

Most previous work in text classification have used various classifiers (most commonly, Support Vector Machines (SVMs) (Cortes and Vapnik, 1995) relying on different sets of features such as: constructed statistics (e.g., bag-of-words (word counts)), lexical TF-IDF, Latent Dirichlet Allocation (LDA) topics (Resnik et al., 2015; Rumshisky et al., 2016)), various linguistic and metadata features (Gkotsis et al., 2016; Bullard et al., 2016).

Recently, CNNs were actively exploited for text classification in the medical domain (Baker and Korhonen, 2017; Yates et al., 2017). For instance, Yates et al. (2017) made an attempt at hierarchical classification. They merge outputs of several CNNs per post to create a representation (roughly, a feature set learned automatically) of the user activity across his/her posts.

CNNs learn to extract a hierarchy of crucial text elements. RNNs, on the other hand, handle text as a sequence. This property of RNNs can be especially beneficial to analyze health-related text, for which the order of described events can be important.

RNNs have been successfully used for document representation and consequently applied to a series of downstream NLP tasks such as topic labeling, summarization, and question answering (Li et al., 2015; Yang et al., 2016; Liu and Lapata, 2017).

As RNN architectures typically exploit an attention mechanism for hierarchical analysis, we also study whether this mechanism can provide insight into which words and sentences contribute to classification decisions. The mechanism opens a range of attractive, less costly modeling perspectives, for instance, in an attempt to replace recursion by Vaswani et al. (2017). One of the side benefits of using an attention mechanism is that the results of its application can be interpreted and provide a powerful tool for further text analysis.

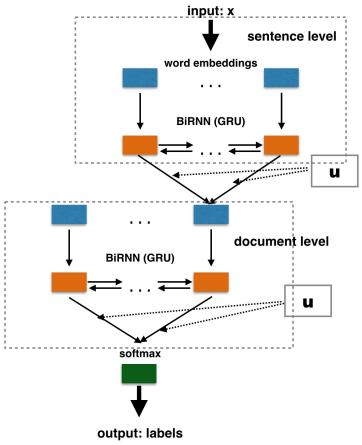

Figure 1: Hierarchical document-level architecture

3 Document-Level RNN Architecture

In our work we reproduce the hierarchical document classification architecture (HIERRNN) as proposed by Yang et al. (2016). This architecture progressively builds a document representation from its sentence representations, which in turn are composed of the representations of the words they contain. Those document representations are directly used by the architecture to make classification decisions.

To do so, the architecture implies a series of RNN encoders. The **encoder** reads an input sequence of words $X = \{x_1 \dots x_J\}$ and calculates a forward sequence of hidden states $(\overrightarrow{h}_1, ..., \overrightarrow{h}_J)$, and a backward sequence of hidden states $(\overleftarrow{h}_1, ..., \overleftarrow{h}_J)$. The hidden states $\overrightarrow{h}_j$ and $\overleftarrow{h}_j$ are concatenated to obtain the resulting representation h_j.

To be more precise, the architecture contains bidirectional encoders, modeling sentences of a document $d = \{x_1 \dots x_T\}$. Each sentence vector can be computed out of word representations: average, maximum, sum etc. We compute a weighted sum of those representations as weighted by the attention mechanism. Those vectors are input to the document encoder. The resulting document vector (again computed out of sentence representations) is in turn input to the `softmax` layer over document labels (see Figure 1).

The attention mechanism is used to weight aggregated representations. More formally, an atten-

tion function consists in mapping a query and a set of key-value pairs to an output. The output is a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of the query with the corresponding key.

To detect both words and sentences that are important to the meaning of a document we employ the **hierarchical attention mechanism**:

$$\alpha_j = \frac{\exp(u_j^\top u_g)}{\sum_{k=1}^{J} \exp(u_k^\top u_g)}, \quad (1)$$

where

$$u_j = t(h_j), \quad (2)$$

where $t(\cdot)$ is a non-linear activation function (`tanh` in our case). The importance of a unit is thus measured as the similarity of u_j to the context vector u_g, jointly learned during the training process. This vector serves a query. The importance weight is normalized importance through a `softmax` function. The document vector is thus computed as follows:

$$v = \sum_{j=1}^{J} \alpha_j h_j \quad (3)$$

4 Experimental Setup

We study the performance of the hierarchical architecture on the task of classifying posts from social media related to mental health.

4.1 Data

We use a dataset of posts from the social media platform Reddit. Each entry has been posted to a so called *subreddit* – a topic-specific community within the platform. We use the posts and subreddits related to 11 mental health problems (i.e. a multiclass classification problem) that have been previously identified and used for text classification (Gkotsis et al., 2016, 2017).[1] In total, the dataset consists of 538,272 posts, with an imbalanced distribution per mental health topic (ranging from 4,360 posts in *addiction* to 197,436 in *depression*). The data and the mental health topics are described in detail in (Gkotsis et al., 2017). The 11 mental health topics are listed in Table 2.

[1]Data was obtained through the corresponding author of these studies and stored on encrypted computers.

4.2 Implementation Details

We implemented our document-level architecture using the `Keras` toolkit with Gated Recurrent Units (GRUs) (Cho et al., 2014) as RNNs. We followed the implementation details in Yang et al. (2016): the word embedding dimensionality is set to 200. The size of the hidden units of the encoder is 50. We set the input vocabulary size to 30K. We limit the sentence length to 70 tokens as standard in downstream NLP tasks (Hewlett et al., 2017). We fixed the size of a document to 17 sentences (empirically chosen value, which corresponds to the third quartile of the overall distribution of sentence length values), shorter documents were extended with dummy sentences. For training, we use a mini-batch size of 70. We use stochastic gradient descent to train all models with momentum of 0.9. We train the system to minimize the categorical cross-entropy loss and choose the best learning rate using grid search.

As our dataset is highly imbalanced we provide the system with class weights computed as inversely proportional to each class frequency.

4.3 Evaluation

We compare our results for HIERRNN with the attention mechanism (`RNN-att`) to two other configurations, where we a) take a maximum of vectors (`RNN-max`) or b) an average of vectors (`RNN-av`) at both word and sentence levels. We also compare our results to a baseline result reported by Gkotsis et al. (2017) for a CNN-based architecture (`CNN`). This architecture is a rather simple architecture with 5 layers: an embedding layer, a convolution layer (a filter window of 5), a max-pooling layer, a fully-connected layer and an output sigmoid layer. The results are directly comparable as performed for the same data split.

In terms of evaluation metrics we use the standard set of precision (PR), recall (RC) and F-measure (FM). In addition, we manually review a random sample of the results from the attention mechanism, and provide a few paraphrased examples (Benton et al., 2017).

5 Results

Results of our experiments are presented in Tables 1 and 2. All the three HIERRNN configurations yield an improvement over `CNN`: with a minor improvement of 1 FM for `RNN-av`, 2 FM for `RNN-max` and the highest improvement of 4

FM for RNN-att. Thus, we believe that considering the sequential characteristic of text, as done by RNN models, can be beneficial for analyzing posts related to mental health.

We should also note the improvement due to the attention mechanism as compared to the maximum and averaging strategies (on average 2.5 FM). Those results are consistent with the results presented by Yang et al. (2016) for other types of texts (e.g., reviews) and other types of labels (e.g., ratings).

As for per class performance, RNN-att improves this performance by 6 FM on average. The improvement in precision is twice as low as the improvement in recall (6% relative change in PR vs. 12% in RC). This difference is particularly remarkable for more rare classes. We tend to attribute this to intrinsic properties of RNNs (see Table 2).[2]

A relatively high performance improvement of 8 FM is observed for the 8 classes of posts (*BPD*, *bipolar*, *schizophrenia*, *selfharm*, *addiction*, *cripplingalcoholism*, *Opiates*, *autism*), which are under represented (on average represent 4% of all the test set posts) and with a relatively low document length (9 sentences on average vs. 11 sentences for all the classes). Except for intrinsic properties of RNNs, our modeling approximation (we limit the document size to 17 sentences to avoid optimization issues) could also contribute to this improvement.

As can be seen from the confusion matrix in Figure 2 the intrinsic overlap of post content across the themes can be misleading for classification: e.g., and again, as shown by Gkotsis et al. (2017), a lot of *Opiates* posts are misclassified as *cripplingalcoholism* and vice versa. However, HIERRNN is in general more precise and reveals less confusion between classes: e.g., the amount of confusion for *schizophrenia* with *depression* has reduced twice as compared to CNN.

One of the advantages of the attention mechanism is that its weights can be visualized and interpreted by humans (which is not always the case with neural network layers). In this work, we focus on the analysis of sentence-level attention weights. This information can be especially helpful for reducing the quantity of analyzed post sentences to create less costly classification solutions.

[2]To confirm this conclusion, we also performed a series of control experiments without assigning class weights, which still resulted in similar results.

	PR/RC/FM
CNN	0.72 / 0.71 / 0.72
RNN-av	0.74 / 0.73 / 0.73
RNN-max	0.74 / 0.74 / 0.74
RNN-att	0.76 / 0.76 / 0.76

Table 1: F-Measure (FM) weighted average results (PR refers to precision, RC – to recall)

Table 3 provides results of our analysis of attention weights distributions. For this analysis we filtered out one-sentence documents. We study how often an absolute sentence position receives a maximum or a minimum weight from the total amount of cases this position is present across documents (a document is long enough). We report top three maximum and minimum positions. We also report average entropy values for the distributions per sentence. [3]

We also report similar statistics for a selection of classes in Table 4.

Our analysis shows that RNN-att is able to distinguish a certain semantic importance pattern: the most attention is paid to the first, then to the second and finally last sentences. The least attention is systematically paid to a sentence after a peak attention at the beginning (4th sentence), to a sentence in the middle (7th position) and to a sentence before the end (14th position).

At the same time, attention weights are quite equally spread between peak positions (average entropy of 1.93). The entropy values tend to increase for the classes that are better represented and for which posts are on average longer (e.g., *depression*, *suicidewatch*). Relevant information is not concentrated in those longer documents and several sentences are likely to be equally important.

Table 5 provides some examples of attention distributions for documents of different lengths and belonging to different classes. So that, for a longer document from *suicidewatch* the most relevance is given to the first 2 sentences containing words like "rejection" and "depression", whereas a neutral sentence "I met this girl." receives a low

[3]Note that this analysis could have been performed in a different way: e.g., for relative positions, first or last sentence; or taking the fixed document length into account. However, such analysis would be biased since dummy sentences from padded documents tend to receive less attention than actual sentences.

				PR/RC/FM		
Theme	%	$\bar{l}_{doc}$	$\bar{l}_{sent}$	CNN	RNN-max	RNN-att
BPD	2%	14	19	0.88 / 0.46 / 0.60	0.84 / 0.52 / 0.64	**0.87 / 0.53 / 0.66**
bipolar	8%	13	18	0.77 / 0.60 / 0.67	0.73 / 0.67 / 0.70	**0.79 / 0.68 / 0.73**
schizophrenia	1%	11	19	0.75 / 0.48 / 0.58	0.78 / 0.60 / 0.67	**0.82 / 0.59 / 0.69**
Anxiety	11%	13	19	0.83 / 0.75 / 0.79	0.79 / 0.81 / 0.80	0.89 / 0.76 / 0.82
depression	37%	16	18	0.70 / 0.77 / 0.73	0.72 / 0.76 / 0.74	0.73 / 0.81 / 0.76
selfharm	3%	11	17	0.70 / 0.58 / 0.64	0.72 / 0.67 / 0.70	**0.76 / 0.67 / 0.71**
suicidewatch	17%	17	17	0.62 / 0.59 / 0.61	0.62 / 0.60 / 0.61	0.65 / 0.61 / 0.63
addiction	0.8%	6	17	0.72 / 0.41 / 0.52	0.76 / 0.41 / 0.53	**0.75 / 0.51 / 0.60**
cripplingalcoholism	8%	7	15	0.68 / 0.76 / 0.72	0.83 / 0.77 / 0.80	**0.73 / 0.86 / 0.79**
Opiates	12%	9	17	0.76 / 0.86 / 0.80	0.82 / 0.89 / 0.85	**0.88 / 0.88 / 0.88**
autism	0.2%	5	18	0.84 / 0.71 / 0.77	0.90 / 0.80 / 0.85	**0.86 / 0.85 / 0.86**
all	100%	11	18	0.72 / 0.71 / 0.72	0.74 / 0.74 / 0.74	**0.76 / 0.76 / 0.76**

Table 2: Multiclass classification evaluation results (we indicate the percentage of posts belonging to a class in the sample; $\bar{l}_{doc}$ refers to average document length in sentences; $\bar{l}_{sent}$ – average sentence length in tokens; FM refers to F-measure; PR – to precision; RC – to recall;)

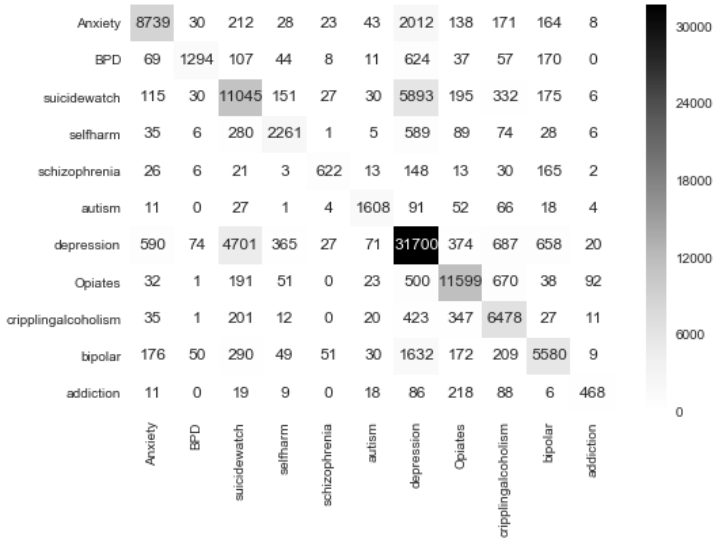

Figure 2: Multiclass classification confusion matrix: RNN-att

max		min		
position	% of occurrences	position	% of occurrences	H
1	86	7	27	1.93
2	13	4	23	
17	1	14	22	

Table 3: Absolute sentence positions that receive the most and the least attention. We provide top three positions with the percentage of their occurrences that received maximum or minimum attention. E.g.: 2nd sentence receives the most attention in 13% of the cases a post contains a 2nd sentence. H refers to entropy

	max		min		
theme	position	% of occurrences	position	% of occurrences	H
schizophrenia	1	75	7	32	1.76
	2	17	4	25	
	3	3.4	14	19	
depression	1	88	7	26	2.04
	2	10	14	22	
	3	0.9	4	22	
suicidewatch	1	87	7	25	2.07
	2	10	4	22	
	17	2.5	14	21%	
cripplingalcoholism	1	83	3	30	1.59
	2	14	4	29	
	6	2	7	28	
Opiates	1	83	7	30	1.66
	2	14	14	26	
	17	4	4	24	

Table 4: Absolute sentence positions that receive the most and the least attention: selection of classes. We provide top three positions with the percentage of their occurrences that received maximum or minimum attention. E.g.: for *Opiates* the 17th sentence receives maximum attention in 4% of the cases a post contains a 17th sentence. H refers to entropy.

weight	*suicidewatch*	weight	*cripplingalcoholism*
0.20	deal with rejection i 'm young .	0.85	best part of my morning that was not an open bottle you left for your sober self , it 's the jar you pissed in .
0.23	i 'm depressed .	0.15	tasted like nothing , cheers !
0.08	i 've already tried to do it .		
0.05	i met this someone .		
0.07	she kinda become everything to me and i just got rejected .		
0.07	i went walking and i was crossing the street hoping for someone to hit me i guess .		
0.05	sorta a stupid way to do it .		
0.09	i 'm back home but i 'm just really sad .		
0.08	i did n't meet anyone for more than 10 years because i thought i could n't handle rejection i now i think i was right .		
0.08	good night everyone .		
weight	*opiates*	weight	*schizophrenia*
0.41	[medication] heloooo , have n't posted here in a long long time after not having used in a while , but now i need some advice .	0.64	hearing voices or are these just thoughts ?
0.37	i bought massive amounts of pills recently , including [medication] , [medication] (ir + er) , which obviously gives me the time of my life .	0.24	i 've always heard random nonsense and noises - phrases that have no meaning and that are connected to nothing .
0.10	can anyone tell me how to stop the prolonged pill release to make it instant ?	0.12	how can i actively understand that these are thoughts and not something wrong with me ?
0.12	thanks !		

Table 5: Paraphrased examples of attention weights distributions over post sentences. Medication names have been replaced with [medication]

weight. For a short document of 2 sentences from *cripplingalcoholism* 3/4 of the weight is concentrated on the 1st sentence. This sentence is especially relevant to the topic and contains keywords such as "beer" and "sober".

Note that, for instance, for a *schizophrenia* post (a class for which performance was significantly improved by 10 FM as compared to CNN) the elaboration of the topic of auditory hallucinations in the first two sentences might have been taken into account by RNNs.

However, RNNs usually require more computational power to be trained than other neural architectures.[4] We believe that such information on attention distributions can be particularly useful for the creation of low-resource models, which could operate with filtered data (e.g., only two first sentences of a post).

6 Discussion and Conclusions

In this paper, we have applied a hierarchical Recurrent Neural Network (RNN) architecture to the classification of posts related to mental health, which is, to our knowledge, is the first attempt of the kind. The ability to classify posts in this manner is the first step towards targeted interventions, e.g. by redirecting posts requiring moderator attention.

Our model progressively builds a document representation: it aggregates important words into sentence vectors and then aggregates important sentence representations to document representations, directly used for inference.

We have shown that the intrinsic ability of RNNs to consider input in its sequence in general, and the hierarchical structure of this architecture specifically can be beneficial for the analysis of health-related online text. We observed a performance improvement of 4 F-measure (FM) as compared to Convolutional Neural Network (CNN) solutions. This improvement is mainly due to the performance improvement for more rare classes (8 FM on average).

We have also shown that the attention mechanism is capable to efficiently distinguish words and sentences of a document relevant for classification decisions. We provided a detailed study of attention distribution patterns at the sentence level and showed that the beginning of a document, as well as the last sentence are the most important. At the same time, attention tends to be equally distributed between those positions.

In the future, we plan to reproduce our study for other types of health-related text, including Electronic Health Records (EHRs), where the sequence of events can be even more important for classification decisions. We also plan to investigate attention weights at the word level and compare those results to the results produced using state-of-the-art weighting techniques, e.g., TF-IDF.

We also plan to systematically compare performance of different attention mechanisms with the purpose of finding a robust solution able to replace the computationally expensive recursion step.

[4] Depending on the type of word and sentence vector approximation, HIERRNN takes around from 30 minutes up to 1 hour to train on a 12G GeForce TITAN X NVIDIA GPU.

References

Simon Baker and Anna Korhonen. 2017. Initializing neural networks for hierarchical multi-label text classification. In *BioNLP 2017*. Association for Computational Linguistics, Vancouver, Canada,, pages 307–315. http://www.aclweb.org/anthology/W17-2339.

Adrian Benton, Glen Coppersmith, and Mark Dredze. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*. Association for Computational Linguistics, Valencia, Spain, pages 94–102. http://www.aclweb.org/anthology/W17-1612.

Joseph Bullard, Cecilia Ovesdotter Alm, Xumin Liu, Qi Yu, and Rubén Proaño. 2016. Towards early dementia detection: Fusing linguistic and non-linguistic clinical data. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA, pages 12–22. http://www.aclweb.org/anthology/W16-0302.

Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *CoRR* abs/1406.1078. http://arxiv.org/abs/1406.1078.

Mike Conway and Daniel OConnor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology* 9:77 – 82. Social media and applications to health behavior. https://doi.org/https://doi.org/10.1016/j.copsyc.2016.01.004.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20(3):273–297. http://dx.doi.org/10.1007/BF00994018.

Munmun De Choudhury, Scott Counts, Eric J. Horvitz, and Aaron Hoff. 2014. Characterizing and predicting postpartum depression from shared facebook data. In *Proceedings of the 17th ACM Conference on Computer Supported Cooperative Work & Social Computing*. CSCW '14, pages 626–638. https://doi.org/10.1145/2531602.2531675.

George Gkotsis, Anika Oellrich, Tim Hubbard, Richard Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*. Association for Computational Linguistics, San Diego, CA, USA, pages 63–73. http://www.aclweb.org/anthology/W16-0307.

George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim J P Hubbard, Richard J B Dobson, and Rina Dutta. 2017. Characterisation of mental health conditions in social media using Informed Deep Learning. *Sci Rep* 7:45141.

G. Gonzalez-Hernandez, A. Sarker, K. O'Connor, and Savova G. 2017. Capturing the patient's perspective: a review of advances in natural language processing of health-related text. *Yearbook of Medical Informatics* (1):214–227. https://doi.org/10.15265/IY-2017-029.

Daniel Hewlett, Llion Jones, Alexandre Lacoste, and izzeddin gur. 2017. Accurate supervised and semi-supervised machine reading for long documents. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2011–2020. https://www.aclweb.org/anthology/D17-1214.

Jiwei Li, Thang Luong, and Dan Jurafsky. 2015. A hierarchical neural autoencoder for paragraphs and documents. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics, Beijing, China, pages 1106–1115. http://www.aclweb.org/anthology/P15-1107.

Yang Liu and Mirella Lapata. 2017. Learning structured text representations. *CoRR* abs/1705.09207. http://arxiv.org/abs/1705.09207.

Jude Mikal, Samantha Hurst, and Mike Conway. 2016. Ethical issues in using twitter for population-level depression monitoring: a qualitative study. *BMC Medical Ethics* 17(1):22. https://doi.org/10.1186/s12910-016-0105-5.

A. Névéol and P. Zweigenbaum. 2017. Making sense of big textual data for health care: Findings from the section on clinical natural language processing. *Yearb Med Inform* 26(01):228–233. https://doi.org/10.15265/IY-2017-027.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 2015. Beyond lda: Exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Association for Computational Linguistics, Denver, Colorado, pages 99–107. http://www.aclweb.org/anthology/W15-1212.

A Rumshisky, M Ghassemi, T Naumann, P Szolovits, V M Castro, T H McCoy, and R H Perlis. 2016. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry* 6(10):e921–. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC5315537.

Abeed Sarker, Karen O'Connor, Rachel Ginn, Matthew Scotch, Karen Smith, Dan Malone, and Graciela Gonzalez. 2016. Social media mining for toxicovigilance: Automatic monitoring of prescription medication abuse from twitter. *Drug Safety* 39(3):231–240. https://doi.org/10.1007/s40264-015-0379-4.

Robert Stewart and Katrina Davis. 2016. 'big data' in mental health research: current status and emerging possibilities. *Social Psychiatry and Psychiatric Epidemiology* 51(8):1055–1072. https://doi.org/10.1007/s00127-016-1266-8.

John Torous and Justin T. Baker. 2016. Why Psychiatry Needs Data Science and Data Science Needs Psychiatry: Connecting With Technology. *JAMA psychiatry* 73(1):3–4. https://doi.org/10.1001/jamapsychiatry.2015.2622.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR* abs/1706.03762. http://arxiv.org/abs/1706.03762.

Akkapon Wongkoblap, A. Miguel Vadillo, and Vasa Curcin. 2017. Researching mental health disorders in the era of social media: Systematic review. *J Med Internet Res* 19(6):e228. https://doi.org/10.2196/jmir.7215.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1480–1489. http://www.aclweb.org/anthology/N16-1174.

Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Copenhagen, Denmark, pages 2968–2978. https://www.aclweb.org/anthology/D17-1322.