World Wide Web manuscript No.
(will be inserted by the editor)

Explainable Depression Detection with Multi-Aspect Features Using a Hybrid Deep Learning Model on Social Media

Hamad Zogan · Imran Razzak · Xianzhi Wang · Shoaib Jameel · Guandong Xu*

Received: date / Accepted: date

Abstract The ability to explain why the model produced results in such a way is an important problem, especially in the medical domain. Model explainability is important for building trust by providing insight into the model prediction. However, most existing machine learning methods provide no explainability, which is worrying. For instance, in the task of automatic depression prediction, most machine learning models lead to predictions that are obscure to humans. In this work, we propose explainable Multi-Aspect Depression Detection with Hierarchical Attention Network **MDHAN**, for automatic detection of depressed users on social media and explain the model prediction. We have considered user posts augmented with additional features from Twitter. Specifically, we encode user posts using two levels of attention mechanisms applied at the tweet-level and word-level, calculate each tweet and words' importance, and capture semantic sequence features from the user timelines (posts). Our hierarchical attention model is developed in such a way that it can capture patterns that leads to explainable results. Our experiments show that **MDHAN** outperforms several popular and robust baseline methods, demonstrating the effectiveness of combining deep learning with multi-aspect features. We also show that our model helps improve predictive performance when detecting depression in users who are posting messages publicly on social media. **MDHAN** achieves excellent performance and ensures adequate evidence to explain the prediction.

Keywords depression detection · social network · deep learning · machine learning · explainability

1 Introduction

Mental illness is a serious issue faced by a large population around the world. In the United States (US) alone, every year, a significant percentage of the adult population is affected by different mental disorders, which include depression mental illness (6.7%), anorexia and bulimia nervosa (1.6%), and bipolar mental illness (2.6%) [1]. Sometimes mental illness has been attributed to

* Corresponding authors

Authors Address:
Hamad Zogan
University of Technology Sydney (UTS), Australia
Jazan University, Saudi Arabia
E-mail: hamad.a.zogan@student.uts.edu.au

Imran Razzak
Deakin University, Australia
E-mail: imran.razzak@deakin.edu.au

Xianzhi Wang
University of Technology Sydney (UTS), Australia
E-mail: Xianzhi.Wang@uts.edu.au

Shoaib Jameel
University of Essex, United Kingdom
E-mail: shoaib.jameel@essex.ac.uk

Guandong Xu
University of Technology Sydney (UTS), Australia
E-mail: Guandong.Xu@uts.edu.au

the mass shooting in the US [2], which has taken numerous innocent lives. One of the common mental health problems is depression that is more dominant than other mental illness conditions worldwide [3]. Diagnosis of depression is usually a difficult task because depression detection needs a thorough and detailed psychological testing by experienced psychiatrists at an early stage [4] and it requires interviews, questionnaires, self-reports or testimony from friends and relatives. Moreover, it is very common among people who suffer from depression that they do not visit clinics to ask help from doctors in the early stages of the problem [5].

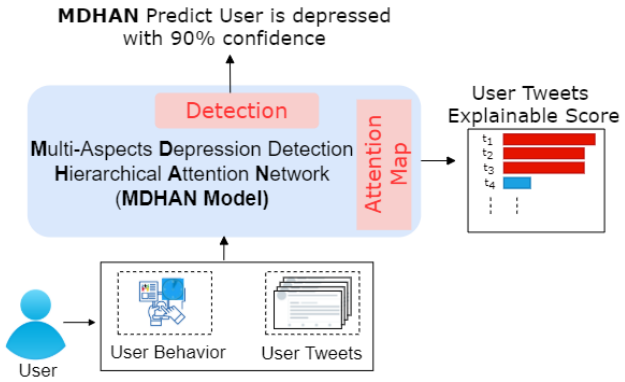


Fig. 1: Explainable depression detection

Individuals and health organizations have shifted away from their traditional interactions, and now meeting online by building online communities for sharing information, seeking and giving the advice to help scale their approach to some extent so that they could cover more affected populations in less time. Besides sharing their mood and actions, recent studies indicate that many people on social media tend to share or give advice on health-related information [6–9]. These sources provide the potential pathway to discover the mental health knowledge for tasks such as diagnosis, medications and claims. It is common for people who suffer from mental health problems too often "implicitly" (and sometimes even "explicitly") to disclose their feelings and their daily struggles with mental health issues on social media as a way of relief [10,11]. Therefore, social media is an excellent resource to automatically discover people who are depressed. While it would take a considerable amount of time to manually sift through individual social media posts and profiles to locate people going through depression, automatic scalable computational methods could provide timely and mass detection of depressed people which could help prevent many major fatalities in the future and help people who genuinely need it at the right moment. Usually, depressed users act differently when they are on social media, producing rich behavioural data, which is often used to extract various features. However, not all of them are related to depression.

Recently, deep learning has been successfully applied to several application problems, such as stock market predictions [12,13], traffic flow and traffic accident risk predictions [14–16], and mental illness detections [17]. Moreover deep learning has been applied for depression detection on social media and showed significantly better performance than traditional machine learning methods. Hamad et. al. [18] presented a computational framework for automatic detection of the depressed user that initially selects relevant content through a hybrid extractive and abstractive summarization strategy on the sequence of all user tweets leading to a more fine-grained and relevant content, which then is forwarded to deep learning framework comprising of unified learning machinery of the convolutional neural network coupled with attention-enhanced gated recurrent units leading to better empirical performance than existing strong baseline methods. Even though recent work showed the effectiveness of deep learning methods for depression detection, most of the existing machine learning methods provide no explainability for depression prediction, hence their predictions are obscure to humans which reduces the trust in the deep learning models. An explainable model provides insights into how a deep learning model can be improved and supports understanding. Thus, to engenders the appropriate user trust and provide the reason behind the decision, we aim to develop an explainable deep learning-based solution for depression detection by utilizing multi-aspect features from the diverse behaviour of the depressed user in social media. Apart from the latent features derived from lexical attributes, we notice that the dynamics

of tweets, i.e. tweet timeline provides a crucial hint reflecting depressed user emotion change over time. To this end, we propose a hybrid model, **M**ulti-aspect **D**epression Detection **H**ierarchical **A**ttention **N**etwork **MDHAN** to boost the classification of depressed users using multi-aspect features and word embedding features. Figure 1 illustrate the effectiveness of explainability in improving user trust. The model can derive new deterministic feature representations from training data and produce superior results for detecting depression-level of Twitter users, and derive explanations from a user posts content. Besides, we also studied the performance of our model when we used the two components of user posts and his multi-aspect features separately. We found that model performance deteriorated when we used only multi-aspect features. We further show when we combined the two attributes, our model led to better performance. Our model is based on explainable depression detection, which can learn explainable information from a user's tweets. The attention map in Figure 1 returns a user's tweets with explainable scores where the higher the score, the more likely tweet that is important and contributed to depression classification. To summarize, our study makes the following **key contributions**:

1. a novel explainable depression detection framework using deep learning of the textual, behavioural, temporal, and semantic aspect features from social media. To the best of our knowledge, this is the first work of using multi-aspect of topical, temporal and semantic features jointly with word embeddings in deep learning for depression detection.
2. introducing the prospective of viewing explainability of model for depression detection and building a pipeline aided with explainability based on hierarchical attention networks to explain the prediction of depression detection.
3. Extensive experiments are conducted on benchmark depression twitter dataset, which shows the superiority of our proposed method when compared to baseline methods.

The rest of our paper is organized as follows. **Section 2** reviews the related work to our paper, and in **Section 3** we formulate our problem, and present our explainable model for detection depression, and describes the different attributes that we extracted for our model. **Section 4** reports experiments and results. Finally, **Section 5** concludes this paper.

2 Related Work

In this section, we will discuss closely related literature and mention how they are different from our proposed method. In general, just like our work, most existing studies focus on user behaviour to detect whether a user suffers from depression or any mental illness. We will also discuss other relevant literature covering word embeddings and hybrid deep learning methods which have been proposed for detecting mental health from online social networks and other resources including public discussion forums.

Understanding depression on online social networks could be carried out using two complementary approaches which are widely discussed in the literature, and they are:

- Post-level behavioural analysis
- User-level behavioural analysis

2.1 Post-level behavioural analysis

Methods that use this kind of analysis mainly target the textual features of the user post that is extracted in the form of statistical knowledge such as those based on count-based methods [19]. These features describe the linguistic content of the post which are discussed in [20, 21]. For instance, in [20] the authors propose a classifier to understand the risk of depression. Concretely, the goal of the paper is to estimate that there is a risk of user depression from their social media posts. To this end, the authors collect data from social media for a year preceding the onset of depression from user profiles and distil behavioural attributes to be measured relating to social engagement, emotion, language and linguistic styles, ego network, and mentions of antidepressant medications. The authors collect their data using crowd-sourcing tasks, which is not a scalable strategy, on Amazon Mechanical Turk. In their study, the crowd workers were asked to undertake a standardized clinical depression survey, followed by various questions on their depression history and demographics. While the authors have conducted thorough quantitative and qualitative studies, they are disadvantageous in that it does not scale to a large set of users and does not consider

the notion of text-level semantics such as latent topics and semantic analysis using word embeddings. Our work is both scalable and considers various features which are jointly trained using a novel hybrid deep learning model using a multi-aspect features learning approach. It harnesses high-performance Graphics Processing Units (GPUs) and as a result, has the potential to scale to large sets of instances. In Hu et al., [21] the authors also consider various linguistic and behavioural features on data obtained from social media. Their underlying model relies on both classification and regression techniques for predicting depression while our method performs classification, but on a large scale using a varied set of crucial features relevant to this task.

To analyze whether the post contains positive or negative words and/or emotions, or the degree of adverbs [22] used cues from the text, for example, *I feel a little depressed* and *I feel so depressed*, where they capture the usage of the word *"depressed"* in the sentences that express two different feelings. The authors also analyzed the posts' interaction (i.e., on Twitter (retweet, liked, commented)). Some researchers studied post-level behaviours to predict mental problems by analysing tweets on Twitter to find out the depression-related language. In [23], the authors have developed a model to uncover meaningful and useful latent structure in a tweet. Similarly, in [24], the authors monitored different symptoms of depression that are mentioned in a user's tweet. In [25], they study users' behaviour on both Twitter and Weibo. To analyze users' posts, they have used linguistic features. They used a Chinese language psychological analysis system called TextMind in sentiment analysis. One of the interesting post-level behavioural studies was done by [24] on Twitter by finding depression relevant words, antidepressants, and depression symptoms. In [26] the authors used post-level behaviour for detecting anorexia; they analyze domain-related vocabulary such as anorexia, eating disorder, food, meals and exercises.

2.2 User-level behaviours

There are various features to model users in social media as it reflects overall behaviour over several posts. Different from post-level features extracted from a single post, user-level features extract from several tweets during different times [22]. It also extracts the user's social engagement presented on Twitter from many tweets, retweets and/or user interactions with others. Generally, posts' linguistic style could be considered to extract features [21, 27]. The authors in [24] extracted six depression-oriented feature groups for a comprehensive description of each user from the collected data set. The authors used the number of tweets and social interactions as social network features. For user profile features, they have used user shared personal information in a social network. Analysing user behaviour looks useful for detecting eating disorders. In Wang et al., [28] they extracted user engagement and activities features on social media. They have extracted linguistic features of the users for psychometric properties which resembles the settings described in [25, 26] where the authors have extracted 70 features from two different social networks (Twitter and Weibo). They extracted features from a user profile, posting time and user interaction features such as several followers and followee. Similarly, Wong et al. combined user-level and post-level semantics and cast their problem as multiple instances learning setups. The advantage that this method has is that it can learn from user-level labels to identify post-level labels [29].

Recently, Lin et al. [30] applied a CNN-based deep learning model to classify Twitter users based on depression using multi-modal features. The framework proposed by the authors has two parts. In the first part, the authors train their model in an offline mode where they exploit features from Bidirectional Encoder Representations from Transformers (BERT) and visual features from images using a CNN model. The two features are then combined, just as in our model, for joint feature learning. There is then an online depression detection phase that considers user tweets and images jointly where there is a feature fusion at a later stage. In another recently proposed work [31], the authors use visual and textual features to detect depressed users on Instagram posts than Twitter. Their model also uses multi-modalities in data, but keep themselves confined to Instagram only. While the model in [32] showed promising results, it still has a certain disadvantage. For instance, BERT vectors for masked tokens are computationally demanding to obtain even during the fine-tuning stage, unlike our model which does not have to train the word embeddings from scratch. Another limitation of their work is that they obtain sentence representations from BERT, for instance, BERT imposes a 512 token length limit where longer sequences are simply truncated resulting in some information loss, where our model has a much longer sequence length which we can tune easily because our model is computationally cheaper to train. We have proposed a hybrid model that considers a variety of features, unlike these works. While we have not specifically

used visual features in our work, using a diverse set of crucial relevant textual features is indeed reasonable than just visual features. Of course, our model has the flexibility to incorporate a variety of other features including visual features.

Multi-modal features from the text, audio, images have also been used in [33], where a new graph attention-based model embedded with multi-modal knowledge for depression detection. While they have used the temporal CNN model, their overall architecture has experimented on small-scale questionnaire data. For instance, their dataset contains 189 sessions of interactions ranging between 7-33min (with an average of 16 min). While they have not experimented with their method with short and noisy data from social media, it remains to be seen how their method scales to such large collections. Xezonaki et al., [34] propose an attention-based model for detecting depression from transcribed clinical interviews than from online social networks. Their main conclusion was that individuals diagnosed with depression use affective language to a greater extent than those who are not going through depression. In another recent work [35], the authors discuss depression among users during the COVID-19 pandemic using LSTM and fastText [36] embeddings. In [37], the authors also propose a multi-model RNN-based model for depression prediction but apply their model on online user forum datasets. Trotzek et al., [38] study the problem of early detection of depression from social media using deep learning where they leverage different word embeddings in an ensemble-based learning setup. The authors even train a new word embedding on their dataset to obtain task-specific embeddings. While the authors have used the CNN model to learn high-quality features, their method does not consider temporal dynamics coupled with latent topics, which we show to play a crucial role in overall quantitative performance. Farruque et al., [39] study the problem of creating word embeddings in cases where the data is scarce, for instance, depressive language detection from user tweets. The underlying motivation of their work is to simulate a retrofitting-based word embedding approach [40] where they begin with a pre-trained model and fine-tune the model on domain-specific data.

Opinions and emotions play an important role in detecting depression in social media product feedback, services, and other topics. The analysis of emotions in users' posts has continued to be one of the leading research directions. Prior researches [20, 41, 42] have investigated how emotions and affective states play a role in people's interactions with technology. Recent research in depression identification has shown that excessive self-focused language and negative emotions are key indicators for detecting depressed people [43, 44]. De Choudhury et al. [45] collected a Twitter dataset that included postings from people who had been diagnosed with depression. They studied the sentiment, emotion and linguistic of these tweets. They found interesting differences in the usage of words associated with negative emotions for the depressed user's tweets. Additionally, Twitter data analysis reveals that moms' emotional expression, social engagement and linguistic style of moms who experience postpartum depression alter before their baby is even born [20].

Recent studies have started to target depressed users online, extracting features representing user behaviours and classifying these features into different groups, such as the number of posts, posting time distribution, and several followers and followee. Peng et. al. extracted different features and classified them into three groups, user profile, user behaviour and user text and used multi-kernel SVM for classification [46]. The above-mentioned works have some limitations. They mainly focused on studying user behaviour than taking cues from user-generated content such as the text they share which make it extremely difficult to achieve high performance in classification. These models also cannot work well to detect depressed users at the user level, and as a result, they are prone to incorrect prediction. Our novel approach combines user behaviour with user history posts. Besides, our strategy to select salient content using automatic summarization helps our model only focus on the most important information. Although recent deep learning methods showed significant performance for depression detection, most of the existing models do not explain prediction since explainability and effectiveness could sometimes conflict. The explainable model can provide deep insight into how a deep learning model can be improved and supports understanding. Therefore, to provide some details and explain user tweets or reasons to make a decision functioning clear or easy to understand, we aim to develop an explainable deep learning-based approach for depression detection. Our proposed model utilized multi-aspect features from the diverse behaviour of the depressed user and his posts on social media.

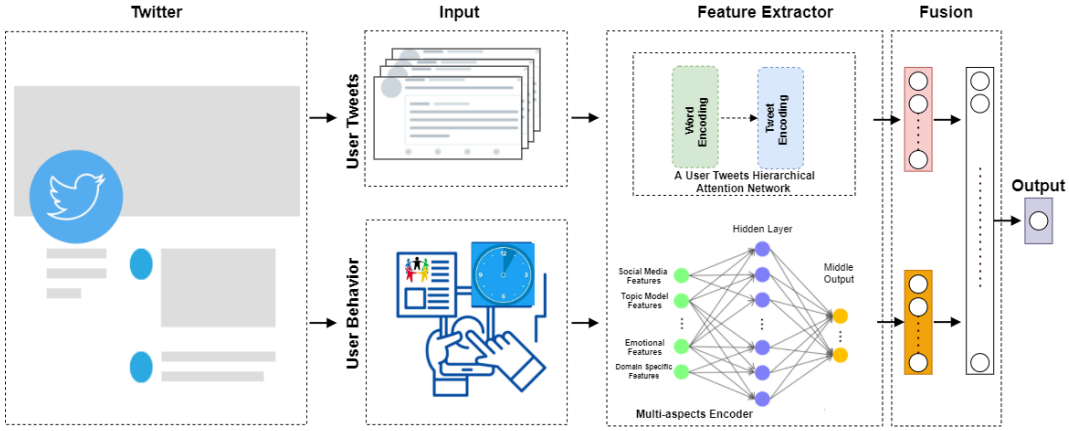


Fig. 2: Overview of our proposed model MDHAN: We predict depressed user by fusing two kinds of information: (1) User tweets. (2) User Behaviours.

2.3 Explainable Deep Learning

Deep neural networks help people make better decisions in various industries by producing more accurate and insightful predictions based on vast amounts of data. However, unlike interpretable machine learning methods [47,48], deep learning models (DNNs) learned representations are typically not interpretable by humans [49]. As a result, understanding the representations acquired by neurones in intermediate levels of DNNs is important to the explanation of deep neural networks (DNNs) [50,51]. Meanwhile, concerns about the nature and operation of the deep neural network's black box have grown, driving an increase in curiosity in deconstructing its essential components and understanding its functions. Therefore, explainability has lately received a lot of attention, owing to the requirement to explain the internal mechanics of a deep learning system [52, 53]. Many recent studies have focused on improving the transparency of deep neural networks to be adequately understood and be reliable. Attention-based methods can improve model transparency and have shown to be effective in various Natural Language Processing (NLP) tasks, including entity recognition, machine translation systems and text classification [54, 55]. Moreover, for document classification [52] and time series forecasting and classification [56], a variety of approaches for designing explainable neural networks employing attention processes have been investigated. In this paper, we propose using hierarchical attention to improve depression detection by capturing the explainability of depressed user tweets.

3 Explainable Deep Depression Detection

Suppose we have a set U of labelled users from both depression or non-depression samples. Let A be a user posts $A = [t_1, t_2, ..., t_L]$ consisting L tweets, where L is the total number of tweets per user, each tweet t_i contains n-words $t_i = [w_{i1}, w_{i2}, ..., w_{iN}]$ where N is the total number of words per tweet. Let M be the features in total for a user $\{m_i\}_{i=1}^{M}$, and let $\{1, 2, ..., S\}$ be a finite set of available aspects features, so we denote M_s as the dimension of S^{th} aspect. Therefore, once we have a user tweets A and a set of related user behaviours feature M. Our depression detection function is represented as follows:

$$f(A, M) \rightarrow \hat{y} \tag{1}$$

The model has been designed in such a way that it maximizes prediction accuracy. In our problem, we treat depression detection as the binary classification problem, i.e., user can be depressed ($\hat{y}$ = 1) or not-depressed ($\hat{y}$ = 0). Due to the complexity of user posts and the diversity of their behaviour on social media, we propose a hybrid model based on Hierarchical Attention Networks (HAN) that combines with Multilayer Perceptron (MLP) to detect depression through social media as depicted in Figure 2. For each user, the model takes two inputs for the two attributes. First, the four aspects feature input that represents the user behaviour vector runs into MLP, capturing distinct and latent features and correlation across the features matrix. The second input represents each user input tweet that will be replaced with its embedding and fed to Hierarchical Attention

Networks (HAN) to learn some representation features through a hierarchical word and tweet level encoding. The output in the middle of both attributes is concatenated to represent one single vector feature that fed into an activation layer of sigmoid for prediction. In the following sections, we will discuss the following two existing separate architectures.

3.1 Feature Selection

From the depression criteria and online behaviours on social media, we extracted a comprehensive set of depression-oriented features inspired by offline symptoms. Each feature group represents a single aspect. While we did not exploit multimedia features such as images or videos, we used a rich set of features to model multiple aspect. We introduce this attribute type where the goal is to calculate the attribute value corresponding to each features aspect for each user. We mainly consider four major aspects as listed below. These features are extracted respectively for each user as follows:

3.1.1 ***Social Information and Interaction***

From this attribute, we extracted several features embedded in each user profile. These are features related to each user account as specified by each feature name. Most of the features are directly available in the user data, such as the number of users following and friends, favourites, etc.

Moreover, the extracted features relate to user behaviour on their profile. For each user, we calculate their total number of tweets, the total length of all tweets and the number of retweets. We further calculate posting time distribution for each user, by counting how many tweets the user published during each of the 24 hours a day. Hence it is a 24-dimensional integer array. To get posting time distribution for each tweet, we extract two digits as hour information, then go through all tweets of each user and track the count of tweets posted in each hour of the day.

3.1.2 ***Emojis Sentiment***

Emojis allow users to express their emotions through simple icons and non-verbal elements. It is useful to get the attention of the reader. Emojis could give us a glance at the sentiment of any text or tweet, and it is essential to differentiate between positive and negative sentiment text [57]. User tweets contain a large number of emojis which can be classified into positive, negative and neutral. For each positive, neutral, and negative type, we count their frequency in each tweet. Then we sum up the numbers from each user's tweets to get the sum for each user. So the final output is three values corresponding to positive, neutral and negative emojis by the user. We also consider Voice Activity Detection (VAD) features. These features contain Valance, Arousal and Dominance scores. For that, we count First Person Singular and First Person Plural. Using affective norms for English words, a VAD score for 1030 words are obtained. We create a dictionary with each word as a key and a tuple of its (valance, arousal, dominance) score as value. Next, we parse each tweet and calculate the VAD score for each tweet using this dictionary. Finally, for each user, we add up the VAD scores of tweets by that user, to calculate the VAD score for each user.

3.1.3 ***Topic Distribution***

Topic modelling belongs to the class statistical modelling frameworks which help in the discovery of abstract topics in a collection of text documents. It gives us a way of organizing, understanding and summarizing collections of textual information. It helps find hidden topical patterns throughout the process, where the number of topics is specific by the user apriori. It can be defined as a method of finding a group of words (i.e. topics) from a collection of documents that best represent the latent topical information in the collection. In our work, we applied the unsupervised Latent Dirichlet Allocation (LDA) [58] to extract the most latent topic distribution from user tweets. To calculate topic level features, we first consider the corpus of all tweets of all depressed users. Next, we split each tweet into a list of words and assemble all words in decreasing order of their frequency of occurrence, and common English words (stopwords) are removed from the list. Finally, we apply LDA to extract the latent $K = 25$ topics distribution, where K is the number of topics. We have found experimentally $K = 25$ to be a suitable value. While there are tuning strategies and strategies based on Bayesian non-parametric [59], we have opted to use a simple, popular, and computationally efficient approach that helps give us the desired results.

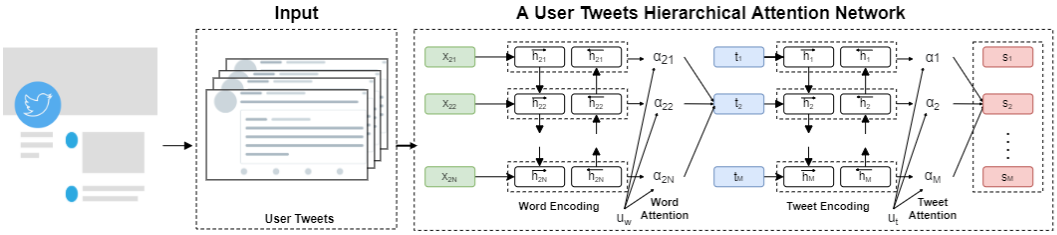


Fig. 3: An illustration of hierarchical attention network that we used to encode user tweets

3.1.4 Domain-specific features

1- Depression symptom counts: It is the count of depression symptoms occurring in tweets, as specified in nine groups in DSM-IV criteria for a depression diagnosis. The symptoms are listed in Appendix 7. We count how many times the nine depression symptoms are mentioned by the user in their tweets. The symptoms are specified as a list of nine categories, each containing various synonyms for the particular symptom. We created a set of seed keywords for all these nine categories, and with the help of the pre-trained word embedding, we extracted the similarities of these symptoms to extend the list of keywords for each depression symptom. Furthermore, we scan through all tweets, counting how many times a particular symptom is mentioned in each tweet. 2- Antidepressants: We also focused on the antidepressants, and we created a lexicon of antidepressants from the "Antidepressant" Wikipedia page which contains an exhaustive list of items and is updated regularly, in which we counted the number of names listed for antidepressants. The medicine names are listed in Appendix 8.

3.2 User Tweets Encoder Using RNN

Recently, researchers find that the HAN [52,60] can generate explanations by considering the most important words and sentences in a document. A depressed user could often have different linguistic style posts, including depressive language use, and mentions of antidepressants and symptoms, which can help detect depression. Additionally, a social media post contains linguistic prompts with different levels of word-level and tweet-level. Every word in a tweet and every tweet of a user is equally important to understand a depressed user in social media. For example, "My dad doesn't even seem to believe I'm really hurt!", the word "hurt" contributes more signals to decide whether the tweet is depressed rather than other words in the tweet. So in this way, HAN performs better in predicting the class of given user tweets. Inspired by [52], we proposed Hierarchical Attention Network to learn user tweets representation as depicted in Figure 3. We consider U be a user made M tweets $T = [t_1, t_2, ..., T_M]$ each tweet $t_i = [w_1, w_2, ..., w_N]$ contains N_i words. Each tweet is represented by the sequence of d-dimensional embeddings of their words, $t_i = [w_{11}, ..., w_{MN}]$. And we represent each word as the input layer a fixed-size vector from pre-trained word embeddings.

3.2.1 ***Word Encoder***

A bidirectional Gated Recurrent Unit (biGRU) is first used as the word level encoder to capture annotations' contextual information. GRU is a Recurrent Nural Network (RNN) that can capture sequential information and sentences' long-term dependency. Only two gate functions are used which are reset and update gates. Update gate has been used to monitor the degree to which the previous moment's status information has been transported into the current state. The higher the update gate value, the more the previous moment's status information is carried forward. The reset gate has been used to monitor the degree to which the previous moment's status information is overlooked. The smaller the reset gate value, the more neglected the context will be. Both the preceding and the following words influence the current word in the sequential textual data, so we use the BiGRU model to extract the contextual features. The BiGRU consists of a forward $\overrightarrow{GRU}$ and a backward $\overleftarrow{GRU}$ that are used, respectively, to process forward and backward data. The annotation w_{ij} represent the word j in a sentence i that contains N-words. Each word of user post (tweet) will convert to a word embedding x_{ij} utilising GloVe [61].

$$\overrightarrow{h_{ij}^{w}} = \overrightarrow{GRU}\left(x_{ij}, \overrightarrow{h_{i(j-1)}}\right), j \in \{1, ..., N\} \tag{2}$$

$$\overleftarrow{h_{ij}^{w}} = \overleftarrow{GRU}\left(x_{ij}, \overleftarrow{h_{i(j-1)}}\right), j \in \{N, ..., 1\} \tag{3}$$

The combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h_{ij}^{w}}$ and $\overleftarrow{h_{ij}^{w}}$ is represented as $h_{ij}^{w} = \left[\overrightarrow{h_{ij}^{w}} \oplus \overleftarrow{h_{ij}^{w}}\right]$. Which carries the complete tweet information centred around x_{ij}.

We describe the attention mechanism. It is crucial to introduce a vector u_{ij} for all words, which is trainable and expected to capture global words. The h_{ij}^{w} annotations create the basis for attention that starts with another hidden layer by letting the model learn and randomly initialized biases (b_w) and weights (W_w) through training. the annotations u_{ij} will be represented as follows:

$$u_{ij} = tanh(W_w h_{ij}^{w} + b_w) \tag{4}$$

The product $u_{ij}u_w$ (u_w is randomly initialized) expected to signal the importance of the j word and normalized to an importance weight per word α_{ij} by a softmax function:

$$\alpha_{ij} = \frac{\exp(u_{ij}u_w)}{\sum_{j} \exp(iju_w)} \tag{5}$$

Finally, a weighted sum of word representations concatenated with the annotations previously determined called the tweet vector v_i, where α_t indicating importance weight per word:

$$v_i = \sum\nolimits_{t} \alpha_{ij} h_{ij}^{w} \tag{6}$$

3.2.2 ***Tweet Encoder***

In order to learn the tweet representations h_i^t from a learned tweet vector v_i, we capture the information of context at the tweet level. Similar to the word encoder component, the tweet encoder employs the same BiGRU architecture. Hence the combination of the hidden state that is obtained from the forward GRU and the backward GRU $\overrightarrow{h_i^t}$ and $\overleftarrow{h_i^t}$ is represented as $h_i^t = \left[\overrightarrow{h_i^t} \oplus \overleftarrow{h_i^t}\right]$. Which capture the coherence of a tweet concerning its neighbouring tweets in both directions. Following that, we want to find user tweets that might explain why someone is sad. They should also help identify depression since they give good explainability. Since a user tweets may not be equally important in determining and explaining whether a user is depressed, we use attention over user tweets to capture the semantic affinity of tweets and learn their attention weights based on their relevance to the depression, allowing more reliable and explainable predictions. we will capture the related tweets in the formed vector $\hat{t}$ by using tweet level attention layer. The product $u_i u_s$ is expected to signal the importance of the i tweet and normalized to an importance weight per tweet α_i. Finally, s_i will be a vector that summarizes all the tweet information in a user post:

$$s_i = \sum\nolimits_{t} \alpha_i h_i^t \tag{7}$$

3.3 Multi-Aspect Encoder

Suppose the input which resembles a user behaviour be represented as $[m_1, m_2, ..., m_M\ ]$ where M is the total number of features and M_s is the dimension of S^{th} aspect. Hence, to obtain fine-grained information from user behaviours features, the multi-aspect features are fed through a one-layer MLP to get a hidden representation m_i:

$$p_i = f\left(b + \sum_{i=1}^{M} W_i m_i\right) \tag{8}$$

where f stands for the nonlinear function and the outcome of behaviour modelling p_i is the high-level representation that captures the behavioural semantic information and plays a critical role in depression diagnosis.

3.4 Classification Layer

At the classification layer, we need to predict whether the user is depressed or not depressed. So far, we have introduced, how we encode user multi aspect behaviours features (p) and how we can encode user tweets by modelling the hierarchical structure from word level and tweet level (s). Then from both components, we construct the feature matrix of user behaviours features and user tweets:

$$p = p_1, p_2, ..., p_M \in \mathbb{R}^{1d \times M} \tag{9}$$

$$s = s_1, s_2, ..., s_n \in \mathbb{R}^{2d \times n} \tag{10}$$

We further unify these components together, which is denoted as $[p, s]$. The output of such a network is typically fed to a sigmoid layer for classification:

$$\hat{y} = sigmoid(b_f + [p, s]W_f) \tag{11}$$

where where $\hat{y}$ is the predicted probability vector with $\hat{y_0}$ and $\hat{y_1}$ indicate the predicted probability of label being 0 (not depressed) and 1 (depressed user) respectively. Then, we aim to minimize the cross-entropy error for each user with ground-truth label y:

$$\text{Loss} = -\sum_i y_i \cdot \log \hat{y}_i$$

where $\hat{y}_i$ is the predicted probability and y_i is the ground truth label (either depression or non-depression) user.

3.5 Explainability

We aim to select user tweets that can explain why a user is depressed. As they provide a reasonable explanation, they should also help detect depression. The hierarchical attention that we explained in the previous sections in the word and tweet encoding is a suitable mechanism for giving high weights of user tweets representations. Besides, the explainability degree of user tweets are learned through the attention weight. Since varied words have different weights in each tweet based on the attention map, it indicates that our model can extract important and long-range contextual information from a tweet. Generally, the attention map of our model can select the most contributed words that identify a depressed and their corresponding tweets. Therefore, user tweets with high attention weight are essential and likely explain why a user is depressed.

4 Experiments and Results

In this section, we present the experimental evaluation to validate the performance of **MDHAN**. First will we will introduce datasets and evaluation Metrics and experimental settings, followed by the experimental results.

4.1 Comparative Methods

We compare our model with the following classification methods:

- **MDL: Multimodal Dictionary Learning Model** is to detect depressed users on Twitter [24]. They use dictionary learning to extract latent data features and sparse representations of a user.
- **SVM: Support Vector Machines** is a popular and strong classifier that has been applied on a wide range of classification tasks [62] and it remains a strong baseline.
- **NB: Naive Bayes** is a family of probabilistic algorithms based on applying Bayes' theorem with the "naive" assumption of conditional independence between instances [63].
- **BiGRU**:We applied **Bidirectional Gated Recurrent Unit** [64] with attention mechanism to obtain user tweets representations, which we then used for user tweets classification.

Description	Depressed	Non-Depressed
Numer of users	2159	2049
Numer of tweets	447856	1349447

Table 1: Summary of labelled data used to train MDHAN model

- **MBiGRU**: Hybrid model based on MLP and BiGRU for multi-aspect features for the user behaviour and the user's online timeline (posts).
- **CNN**: We utilized **Convolutional Neural Networks** [65] with an attention mechanism to model user tweets, which can capture the semantics of different convolutional window sizes for depression detection.
- **MCNN**: Hybrid model based on MLP and CNN for multi-aspect features for the user behaviour and the user's online timeline (posts).
- **HAN**: A hierarchical attention neural network framework [52], it used on user posts for depression detection. The network encodes first user posts with word-level attention on each tweet and tweet-level attention on each user post.
- **MDHAN**: The proposed model in this paper.

4.2 Datasets

Recent research conducted by Shen et al. [24] is one such work that has collected large-scale data with reliable ground truth data, which we aim to reuse. To exemplify the dataset further, the authors collected three complementary data sets, which are:

- Depression data set: Each user is labelled as depressed, based on their tweet content between 2009 and 2016.
- Non-depression data set: Each user is labelled as non-depressed and the tweets were collected in December 2016.
- Depression-candidate data set: The authors collected are labelled as depression-candidate, where the tweet was collected if contained the word "depress".

Data collection mechanisms are often loosely controlled, impossible data combinations, for instance, users labelled as depressed but have provided no posts, missing values, among others. After data has been crawled, it is still not ready to be used directly by the machine learning model due to various noise still present in data, which is called the "raw data". The problem is even more exacerbated when data has been downloaded from online social media such as Twitter because tweets may contain spelling and grammar mistakes, smileys, and other undesirable characters. Therefore, a pre-processing strategy is needed to ensure satisfactory data quality for computational modal to achieve reliable predictive analysis.

To further clean the data we used Natural Language processing ToolKit (NLTK). This package has been widely used for text pre-processing [66] and various other works. It has also been widely used for removing common words such as stop words from text [23, 67]. We have removed the common words from users tweets (such as "the", "an", etc.) as these are not discriminative or useful enough for our model. These common words sometimes also increase the dimensionality of the problem which could sometimes lead to the "curse-of-dimensionality" problem and may have an impact on the overall model efficiency. To further improve the text quality, we have also removed non-ASCII characters which have also been widely used in literature [27].

Pre-processing and removal of noisy content from the data helped get rid of plenty of noisy content from the dataset. We then obtained high-quality reliable data which we could use in this study. Besides, this distillation helped reduce the computational complexity of the model because we are only dealing with informative data which eventually would be used in modelling. We present the statistics of this distilled data below:

- Number of users labelled positive tweets: 5899.
- Number of tweets from positive users: 508786.
- Number of users labelled negative: 5160.
- Number of tweets from negative users: 2299106.

To further mitigate the issue of sparsity in data, we excluded those users who have posted less than ten posts and users who have less than 5000 followers, therefore we ended up with 2159 positive users and 2049 negative users.

Matric	SVM	NB	MDL	BiGRU	MBiGRU	CNN	MCNN	HAN	**MDHAN**
Accuracy	0.644	0.636	0.787	0.764	0.786	0.806	0.871	0.844	**0.895**
Precision	0.724	0.724	0.790	0.766	0.789	0.817	0.874	0.870	**0.902**
Recall	0.632	0.623	0.786	0.762	0.787	0.804	0.870	0.840	**0.892**
F1-score	0.602	0.588	0.786	0.763	0.786	0.803	0.870	0.839	**0.893**

Table 2: Performance comparison of MDHAN against the baselines for depression detection on [24] Dataset

For our experiments, we have used the datasets as mentioned in section (3). They provide a large scale of data, especially for labelled negative and candidate positive, and in our experiments, we used the labelled dataset. We preprocess the dataset by excluding users who have their posting history comprising of less than ten posts or users with followers more than 5000, or users who tweeted in other than English so that we have sufficient statistical information associated with every user. We have thus considered 4208 users (51.30% depressed and 48.69 % non-depressed users) as shown in Table 1. For evaluation purposes, we split the dataset randomly into training (80%) and test (20%), and we have reported our experimental results after performing five fold cross-validation.

4.3 Experimental Setting and Evaluation Metrics

For parameter configurations, the word embeddings are initialized with the Glove [61] with a dimension of 100 on the training set of each dataset to initialize the word embeddings of all the models, including baselines. The hidden dimension has been set to 100 in our model and other neural models, also, the dropout is set to 0.5. All the models are trained to use use the Adam optimization algorithm [68] with a batch size of 16 and an initial learning rate of 0.001. Then we trained our model for 10 iterations, with a batch size of 16. The number of iterations was sufficient to converge the model and our experimental results further cement this claim where we outperform existing strong baseline methods, and the training epoch is set to 10. We used python 3.6.3 and Tensorflow 2.1.0 to develop our implementation. We rendered the embedding layer to be not trainable so that we keep the features representations, e.g., word vectors and topic vectors in their original form. We used one hidden layer and a max-pooling layer of size 4 which gave a better performance in our setting. Finally, we employ traditional popular metrics such as precision, recall, F1, and accuracy.

4.4 Experimental Results

In our experiments, we study our model attributes including the quantitative performance of our hybrid model. For the multi-aspect features and user's timeline semantic features attribute, we will use both these attributes jointly. After grouped user behaviour in social media into a multi-aspect attribute, we evaluate the performance of the model. First, we examine the effectiveness of using the multi-aspect features only for depression detection with different classifiers. Second, we showed how the model performance increased when we utilize multi-aspect features with hierarchical attention network MDHAN. We summarise the results in Table 2 as follows:

- Naive Bayes obtain the lowest F1 score, which demonstrates that this model has less capability to classify tweets when compared with other existing models to detect depression. The reason for its poor performance could be that the model is not robust enough to sparse and noisy data.
- MDL model outperforms SVM, NB and BiGRU, and obtains better accuracy than these three methods. Since this is a recent model specially designed to discover depressed users, it has captured the intricacies of the dataset well and learned its parameters faithfully leading to better results.
- we can observe the evolving when we integrate The multi-aspect features with user posts and that better helped to analyze a user that seems to be depressed as shown in the performance of MBiGRU, MCNN MDHAN.
- We can see our proposed model MDHAN improved the depression detection up to 10% on F1-Score, compared to MDL model and 5% compared to HAN model. This suggests that our model outperforms a strong model. The reason why our model performs well is primarily that

it leverages a rich set of features which is jointly learned in the estimation of the consolidated parameters resulting in a robust model.
– Furthermore, MDHAN achieved the best performance with 89% in F1, indicating that combining HAN with multi-aspect strategy for user timeline semantic features strategy is sufficient to detect depression in Twitter. We can also deduce from the table that our model consistently outperforms all existing and strong baselines.

4.5 Comparison and Discussion

To get a better look at our model performance, We have compared the effectiveness of each of the two attributes of our model. Therefore, we test the performance of the model with a different attribute, we build the model to feed it with each attribute separately and compare how the model performs. First, we test the model using only the multi-aspect attribute, we can observe in Fig 4 the model perform less optimally when we used **M**LP for **M**ulti-aspect features (MM). In contrast, the model performs better when we use only HAN with word embedding attributes. This signifies that extracting semantic information features from user tweets is crucial for depression detection. Thus, we can see the MDHAN model performance increased when combined both MM and HAN, and outperforms when using each attribute independently. One of the key parameters in MDHAN is the number of tweets for each user; we eventually observed that MDHAN reached optimal performance when using 200 tweets as the maximum number of tweets. Figure 5 illustrates the performance of our model concerning the number of tweets.

To further analyze the role played by each aspect features and contribution of the user behavioural attributes and user posts attribute, we removed the four aspects separately as following: the domain-specific feature and denote as *MDHAN - D*, emotion feature and denote as *MDHAN - E*, the social network feature and denote this model as *MDHAN - S* and topic feature which we denote as *MDHAN - T*. We can see in Figure 6 that our model performance deteriorates as we remove the topic feature from the MDHAN model and degrades more without the social network features. To dive deeper and understand the effectiveness of each aspect, we combine each aspect separately with HAN and denote them respectively as following: *D+HAN*, *E+HAN*, *S+HAN* and *T+HAN*. As shown in Figure 7, we could see that MDHAN with four aspects outperforms the others, which means that each aspect does contribute to depression detection.

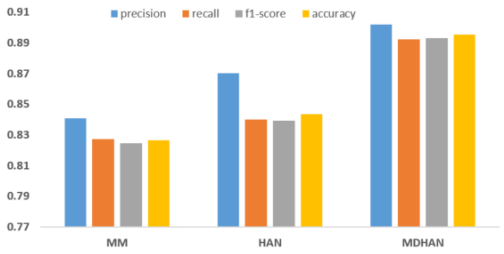


Fig. 4: Effectiveness comparison between MDHAN with different attributes.

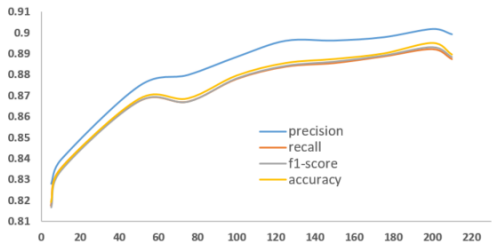


Fig. 5: Model vs number of tweets

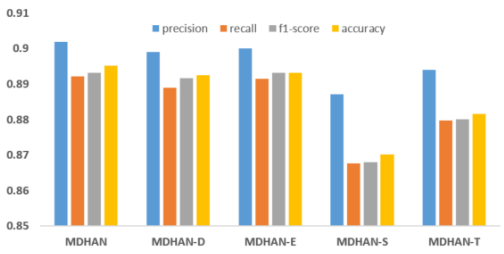


Fig. 6: Comparisons of various attributes

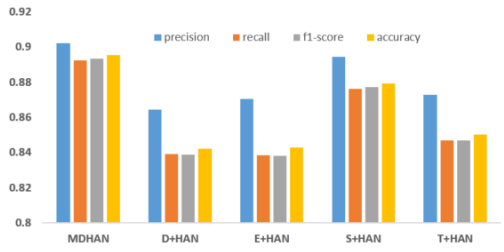


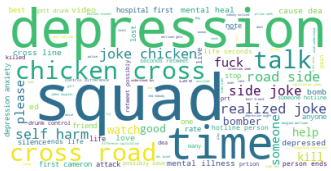
Fig. 7: Comparison of various use of attributes

Attention	Tweets
	One in four experience mental illness and yet hardly anyone I talk to seems to know anything about mental illness at all. Stigma abound.
	"MOST patients referred for talking therapies SHOULD start treatment within 6wks (max 18)." MOST? SHOULD? And howd we go from 6 to 18? mh
	"MANY patients who experience psychosis for the first time will get treatment within 2 weeks." MANY? WTF is "MANY"? mh
	Its ALWAYS time to talk about mental illness and it HAS been "time to talk" for many years now. The problem is that nobodys listening. mh
	6wks isnt enough anyway. If youre in physical pain, they give you treatment on the day of your GP appt. But for mental distress, wait 6wks
	Dont tell me what youll do in future if I vote for you. Show me what youve already done for us & then Ill consider voting for you.
	I cannot listen to this nonsense coming out of the mouths of people who have had the power and the opportunity to make changes ALREADY.
	Then they prance in acting like the saviour, announcing that theres a problem and theyre going to fix it - but not now. Next year. Maybe.

Fig. 8: Explainability via visualization of attention score in MDHAN

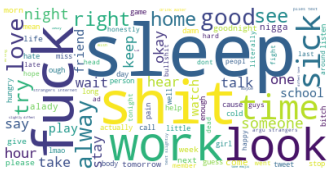
4.6 Case Study

To illustrate the importance of MDHAN for explaining depression detection results, we visualize the attention map for an example of a depressed user to show the words and tweets captured by MDAHN in Figure 8. The words and tweet weights are indicated by the red in this example, and the words and tweet are more important by attention weight if the colour is darker. Varied words have different weights in each tweet based on the attention map. It indicates that our model can extract important and long-range contextual information from a tweet. Generally, the attention map of our model can select the most contributed words that identify a depressed user, like mental, patients, therapies and illness, and their corresponding tweets. Tweets containing some words that have not contributed to classifying a depressed user and low attention weight will be neglected, for example, in the figure, we will notice that the first tweet has got the most attention, and the same goes for the words: mental and illness that had the highest weights when determining the prediction of class depression. The figure demonstrates that the attention map gives higher weights to explainable depression tweets; for instance, the tweet *"One in four experience mental illness ... "* gained the highest attention score among all the user tweets. Moreover, MDAHN can give higher weights to explainable tweets than those interfering and unrelated tweets, which can help select more related tweets and to be a more important feature to detect the depressed user.

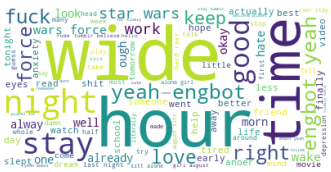


(a) A word cloud for Suicidal ideation.

(b) A word cloud for Worthlessness.



(c) A word cloud for Loss of Energy.



(d) A word cloud for Insomnia.

(e) A word cloud for depression mood.

Fig. 9: A word cloud depicting the most influencing symptoms.

To further investigate the five most influencing symptoms among depressed users, we collected all the tweets associated with these symptoms. Then we created a tag cloud [69] for each of these five symptoms, to determine what are the frequent words and importance that related to each symptom as shown in Figure 9 where larger font words are relatively more important than rest in the same cloud representation. This cloud gives us an overview of all the words that occur most frequently within each of these five symptoms.

5 Conclusion

We have proposed explainable Multi-Aspect Depression Detection with Hierarchical Attention Network (MDHAN) for detecting depressed users through social media analysis by extracting features from the user behaviour and the user's online timeline (posts). We have used a real-world data set for depressed and non-depressed users. Our main contribution is a novel hybrid computational model that can not only effectively model the real-world data but can also help derive explanations from them. We assign the multi-aspect attribute which represents the user behaviour into the MLP and user timeline posts into HAN to calculate each tweet and words' importance and capture semantic sequence features from the user timelines (posts). Our model shows that training this hybrid network improves classification performance and identifies depressed users outperforming other strong methods and ensures adequate evidence to explain the prediction. In the future, We will analyze users' tweets by considering topics and sentiments simultaneously to provide supporting evidence for each Depression DSM-IV criteria. Moreover, we will go beyond social media content and use URLs, images, and a mix of short and long user-generated content with traditional web pages. This would help give more contextual knowledge to the model that will help us focus on a task where our model not only detects depression but also automatically suggests the possible diagnosis.

6 Declaration

This work was supported by Australian Research Council (ARC) under Grant No. DP200101374 and LP170100891. The Authors declare that they has received research support from University of Technology Sydney. Shoaib Jameel is supported by Global Challenges Research Fund (grant number G004) and NVIDIA Academic Hardware Grant Program. The authors declare that they have no conflict of interest.

References

1. Kathleen Ries Merikangas, Jian-ping He, Marcy Burstein, Sonja A Swanson, Shelli Avenevoli, Lihong Cui, Corina Benjet, Katholiki Georgiades, and Joel Swendsen. Lifetime prevalence of mental disorders in us adolescents: results from the national comorbidity survey replication–adolescent supplement (ncs-a). *Journal of the American Academy of Child & Adolescent Psychiatry*, 49(10):980–989, 2010.
2. Jonathan M Metzl and Kenneth T MacLeish. Mental illness, mass shootings, and the politics of american firearms. *American journal of public health*, 105:240–249, 2015.
3. Aqsa Zafar and Sanjay Chitnis. Survey of depression detection using social networking sites via data mining. In *2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*, pages 88–93. IEEE, 2020.
4. Esteban Andrés Ríssola, Mohammad Aliannejadi, and Fabio Crestani. Beyond modelling: Understanding mental disorders in online social media. In *European Conference on Information Retrieval*, pages 296–310. Springer, 2020.
5. Maria Li Zou, Mandy Xiaoyang Li, and Vincent Cho. Depression and disclosure behavior via social media: A study of university students in china. *Heliyon*, 6(2):e03368, 2020.
6. Carleen Hawn. Take two aspirin and tweet me in the morning: how twitter, facebook, and other social media are reshaping health care. *Health affairs*, 28:361–368, 2009.
7. Linda Neuhauser and Gary L Kreps. Rethinking communication in the e-health era. *Journal of Health Psychology*, 8(1):7–23, 2003.
8. Daniel Scanfeld, Vanessa Scanfeld, and Elaine L Larson. Dissemination of health information through social networks: Twitter and antibiotics. *American journal of infection control*, 38(3):182–188, 2010.
9. Kyle W Prier, Matthew S Smith, Christophe Giraud-Carrier, and Carl L Hanson. Identifying health-related topics on twitter. In *International conference on social computing, behavioral-cultural modeling, and prediction*, pages 18–25. Springer, 2011.
10. Minsu Park, Chiyoung Cha, and Meeyoung Cha. Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, volume 2012, pages 1–8, 2012.
11. Krishna C Bathina, Marijn ten Thij, Lorenzo Lorenzo-Luaces, Lauren A Rutter, and Johan Bollen. Depressed individuals express more distorted thinking on social media. *arXiv preprint arXiv:2002.02800*, 2020.

12. Huihui Ni, Shuting Wang, and Peng Cheng. A hybrid approach for stock trend prediction based on tweets embedding and historical prices. *World Wide Web*, 24(3):849–868, 2021.
13. Nhi NY Vo, Xuezhong He, Shaowu Liu, and Guandong Xu. Deep learning for decision making and the optimization of socially responsible investments and portfolio. *Decision Support Systems*, 124:113097, 2019.
14. Aniekan Essien, Ilias Petrounias, Pedro Sampaio, and Sandra Sampaio. A deep-learning model for urban traffic flow prediction with traffic events mined from twitter. *World Wide Web*, 24(4):1345–1368, 2021.
15. Fucheng Wang, Jiajie Xu, Chengfei Liu, Rui Zhou, and Pengpeng Zhao. On prediction of traffic flows in smart cities: a multitask deep learning based approach. *World Wide Web*, 24(3):805–823, 2021.
16. Patara Trirat and Jae-Gil Lee. Df-tar: A deep fusion network for citywide traffic accident risk prediction with dangerous driving behavior. In *Proceedings of the Web Conference 2021*, pages 1146–1156, 2021.
17. Jina Kim, Jieon Lee, Eunil Park, and Jinyoung Han. A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, 10(1):1–6, 2020.
18. Hamad Zogan, Imran Razzak, Shoaib Jameel, and Guandong Xu. Depressionnet: Learning multi-modalities with user post summarization for depression detection on social media. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 133–142, 2021.
19. Rémi Lebret and Ronan Collobert. Rehabilitation of count-based models for word vector representations. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 417–429. Springer, 2015.
20. Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3267–3276, 2013.
21. Quan Hu, Ang Li, Fei Heng, Jianpeng Li, and Tingshao Zhu. Predicting depression of social media user on different observation windows. In *2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*, volume 1, pages 361–364. IEEE, 2015.
22. Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196. ACM, 2015.
23. Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. 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*, pages 99–107, 2015.
24. Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844, 2017.
25. Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat-Seng Chua, and Wendy Hall. Cross-domain depression detection via harvesting social media. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 1611–1617. International Joint Conferences on Artificial Intelligence Organization, 2018.
26. Diana Ramírez-Cifuentes, Marc Mayans, and Ana Freire. Early risk detection of anorexia on social media. In *International Conference on Internet Science*, pages 3–14. Springer, 2018.
27. Amir Hossein Yazdavar, Hussein S Al-Olimat, Monireh Ebrahimi, Goonmeet Bajaj, Tanvi Banerjee, Krishnaprasad Thirunarayan, Jyotishman Pathak, and Amit Sheth. Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, pages 1191–1198. ACM, 2017.
28. Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 91–100. ACM, 2017.
29. Akkapon Wongkoblap, Miguel A Vadillo, and Vasa Curcin. Modeling depression symptoms from social network data through multiple instance learning. *AMIA Summits on Translational Science Proceedings*, 2019:44, 2019.
30. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.
31. Chun Yueh Chiu, Hsien Yuan Lane, Jia Ling Koh, and Arbee LP Chen. Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, 56(1):25–47, 2021.
32. Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. Sensemood: Depression detection on social media. In *Proceedings of the 2020 International Conference on Multimedia Retrieval*, pages 407–411, 2020.
33. Wenbo Zheng, Lan Yan, Chao Gou, and Fei-Yue Wang. Graph attention model embedded with multi-modal knowledge for depression detection. In *2020 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2020.
34. Danai Xezonaki, Georgios Paraskevopoulos, Alexandros Potamianos, and Shrikanth Narayanan. Affective conditioning on hierarchical networks applied to depression detection from transcribed clinical interviews. *arXiv preprint arXiv:2006.08336*, 2020.
35. JT Wolohan. Estimating the effect of COVID-19 on mental health: Linguistic indicators of depression during a global pandemic. In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.
36. Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in pre-training distributed word representations. *arXiv preprint arXiv:1712.09405*, 2017.
37. Anu Shrestha, Edoardo Serra, and Francesca Spezzano. Multi-modal social and psycho-linguistic embedding via recurrent neural networks to identify depressed users in online forums. *NetMAHIB*, 9(1):22, 2020.
38. Marcel Trotzek, Sven Koitka, and Christoph M Friedrich. Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):588–601, 2018.
39. Nawshad Farruque, Osmar Zaiane, and Randy Goebel. Augmenting semantic representation of depressive language: From forums to microblogs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 359–375. Springer, 2019.

40. Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*, 2014.
41. Jian Zhao, Liang Gou, Fei Wang, and Michelle Zhou. Pearl: An interactive visual analytic tool for understanding personal emotion style derived from social media. In *2014 IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 203–212. IEEE, 2014.
42. Kiichi Tago, Kosuke Takagi, Seiji Kasuya, and Qun Jin. Analyzing influence of emotional tweets on user relationships using naive bayes and dependency parsing. *World Wide Web*, 22(3):1263–1278, 2019.
43. Mario Ezra Aragón, Adrian Pastor López-Monroy, Luis Carlos González-Gurrola, and Manuel Montes. Detecting depression in social media using fine-grained emotions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1481–1486, 2019.
44. Nikhita Vedula and Srinivasan Parthasarathy. Emotional and linguistic cues of depression from social media. In *Proceedings of the 2017 International Conference on Digital Health*, pages 127–136, 2017.
45. 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*, pages 47–56, 2013.
46. Zhichao Peng, Qinghua Hu, and Jianwu Dang. Multi-kernel svm based depression recognition using social media data. *International Journal of Machine Learning and Cybernetics*, 10(1):43–57, 2019.
47. F Doshi-Velez and B Kim. Towards a rigorous science of interpretable machine learning, corr abs/1702.08608. *arXiv preprint arXiv:1702.08608*, 2017.
48. Nhi NY Vo, Shaowu Liu, Xitong Li, and Guandong Xu. Leveraging unstructured call log data for customer churn prediction. *Knowledge-Based Systems*, 212:106586, 2021.
49. Mengnan Du, Ninghao Liu, and Xia Hu. Techniques for interpretable machine learning. *Communications of the ACM*, 63(1):68–77, 2019.
50. Ninghao Liu, Hongxia Yang, and Xia Hu. Adversarial detection with model interpretation. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1803–1811, 2018.
51. Ninghao Liu, Mengnan Du, and Xia Hu. Representation interpretation with spatial encoding and multimodal analytics. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 60–68, 2019.
52. Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 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*, pages 1480–1489, 2016.
53. Hongxu Chen, Yicong Li, Xiangguo Sun, Guandong Xu, and Hongzhi Yin. Temporal meta-path guided explainable recommendation. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 1056–1064, 2021.
54. Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 189–198, 2017.
55. Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.
56. Phongtharin Vinayavekhin, Subhajit Chaudhury, Asim Munawar, Don Joven Agravante, Giovanni De Magistris, Daiki Kimura, and Ryuki Tachibana. Focusing on what is relevant: Time-series learning and understanding using attention. In *2018 24th International Conference on Pattern Recognition (ICPR)*, pages 2624–2629. IEEE, 2018.
57. Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. Sentiment of emojis. *PLOS ONE*, 10:1–22, 12 2015.
58. David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.
59. Yee W Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in neural information processing systems*, pages 1385–1392, 2005.
60. Dawei Cong, Yanyan Zhao, Bing Qin, Yu Han, Murray Zhang, Alden Liu, and Nat Chen. Hierarchical attention based neural network for explainable recommendation. In *Proceedings of the 2019 on International Conference on Multimedia Retrieval*, pages 373–381, 2019.
61. Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
62. Christian Karmen, Robert C Hsiung, and Thomas Wetter. Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods. *Computer methods and programs in biomedicine*, 120(1):27–36, 2015.
63. Andrew Y Ng and Michael I Jordan. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pages 841–848, 2002.
64. Kyunghyun Cho, Bart van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar, October 2014. Association for Computational Linguistics.
65. Yoon Kim. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar, October 2014. Association for Computational Linguistics.
66. Krystian Horecki and Jacek Mazurkiewicz. Natural language processing methods used for automatic prediction mechanism of related phenomenon. *Lecture Notes in Artificial Intelligence (Subseries of Lecture Notes in Computer Science)*, 9120:13–24, 06 2015.
67. M. Deshpande and V. Rao. Depression detection using emotion artificial intelligence. In *2017 International Conference on Intelligent Sustainable Systems (ICISS)*, pages 858–862, Dec 2017.
68. Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

69. Fernanda B Viégas and Martin Wattenberg. Timelines tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008.

7 Appendix A

List of depression symptoms as per DSM-IV:

1. Depressed mood.
2. iminished interest.
3. Weight or appetite change
4. Insomnia, hypersomnia.
5. Psychomotor retardation, psychomotor impairment.
6. Fatigue or loss of energy
7. Feelings worthlessness, guilt.
8. Diminished ability to think, indecisiveness.
9. Suicidal tendency.

8 Appendix B

List of antidepressant medicine names

Citalopram	Celexa	Cipramil	Escitalopram	Lexapro	Cipralex
Fluoxetine	Prozac	Sarafem	Fluvoxamine	Luvox	Faverin
Paroxetine	Paxil	Seroxat	Sertraline	Zoloft	Lustral
Desvenlafaxine	Pristiq	Duloxetine	Cymbalta	Levomilnac.	Fetzima
Milnacipran	Ixel	Savella	Venlafaxine	Effexor	Vilazodone
Viibryd	Vortioxetine	Trintellix	Nefazodone	Dutonin	Nefadar
Serzone	Trazodone	Desyrel	Atomoxetine	Strattera	Reboxetine
Edronax	Teniloxazine	Lucelan	Metatone	Viloxazine	Vivalan
Bupropion	Wellbutrin	Amitriptyline	Elavil	Endep	Trifluoperazine
Amioxid	Ambivalon	Equilibrin	Clomipramine	Anafranil	Desipramine
Norpramin	Pertofrane	Dibenzepin	Noveril	Victoril	Dimetacrine
Istonil	Dosulepin	Prothiaden	Doxepin	Adapin	Sinequan
Imipramine	Tofranil	Lofepramine	Lomont	Gamanil	Melitracen
Dixeran	Melixeran	Trausabun	Nitroxazepine	Sintamil	Nortriptyline
Pamelor	Aventyl	Noxiptiline	Agedal	Elronon	Nogedal
Opipramol	Insidon	Pipofezine	Azafen	Azaphen	Protriptyline
Vivactil	Trimipramine	Surmontil	Amoxapine	Asendin	Maprotiline
Ludiomil	Mianserin	Tolvon	Mirtazapine	Remeron	Setiptiline
Tecipul	Mianserin	mirtazapine	setiptiline	Isocarboxazid	Marplan
Phenelzine	Nardil	Tranylcyp.	Parnate	Selegiline	Eldepryl
Zelapar	Emsam	Caroxazone	Surodil	Timostenil	Metralindole
Inkazan	Moclobemide	Aurorix	Manerix	Pirlindole	Pirazidol
Toloxatone	Humoryl	Eprobemide	Befol	Minaprine	Brantur
Cantor	Bifemelane	Alnert	Celeport	Agomelatine	Valdoxan
Esketamine	Spravato	Ketamine	Ketalar	Tandospirone	Sediel
Tianeptine	Stablon	Coaxil	Indeloxazine	Elen	Noin
Medifoxamine	Clédial	Gerdaxyl	Oxaflozane	Conflictan	Pivagabine
Tonerg	Ademetionine	Aurorix	SAMe	Heptral	Transmetil
Samyl	Hypericum per.	St. John's Wort	SJW	Jarsin	Kira
Movina	Oxitriptan	Kira	5-HTP	Cincofarm	Levothym
Triptum	Rubidium chl.	Rubinorm	Tryptophan	Tryptan	Optimax
Aminomine	Magnesium	Noveril	Solian	Aripiprazole	Abilify
Brexpiprazole	Rexulti	Lurasidone	Latuda	Olanzapine	Zyprexa
Quetiapine	Seroquel	Risperidone	Risperdal	Buspirone	Buspar
Lithium	Eskalith	Lithobid	Modafinil	Thyroxine	Triiodoth.
Minocycline	Amitriptyline	chlordiaz.	Limbitrol	Parmodalin	Aurorix
Perphenazine	Etafron	Flupentixol	melitracen	Deanxit	Surodil