SenseMood: Depression Detection on Social Media

Chenhao Lin*
Xi'an Jiaotong University, China
linchenhao@xjtu.edu.cn

Pengwei Hu*
IBM Research, China
hupwei@cn.ibm.com

Hui Su
Pattern Recognition Center, Wechat AI, Tencent Inc, China
aaronsu@tencent.com

Shaochun Li
Jing Mei
lishaoc@cn.ibm.com
meijing@cn.ibm.com
IBM Research, China

Jie Zhou
Pattern Recognition Center, Wechat AI, Tencent Inc, China

Henry Leung
University of Calgary, Canada
leungh@ucalgary.ca

ABSTRACT

More than 300 million people have been affected by depression all over the world. Due to the medical equipment and knowledge limitations, most of them are not diagnosed at the early stages. Recent work attempts to use social media to detect depression since the patterns of opinions and thoughts expression of the posted text and images, can reflect users' mental state to some extent. In this work, we design a system dubbed SenseMood to demonstrate that the users with depression can be efficiently detected and analyzed by using proposed system. A deep visual-textual multimodal learning approach has been proposed to reveal the psychological state of the users on social networks. The posted images and tweets data from users with/without depression on Twitter have been collected and used for depression detection. CNN-based classifier and Bert are applied to extract the deep features from the pictures and text posted by users respectively. Then visual and textual features are combined to reflect the emotional expression of users. Finally our system classifies the users with depression and normal users through a neural network and the analysis report is generated automatically.

CCS CONCEPTS

• **Applied computing** → **Life and medical sciences**; • **Information systems** → *Users and interactive retrieval.*

KEYWORDS

Multimodal Learning, Depression Detection, Deep Neural Network

1 INTRODUCTION

Depression, as described in the world health organization's comprehensive mental health action plan 2013-2020 [21], is one of the most common mental disorders. More than 300 million people worldwide suffer from chronic depression, making it the leading cause of acquired disability worldwide. Accurate diagnosis of patients with depression is the premise of treatment, however, the depression patients have to take the initiative to contact with mental health professionals to have the opportunity to get a diagnosis. In clinical diagnosis, psychologists usually refer to standard diagnostic guidelines for diseases, such as PHQ [10], and conduct face-to-face interviews. Although this is the most effective way to diagnose depression, more than 70% of people with early depression go untreated, worsening their condition because that most people lack medical knowledge and do not realize the risks of the disease, or they are ashamed of the disease and do not seek medical treatment.

Many studies have shown that depression is a preventable disease [9, 16, 17], and early identification and timely treatment can reduce the adverse effects of this disease [20]. However, tools and services for early detection and treatment of depression remain limited in passive situations. Improved computational methods make it possible to automatically detect depression-related indicators from interpersonal communication and provide better screening for early depression. Social media is currently one of the most significant public communication platform for humanity, and its emergence offers a promising opportunity for early detection and intervention of mental illness.

Social media occupies an important part of people's daily life. A lot of multimedia content, mainly short texts and photos are continuously shared on social media. Compared with traditional interpersonal communication, information posted on the Internet can be quickly transmitted by acquaintances and viewed by strangers. This mechanism avoids direct contact with humans and enhances the users' desire to express their feelings. Thus, researchers began to analyze users' social network characteristics [6–8], and online emotion detection has been proved to be effective on Twitter [1, 26]. Park et al. obtained a small number of tweets from Twitter users and conducted face-to-face interviews with them to explore the language used to describe depression [19]. These methods are effective, but expensive and time consuming, and there is no guarantee that they can be applied on a large scale with robust results.

*Both authors contributed equally to this research.

ACM Reference Format:
Chenhao Lin, Pengwei Hu, Hui Su, Shaochun Li, Jing Mei, Jie Zhou, and Henry Leung. 2020. SenseMood: Depression Detection on Social Media. In *Proceedings of the 2020 International Conference on Multimedia Retrieval (ICMR '20), June 8–11, 2020, Dublin, Ireland.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3372278.3391932

ICMR '20, June 8–11, 2020, Dublin, Ireland
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7087-5/20/06...$15.00
https://doi.org/10.1145/3372278.3391932

In [3], Choudhury et al. extracted several feature sets, such as engagement and emotional attributes, for large-scale detection of depressed users on Twitter. Yates et al. [25] applied CNN to realize text classification labeled with depression on tweets. However, this kind of approach only considered the clues contained in the text and does not analyze users from multiple perspectives. Many offline studies have proved that audio-visual features can be used to indicate depression [2, 11], and the enhancement of video signals in conversation can be used to detect depression objectively [5, 24]. Inspired by this work, the recent study used visual social media, to encode predictors of depression [21]. Shen et al. [22] also attempted to combine various demographic information of the user, to detect the user's depressive state. However, text and images posted by users are two of the most critical signals on social media, and until now have not been seen as uniform input signals to detect depression on social media. As a result, there is an urgent need for a systematic multimodal framework capable of detecting depression risk from users' textual and visual tweets. The limitations of the existing work are also the problems we aimed to address: 1) the most existing datasets did not fully collect/user the text and visual signals. 2) the most previous work did not solve the problem of multimodal feature representation of social media. 3) it is challenging to integrate the visual and textual representation of social media.

In this paper, we demonstrate a system for efficient detection and analysis of the users with depression by using social media data. In this system, we propose a deep visual-textual multimodal learning approach to learn robust features from normal users and users with depression. We first collect all the user-posted images, together with all the text tweets from a public dataset, as the input signals for depression mining. Then, a CNN is trained and a BERT is used to extract feature representation from images and text signals respectively. Finally, we use a neural network combine this two kinds of features and to obtain the multimodal fusion output. We also conduct a series of experiments on the collected twitter dataset to verify the effectiveness of our system and proposed approach. The experimental results indicate that our system has the potential to accurately and timely detect depression of the users on social media. It helps to detect users' risk of depression and remind them to take active preventive measures.

2 SYSTEM

2.1 System Overview

SenseMood is a system designed to efficiently detect and analyze the potential users with depression on social media. The technologies involved in our system, which take full use of users visual and textual data, have outperformed the existing approaches. As a result, through the social media, our system can help the users themselves and medical or psychological experts to detect users' risk of depression and remind them to take active preventive measures. Figure 1 illustrates the architecture of our system. It mainly includes three parts, offline training, online detection and online analysis. During the offline training, we applied the proposed deep visual-textual multimodal learning approach to extract visual and textual features from users posted images and tweets. More robust multimodal feature was learned and used to classify the users into users with depression and normal users through neural network.

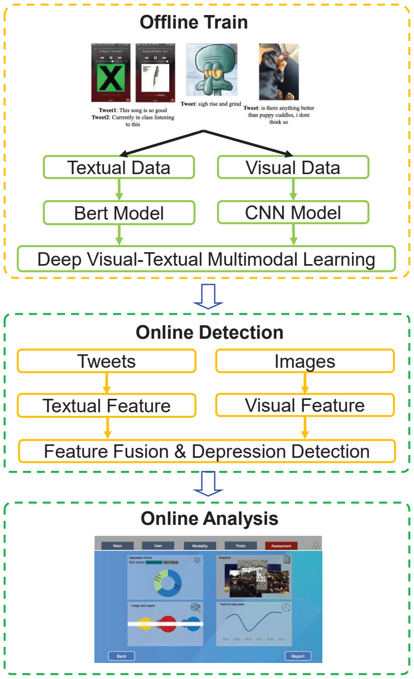

Offline Train
Tweet1: This song is so good
Tweet2: Currently in class listening to this
Tweet: sigh rise and grind
Tweet: is there anything better than puppy cuddles, i dont think so
Textual Data
Visual Data
Bert Model
CNN Model
Deep Visual-Textual Multimodal Learning
Online Detection
Tweets
Images
Textual Feature
Visual Feature
Feature Fusion & Depression Detection
Online Analysis
Main
User
Modality
Posts
Assessment
Back
Report

Figure 1: Illustration of proposed SenseMood system.

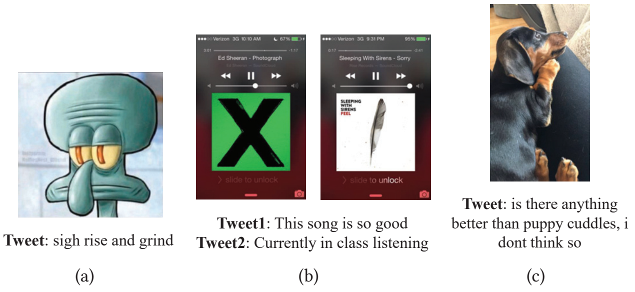
The details of offline training part will be introduced in the next section. For online detection part, the posted images and tweets from each input user were fed into the network for feature extraction, feature fusion and depression detection. Then our system outputted the detection results and analysis report.

2.2 Main Features

Our system provides several facilities to demonstrate the process and results of depression detection. We first provide a preview of sample users and their posted images and tweets. The sample users include the normal users and users with depression. The evidence of users with depression, i.e. 'I've got depression' tweet, has been removed.

Then the online detection function is used to process the selected user with their images and tweets based on trained model, to generate the prediction result. In addition to detecting depression, our system also provides the function to automatically generate depression analysis reports to visually and quantitatively illustrate the reason why such user is classified into depression or non-depression group. Such report helps the experts and users to understand the user's psychological expression.

The probability of the predicted results of the input user is illustrated in the analysis report. We also illustrate the probability curve based on each tweet and related images posted by the input user in chronological order. It indicates the risk and trend of the input user who may be suffer from depression. Such analysis can be used for prevention purpose. Our report also provides some key words/sentences which are extracted from the tweets of predicted users with depression. These key words/sentences imply that the users have high probability to suffer from depression. Our



Tweet: sigh rise and grind

(a)

Tweet1: This song is so good
Tweet2: Currently in class listening

(b)

Tweet: is there anything better than puppy cuddles, i dont think so

(c)

Figure 2: Examples of the tweets and images posted by the users with depression.

report also illustrates visual clues, i.e. sample images posted by the detected user with depression.

In Figure 2, we illustrate several examples of users with depression that correctly predicted by our system. Given space limitations, we only kept one or two images and tweets posted by each user. We find that these users can't be correctly detected by the models with only text features or image features. It indicates that the effective combination of image and text features in the proposed approach can help to detect users with depression. For example, an image with a negative expression combined with words related to depression can convey a strong tendency toward depression, as shown in Figure 2 (a). It is difficult to predict whether the users have depression by using only image features or text features as illustrated in Figure 2 (b) and (c), while they can be correctly predicted by using the proposed approach.

3 METHODOLOGY

In this section, we introduce the key technologies involved in the proposed system. The whole framework of the proposed approach is illustrated in Figure 3.

3.1 User Visual Feature

As a common data type on social media, the user-related images have been proved effective to represent the feelings and emotions of users [22, 23]. Unlike the previous work [22], in which the users' avatars were used to extracted visual features, we consider that both the users' profile images and the posted images can reveal the inner sensibility of users. Therefore, in this work, both of the two kinds of images are collected to generate the visual features.

In order to generate discriminative visual features from the users with and without depression, a binary CNN-based classifier is trained as the feature extractor. Two image data sets are collected to train the classifier and extract visual features separately. The profile images and posted images of the users that their tweets are used for text feature extraction, are collected as the image set V_{fea} for visual feature extraction. A set of profile images and posted images from other users with and without depression are built as V_{train} for classifier training.

For all users with depression, a set of their related images, V_{train_pos} can be acquired. While for other users, the set of images, V_{train_neg} is also collected. Then a deep convolutional neural network is trained. We fine-tune ImageNet pre-trained network using V_{train_pos} and V_{train_neg} data. Then the V_{fea} image set is fed into such binary classifier to generate user visual features. The feature vectors from the fully connected layer of trained CNN classifier are extracted as the deep visual feature representation. Each user-related image sample can be represented by a $1-d$ vectors of length 2048. Let $f(\cdot)$ be the function to generate the feature vector from the trained network. For the user which has multiple profiles or posted images, the mean value of each $f(V_{fea})$ is calculated as the user visual feature, i.e. $f(V_{fea})_m$.

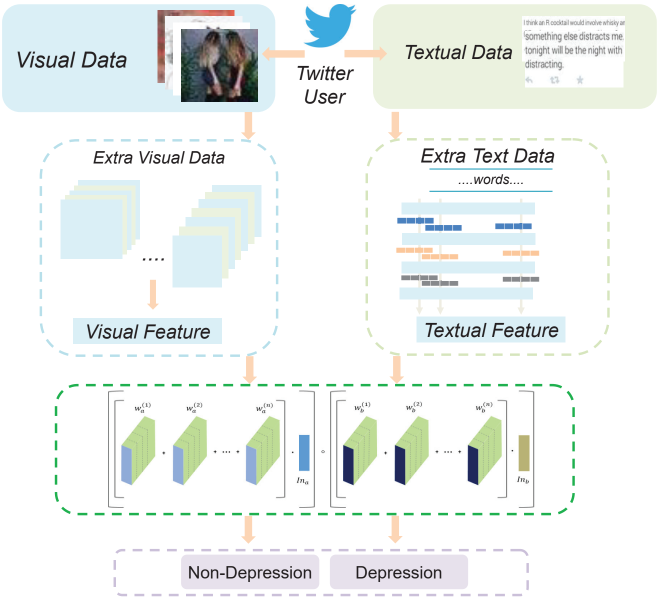


Figure 3: Illustration of proposed framework for depression detection.

3.2 User Textual Feature

Another common and more effective data type for depression detection on social media is text data [3, 22, 25]. In this work, we also attempt to acquire useful information from the user-posted text data for depression detection. Specifically, inspired by the success of BERT [4] in text understanding, we leverage it as the pre-trained model to help obtain contextualized sentence representations for users' Twitter text. A considerable disadvantage of the BERT network structure is that no independent sentence embeddings are computed, which makes it challenging to derive sentence embeddings from BERT. To bypass these limitations, we passed single sentences through BERT and then derived a fixed-sized vector by averaging the outputs from the second-to-last layer [14]. Furthermore, we lock BERT's internal weights, setting their gradients to zero. The text feature vector T_{fea} for each user is the sum of average sentence embeddings. In short, we use the pre-trained BERT to get a meaningful text feature vector T_{fea}. The BERT models are built upon the BERT base uncased model, which has 12 transformer layers, 12 self-attention heads, and with a hidden size 768. And twitter text is all lowercased. Non-ascii letters, urls, @[NAME] are removed. Texts with length more than 512 are thrown away.

3.3 Deep Visual-Textual Multimodal Learning

The users' visual features and text features are combined together to generate a more robust feature representation for more accurate depression detection. Unlike the common method in [12] that directly concatenate the feature vectors as the robust feature representation, reference [13] proposed a multimodal fusion method using low-rank tensor which has been proved efficient on several datasets. In this work, we also apply a similar approach to generate visual-text feature representation.

We define z_{vt} as the input feature tensor which is formulated by,

$$z_{vt} = f(\overset{\text{Vision}}{V_{fea}})_m \otimes f(\overset{\text{Text.}}{T_{fea}}) \tag{1}$$

where ⊗ donates the tensor outer product over input visual and text feature vector. The input tensor $z_{vt} \in \mathcal{R}^{d_1 \times d_2 \times \ldots \times d_M}$, where M is the number of modalities. f_{vt} is defined as the output fusion feature. For extracted visual feature $f(V_{fea})_m$ and $f(T_{fea})$, we can further train a multimodal fusion model using multiple basic linear functions (layers) $g(\cdot)$.

$$f_{vt} = g(z_{vt}; W, b) = W \cdot z_{vt} + b \tag{2}$$

where W is the weight of one linear function (layer) and b is the bias. The z_{vt} is an order-M tensor and the weight W will be a tensor of order-$(M+1)$ in $\mathcal{R}^{d_1 \times d_2 \times \ldots \times d_M \times d_{z_{vt}}}$. To achieve low-rank multimodal fusion, we follow [13] to decompose the weight W into different modality-specific factors. Let $\mathbf{w}$ be the set of rank $\mathcal{R}$ decomposition factors. A low-rank weight tensor can be recovered by,

$$W = \sum_{i=1}^{r} \otimes_{m=1}^{M} \mathbf{w}^{(i)} \tag{3}$$

Then equation (2) can be computed as follows,

self attention ?

$$f_{vt} = (\sum_{i=1}^{r} \otimes_{m=1}^{M} \mathbf{w}^{(i)}) \cdot z_{vt} \tag{4}$$

Finally, the linear function can be extended into a low-rank format,

$$\begin{aligned} f_{vt} &= (\sum_{i=1}^{r} \otimes \mathbf{w}^{(i)}) \cdot z_{vt} \\ &= (\sum_{i=1}^{r} \otimes \mathbf{w}_v^{(i)} \cdot f(V_{fea})_m) \circ (\sum_{i=1}^{r} \otimes \mathbf{w}_t^{(i)} \cdot f(T_{fea})) \end{aligned} \tag{5}$$

where ∘ denotes the element-wise product.

In such way, the visual-text feature representation can be generated. Unlike the work in [13], which further uses a LSTM layer to convert the text input tensor, we use the basic linear function since the text features have already been encoded and included rich contexts.

4 EVALUATION

4.1 Experimental Setting and Results

We used the Twitter dataset collected in reference [22] to evaluate the proposed approach. This dataset contains three subset, i.e. depression dataset D_1, non-depression dataset D_2 and depression-candidate dataset D_3 which contains potential users with depression. The depression dataset contains 1,402 depressed users and 292,564 tweets and the non-depression dataset contains more than 300 million non-depressed users and more than 10 billion tweets. D_3 includes 36,993 depression-candidate users and over 35 million tweets. We used the provided image URL address in these datasets to acquire profile images and posted images of each user.

Table 1: Performance of our and comparative approaches on the Twitter dataset.

Approach	Accuracy	Precision	Recall	F1-Score
Ref. [23]&ARC-t	-	-	-	0.737
Ref. [23]&MMDT	-	-	-	0.739
Ref. [23]&HFA	-	-	-	0.751
Ref. [23] DNN-FATC	0.781	0.740	0.840	0.785
Ref. [22] MDL	0.848	0.848	0.850	0.849
Ref. [18] Dual-Attention	0.848	0.848	0.848	0.848
Ref. [15] Modality Attention	0.866	0.868	0.862	0.864
Proposed Approach	**0.884**	**0.903**	**0.870**	**0.936**

To extract the users' visual features, we built a subset from D_3 and used it to train a binary classifier. The users whose tweets satisfied the strict pattern "(I'm/ I was/ I am/ I've been) diagnosed/suffered from depression" or (I've/ I have/ I had) depression were labeled as depression and the others were labeled as non-depression. Then 3362 images from users with depression and 4280 images from other users were collected respectively to train and test the classifier. All users with depression in D_1 were used for training and test and the same number of normal users were randomly selected as well. Test set consists of 280 users with depression and 280 normal users. We compared the proposed approach with previous state-of-the-art methods using precision, recall, F1-Measure score and accuracy.

We compared the proposed approach with several state-of-the-art methods for depression detection to validate the effectiveness of the proposed approach. Table 1 illustrates performance comparison in terms of accuracy, precision, recall and F1-score on the Twitter dataset. It can be seen that by using the proposed method, better performance has been achieved than several methods proposed in [23], [22], [18] and [15] for depression detection. Our approach has achieved 88.393% accuracy and 93.599% F1-score which outperforms ref. [15] method by 1.793% and 7.199%.

5 CONCLUSION

In this paper, we demonstrate a system dubbed SenseMood for depression detection using the users' social activities on Twitter. This system includes offline training, online detection and online analysis, which help the users and experts to review potential patient histories and support diagnosis. The proposed approach used in our system extracts the users' deep feature from users' textual information (tweet) and visual information (images). Multimodal feature learning approach is applied to fuse visual and textual feature and classify the users with depression and normal users. A public dataset is used to evaluate the efficiency of the proposed approach. Compared with state-of-the-art methods, our approach significantly improves the performance of depression detection on the social networks.

REFERENCES

[1] Oluwaseun Ajao, Deepayan Bhowmik, and Shahrzad Zargari. 2019. Sentiment aware fake news detection on online social networks. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2507–2511.

[2] Tuka Al Hanai, Mohammad M Ghassemi, and James R Glass. 2018. Detecting Depression with Audio/Text Sequence Modeling of Interviews.. In *Interspeech*. 1716–1720.

[3] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Seventh international AAAI conference on weblogs and social media*.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[5] Stephanie Gillespie, Elliot Moore, Jacqueline Laures-Gore, Matthew Farina, Scott Russell, and Yash-Yee Logan. 2017. Detecting stress and depression in adults with aphasia through speech analysis. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5140–5144.

[6] Pengwei Hu, Tiantian He, Keith CC Chan, and Henry Leung. 2017. Deep Fusion of Multiple Networks for Learning Latent Social Communities. In *2017 IEEE 29th International Conference on Tools with Artificial Intelligence (ICTAI)*. IEEE, 765–771.

[7] Pengwei Hu, Zhaomeng Niu, Tiantian He, and Keith CC Chan. 2018. Learning Deep Representations in Large Integrated Network for Graph Clustering. In *2018 IEEE First International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*. IEEE, 101–105.

[8] Po-Yao Huang, Junwei Liang, Jean-Baptiste Lamare, and Alexander G Hauptmann. 2018. Multimodal filtering of social media for temporal monitoring and event analysis. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*. 450–457.

[9] Felice N Jacka and Nicola J Reavley. 2014. Prevention of mental disorders: evidence, challenges and opportunities.

[10] Kurt Kroenke, Tara W Strine, Robert L Spitzer, Janet BW Williams, Joyce T Berry, and Ali H Mokdad. 2009. The PHQ-8 as a measure of current depression in the general population. *Journal of affective disorders* 114, 1-3 (2009), 163–173.

[11] Genevieve Lam, Huang Dongyan, and Weisi Lin. 2019. Context-aware Deep Learning for Multi-modal Depression Detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 3946–3950.

[12] Chenhao Lin and Ajay Kumar. 2018. A CNN-based framework for comparison of contactless to contact-based fingerprints. *IEEE Transactions on Information Forensics and Security* 14, 3 (2018), 662–676.

[13] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064* (2018).

[14] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. *arXiv preprint arXiv:1903.10561* (2019).

[15] Seungwhan Moon, Leonardo Neves, and Vitor Carvalho. 2018. Multimodal named entity disambiguation for noisy social media posts. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2000–2008.

[16] Ricardo F Muñoz, William R Beardslee, and Yan Leykin. 2012. Major depression can be prevented. *American Psychologist* 67, 4 (2012), 285.

[17] Ricardo F Munoz, Patricia J Mrazek, and Robert J Haggerty. 1996. Institute of Medicine report on prevention of mental disorders: summary and commentary. *American Psychologist* 51, 11 (1996), 1116.

[18] Hyeonseob Nam, Jung-Woo Ha, and Jeonghee Kim. 2017. Dual attention networks for multimodal reasoning and matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 299–307.

[19] Minsu Park, David W McDonald, and Meeyoung Cha. 2013. Perception differences between the depressed and non-depressed users in twitter. In *Seventh International AAAI Conference on Weblogs and Social Media*.

[20] A Picardi, I Lega, L Tarsitani, M Caredda, G Matteucci, MP Zerella, R Miglio, A Gigantesco, M Cerbo, A Gaddini, et al. 2016. A randomised controlled trial of the effectiveness of a program for early detection and treatment of depression in primary care. *Journal of affective disorders* 198 (2016), 96–101.

[21] Andrew G Reece and Christopher M Danforth. 2017. Instagram photos reveal predictive markers of depression. *EPJ Data Science* 6, 1 (2017), 15.

[22] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution.. In *IJCAI*. 3838–3844.

[23] Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat-Seng Chua, and Wendy Hall. 2018. Cross-Domain Depression Detection via Harvesting Social Media.. In *IJCAI*. 1611–1617.

[24] Melissa N Stolar, Margaret Lech, and Nicholas B Allen. 2015. Detection of depression in adolescents based on statistical modeling of emotional influences in parent-adolescent conversations. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 987–991.

[25] Andrew Yates, Arman Cohan, and Nazli Goharian. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848* (2017).

[26] Shumei Zhang, Jia Jia, and Yishuang Ning. 2017. Inferring emotions from heterogeneous social media data: A Cross-media Auto-Encoder solution. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2891–2895.