Multimodal Topic-Enriched Auxiliary Learning for Depression Detection

Minghui An, Jingjing Wang*, **Shoushan Li, Guodong Zhou**
School of Computer Science and Technology, Soochow University, China
mhan@stu.suda.edu.cn,
{djingwang, lishoushan, gdzhou}@suda.edu.cn

Abstract

From the perspective of health psychology, human beings with long-term and sustained negativity are highly probable to be diagnosed with depression. Inspired by this, we argue that the global topic information derived from user-generated contents (e.g., texts and images) is crucial to boost the performance of the depression detection task, though this information has been neglected by almost all previous studies on depression detection. To this end, we propose a new Multimodal Topic-enriched Auxiliary Learning (MTAL) approach, aiming at capturing the topic information inside different modalities (i.e., texts and images) for depression detection. Especially, in our approach, a modality-agnostic topic model is proposed to be capable of mining the topical clues from either the discrete textual signals or the continuous visual signals. On this basis, the topic modeling w.r.t. the two modalities are cast as two auxiliary tasks for improving the performance of the primary task (i.e., depression detection). Finally, the detailed evaluation demonstrates the great advantage of our MTAL approach to depression detection over the state-of-the-art baselines. This justifies the importance of the multimodal topic information to depression detection and the effectiveness of our approach in capturing such information.

1 Introduction

Depression detection is a task of determining a human being is *depressed* or *non-depressed* by automatically analyzing user-generated contents (UGC). Due to its crucial role in assessing mental health, it has recently received considerable attention from several research communities, such as NLP (Shen et al., 2017) and CV (Valstar et al., 2016). These studies mainly utilize UGC (e.g., texts and images) on social media to perform depression detection and achieve promising results since UGC instantly reflects not only the daily lives but also the mental states of users. Despite the progress of prior studies, they always focus on leveraging RNN variants (e.g., GRU) to model texts or images along the timeline (see Figure 1) posted by users, suffering from the problem of ignoring the global topic information inside these texts and images, though obviously this global topic information can mitigate the notorious difficulty of RNN variants w.r.t. modeling long-range dependencies (Dieng et al., 2017).

More importantly, from the point of health psychology, humans plagued by the negative emotion for a prolonged period of time (generally longer than two weeks), leading to their inability of carrying out daily activities, are highly probable to have the depression symptom (American Psychiatric Association and others, 2013). For instance, Figure 1 illustrates a month-long timeline of a *depressed* user. From this figure, we can see that this person is highly possible to be *depressed* since he/she discloses negative emotions lasting for almost a month. This conforms to his/her depression symptom and indicates the importance of considering the global semantics for depression detection.

Inspired by the above observations, this paper hypothesizes that it is desirable to consider the global topic information inside multiple modalities (i.e., texts and images) for depression detection. Still take Figure 1 as an example, leveraging a proper topic model to mine global clues inside the texts, e.g., words

[*] Corresponding Author: Jingjing Wang.

Proceedings of the 28th International Conference on Computational Linguistics, pages 1078–1089
Barcelona, Spain (Online), December 8-13, 2020

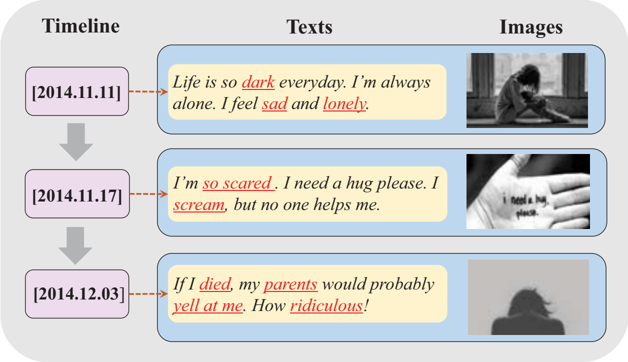

Timeline
Texts
Images
[2014.11.11]
Life is so dark everyday. I'm always alone. I feel sad and lonely.
[2014.11.17]
I'm so scared . I need a hug please. I scream, but no one helps me.
[2014.12.03]
If I died, my parents would probably yell at me. How ridiculous!

Figure 1: An example containing the timeline, texts and corresponding images, posted by a *depressed* user on Twitter.

"*so scared*" and "*parents ... yell at me*" indicating a topic w.r.t. *domestic violence*, can potentially assist the depression prediction. Besides, as reported in Reece and Danforth (2017), the images posted by *depressed* persons can be easily distinguished from those posted by healthy persons. Obviously, also leveraging a proper topic model to mine global clues inside images may more powerfully capture overall differences between *depressed* and *non-depressed* persons, thereby contributing to depression detection.

However, conventional latent topic models (Blei et al., 2003; Dieng et al., 2017) usually focus on processing the text modality composed of discrete textual signals (i.e., words) under the assumption that each topic is a multinomial distribution over the vocabulary. Apparently, these topic models cannot be directly adopted to mine the topic information inside the image modality since the image is composed of continuous visual signals, making the above assumption not applicable. Therefore, an appropriate topic model should not only be capable of capturing the textual topic information inside the texts but also be capable of capturing the visual topic information inside the images for depression detection.

To tackle the above challenges, we propose a multimodal topic-enriched auxiliary learning (MTAL) approach, which can mine both the textual topic information and the visual topic information for depression detection. In particular, a modality-agnostic topic model is first proposed to mine the topical clues from either discrete textual signals or continuous visual signals. Furthermore, the topic modeling w.r.t. the two modalities are cast as two auxiliary tasks for boosting the performance of the primary depression detection task. Third and finally, the primary task is trained alongside the two auxiliary tasks under the architecture of multi-task learning. Experimentation demonstrates that the proposed MTAL approach can significantly outperform several state-of-the-art baselines, including the representative textual depression detection approaches and the state-of-the-art multimodal-based approaches.

2 Related Work

Depression detection is an interdisciplinary research task and has been drawing ever-more attention in NLP with a focus on extracting various types of features from text modality (Choudhury et al., 2013; Nambisan et al., 2015). Compared with the studies on the text modality, the studies on multimodality (e.g., both the text and image modalities) like Gui et al. (2019b) are much less and limited to neglect the topic information inside multiple modalities. In the following, we will first review the depression detection task and then introduce the related studies on neural topic models.

Depression Detection. The ubiquity of social media poses a great opportunity to perform depression detection. Prior studies mainly focus on identifying *depressed* persons by analyzing the generated textual information in social media. Specifically, Choudhury et al. (2013) focus on the differences in word usage for depression detection. Gkotsis et al. (2016) focus on the depth of syntax-parsing trees for depression detection. In recent years, researchers begin to use multimodal information (e.g., the text, speech and image) for depression detection. Specifically, Yin et al. (2019) propose a hierarchical RNN network to extract the features from the vision, speech and text for depression detection. Gui et al. (2019b) propose a reinforced GRU network to capture both the textual and visual information for depression detection. In addition, it is worthwhile to mention that, Resnik et al. (2015) and Shen et al. (2017) also investigate

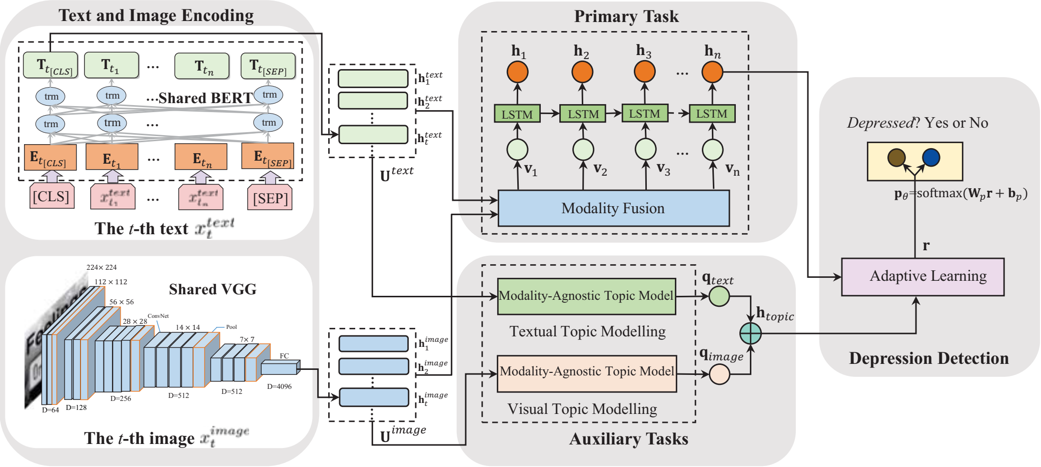


Figure 2: The overall architecture of the proposed Multimodal Topic-enriched Auxiliary Learning (MTAL) approach.

the topic information for depression detection, but is limited to capture this information inside single text modality. Different from them, this paper aims to integrate the topic information inside multiple modalities (i.e., both texts and images).

Neural Topic Models. Traditional topic models, e.g., probabilistic latent semantic analysis (pLSA) (Hofmann, 1999) and latent dirichlet allocation (LDA) (Blei et al., 2003), have been widely leveraged for inferring a low-dimensional latent representation that captures the global semantic information of a text. Recently, based on the variational auto-encoder (VAE) architecture (Kingma and Welling, 2014), Miao et al. (2017) and Srivastava and Sutton (2017) propose the neural topic models (NTM) to mine the topic information inside texts. Gui et al. (2019a) propose a reinforcement learning based neural topic model to alleviate the limitations of traditional topic coherence measures. Wang et al. (2020) also propose a topic-aware multi-task learning model to learn topic-enriched utterance representations in customer service, which is inspirational to our topic-enriched auxiliary learning framework. Unlike the above studies modeling topics under the assumption that the topic-word distribution is a multinomial distribution, Das et al. (2015) model topics with multivariate gaussian distribution over the word embedding space to deal with the new word issue, which is inspirational to our proposed modality-agnostic topic model. However, all the prior topic models rely on word vocabulary and thus are specially-designed for text modality, which cannot be directly adopted to capture the topic information inside images.

Different from all the above studies, this paper proposes a new modality-agnostic topic model to mine the global topics from either the discrete textual signals or the continuous visual signals. On this basis, a MTAL approach is proposed to integrate the multimodal topic information for depression detection. To our best knowledge, this is the first attempt to consider the topic information inside multiple modalities (i.e., both texts and images) for depression detection.

3 Multimodal Topic-Enriched Auxiliary Learning

Figure 2 shows the framework of our proposed Multimodal Topic-enriched Auxiliary Learning (MTAL) approach for depression detection, which consists of one primary task and two auxiliary tasks. The primary task is exactly the depression detection task (introduced in Section 3.1). Two auxiliary tasks are the textual and visual topic modeling respectively (together with the proposed modality-agnostic topic model are introduced in Section 3.2). Finally, a topic-enriched auxiliary learning strategy is proposed to combine the primary task with auxiliary tasks (introduced in Section 3.3).

3.1 Primary Task: Depression Detection

Given all pairs of text and image along the timeline (see Figure 1) posted by a user, the primary task aims at modeling both the text sequence and the corresponding image sequence to perform depression prediction for this user. Figure 2 shows the illustration of the primary task. First of all, given n pairs

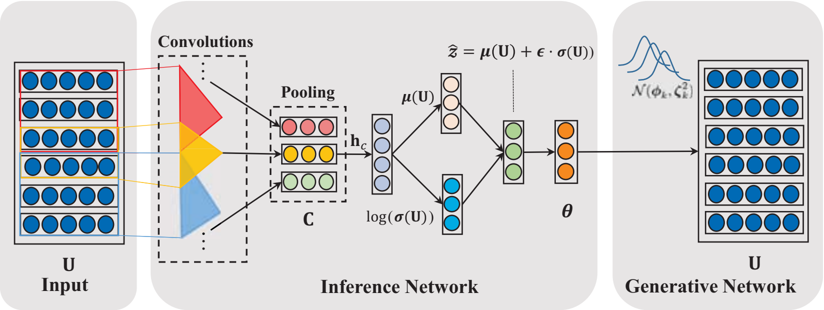

Convolutions
Pooling
C
$\mathbf{h}_c$
$\boldsymbol{\mu}(\mathbf{U})$
$\log(\boldsymbol{\sigma}(\mathbf{U}))$
$\hat{z} = \boldsymbol{\mu}(\mathbf{U}) + \boldsymbol{\epsilon} \cdot \boldsymbol{\sigma}(\mathbf{U}))$
$\boldsymbol{\theta}$
$\mathcal{N}(\phi_k, \zeta_k^2)$
U
Input
Inference Network
U
Generative Network

Figure 3: The framework of the proposed modality-agnostic topic model.

of text and image, each text and each image are encoded with a shared (i.e., parameter sharing) BERT (Devlin et al., 2019) model and a shared VGG (Simonyan and Zisserman, 2015) model respectively.

Text Encoder. As a pre-trained text encoding mechanism, BERT can be fine-tuned to create state-of-the-art models for a range of NLP tasks, e.g., text classification and natural language inference. In our approach, we use BERT-Base (uncased) model as the shared text encoder. Given the t-th text $x_t^{text} = \{x_{t_1}^{text}, x_{t_2}^{text}, ..., x_{t_n}^{text}\}$ of each user, we adopt BERT to encode this text and use the mark "[CLS]" representation $\hat{\mathbf{h}}_t \in \mathbb{R}^{768}$ to compute the text vector $\mathbf{h}_t^{text} \in \mathbb{R}^d$ of x_t^{text} as $\mathbf{h}_t^{text} = \tanh(\mathbf{W}_h\hat{\mathbf{h}}_t + \hat{\mathbf{b}}_h)$. Here, $\mathbf{W}_h \in \mathbb{R}^{d\times 768}$ and $\hat{\mathbf{b}}_h \in \mathbb{R}^d$ are trainable parameters. After all texts in a text sequence are encoded by this BERT, we can obtain a user-generated contents (UGC) matrix, i.e., text matrix $\mathbf{U}^{text} = \{\mathbf{h}_t^{text}\}_{t=1}^n$.

Image Encoder. As a pre-trained image encoding model, VGG has shown the state-of-the-art performance on various computer vision tasks, e.g., image caption and image classification (Simonyan and Zisserman, 2015). In this paper, we use VGG as the shared image encoder. Given the t-th image x_t^{image} of each user, following Gui et al. (2019b), we use the output vector $\hat{\mathbf{h}}_t \in \mathbb{R}^{4096}$ of the first fully connected layer in VGG to compute the image vector $\mathbf{h}_t^{image} \in \mathbb{R}^d$ of x_t^{image} as $\mathbf{h}_t^{image} = \tanh(\mathbf{W}_h\hat{\mathbf{h}}_t + \hat{\mathbf{b}}_h)$. Here, $\mathbf{W}_h \in \mathbb{R}^{d\times 4096}$ and $\hat{\mathbf{b}}_h \in \mathbb{R}^d$ are trainable parameters. After all images in a image sequence are encoded by this VGG, we can obtain another UGC matrix, i.e., image matrix $\mathbf{U}^{image} = \{\mathbf{h}_t^{image}\}_{t=1}^n$.

Modality Fusion. To incorporate both the text sequence and image sequence information, we concatenate the text vector and the image vector at the t-th time-step to obtain the new representation $\mathbf{v}_t \in \mathbb{R}^{2d}$ of the t-th text-image pair. Here, $\mathbf{v}_t = \mathbf{h}_t^{text} \oplus \mathbf{h}_t^{image}$. Finally, this representation $\mathbf{v}_t$ is fed to an LSTM network to obtain the final hidden state $\mathbf{h}_t \in \mathbb{R}^{2d}$ of the text-image pair as $\mathbf{h}_t = \mathrm{LSTM}(\mathbf{v}_t, \mathbf{h}_{t-1}, \mathbf{m}_{t-1})$, where $\mathbf{m}_{t-1}$ denotes the memory cell state at the time-step $t-1$. Finally, we regard vector $\mathbf{h}_n$ at the final time-step n as the output representation of the primary task.

3.2 Auxiliary Tasks: Multimodal Topic Modeling

In this section, we first introduce the proposed modality-agnostic topic model, and then present two types of auxiliary tasks, i.e., the textual topic modeling and the visual topic modeling.

Modality-Agnostic Topic Model. Unlike traditional neural topic models (Miao et al., 2017) focusing on generating an input text represented by a discrete bag-of-words vector, our modality-agnostic topic model aims to generate the intermediate UGC matrix of each modality (e.g., $\mathbf{U}^{text}$ or $\mathbf{U}^{image}$). For clarity, we will omit superscripts $text$ and $image$ of the UGC matrix and take one modality as an example next. Since both the text or image sequence are encoded into the same type of UGC matrix, the proposed topic model could be seen as the modality-agnostic. Similar to Miao et al. (2017), our topic model also adopts the variational auto-encoder architecture, aiming at generating the UGC matrix in an unsupervised setting. Figure 3 shows the workflow of our topic model, consisting of two main components, i.e., the inference network and the generative network.

• **Inference Network** is leveraged to infer the topic distribution $\boldsymbol{\theta}$ from the UGC matrix $\mathbf{U} = [\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n]$, where $\mathbf{u}_t$ is exactly the text vector $\mathbf{h}_t^{text}$ or the image vector $\mathbf{h}_t^{image}$. Specifically, we first aims at estimating $\boldsymbol{\mu}(\mathbf{U})$ and $\boldsymbol{\sigma}(\mathbf{U})$ for parameterizing a diagonal Gaussian distribution $q(\mathbf{z}|\mathbf{U}) = \mathcal{N}(\boldsymbol{\mu}(\mathbf{U}), \boldsymbol{\sigma}^2(\mathbf{U}))$. Wherein, $\mathbf{z} \in \mathbb{R}^K$ (where K is the number of topics) is a latent variable in the topic model. $\boldsymbol{\mu}(\mathbf{U})$ and $\boldsymbol{\sigma}(\mathbf{U})$ are functions of $\mathbf{U}$ which are implemented by neural networks.

More specifically, a convolutional layer is first employed to extract features from the UGC matrix. Formally, we suppose that the width of the kernel is j and the dimension of each row in the UGC matrix is d. A convolutional filter $\mathbf{W}_c \in \mathbb{R}^{d\times j}$ then maps j rows of matrix $\mathbf{U}$ in the receptive field to a single feature map $\mathbf{c}$. A sequence of new features $\mathbf{c} = [c_1, c_2, ..., c_n]$ are computed as $c_i = \tanh(\mathbf{u}_{i;i+j} * \mathbf{W}_c + b_c)$. Here, $b_c \in \mathbb{R}$ is the bias. $\tanh$ is a non-linear activation function. $*$ denotes convolution operation. If there are m_j filters of the same width j, the output features form a feature-map matrix $\mathbf{C} \in \mathbb{R}^{m_j\times n_j}$. We then apply a max-pooling operation over the matrix $\mathbf{C}$, resulting in a fixed-size vector $\mathbf{h}_c \in \mathbb{R}^{m_j}$. The output $\mathbf{h}_c$ can be fed into two different fully connected layers to estimate $\boldsymbol{\mu}(\mathbf{U})$ and $\log(\boldsymbol{\sigma}(\mathbf{U}))$:

$$\boldsymbol{\mu}(\mathbf{U}) = f_\mu(\mathbf{h}_c), \log(\boldsymbol{\sigma}(\mathbf{U})) = f_\sigma(\mathbf{h}_c) \tag{1}$$

where f_μ and f_σ are two different MLP fully connected layers. After obtaining $\boldsymbol{\mu}(\mathbf{U})$ and $\log(\boldsymbol{\sigma}(\mathbf{U}))$, the diagonal Gaussian distribution $q(\mathbf{z}|\mathbf{U})$ can be parameterized. We then sample $\hat{\mathbf{z}}$ from $q(\mathbf{z}|\mathbf{U})$ using a reparameterization trick as described in Kingma and Welling (2014), i.e., $\hat{\mathbf{z}} = \boldsymbol{\mu}(\mathbf{U}) + \boldsymbol{\epsilon}\boldsymbol{\sigma}(\mathbf{U})$. Here, $\boldsymbol{\epsilon}$ is sampled from $\mathcal{N}(\mathbf{0}, \mathbf{I}^2)$. Finally, inspired by the Gaussian softmax proposed by Miao et al. (2017), we compute the topic distribution $\boldsymbol{\theta} \in \mathbb{R}^K$ as follows:

$$\boldsymbol{\theta} = \mathrm{softmax}(\mathbf{W}_\theta \hat{\mathbf{z}} + \mathbf{b}_\theta) \tag{2}$$

where $\mathbf{W}_\theta \in \mathbb{R}^{K\times K}$ and $\mathbf{b}_\theta \in \mathbb{R}^K$ are trainable parameters.

• **Generative Network** is leveraged to parameterize $p(\mathbf{U}|\boldsymbol{\phi}_{1:K}, \boldsymbol{\zeta}_{1:K})$, which is a conditional probability distribution of the UGC matrix $\mathbf{U}$ given the trainable parameters ($\boldsymbol{\phi}_k$ and $\boldsymbol{\zeta}_k$) for the k-th topic. Different from the neural topic model proposed by Miao et al. (2017) which defines a multinomial distribution for each topic over the word vocabulary, our model defines an independent diagonal Gaussian distribution $\mathcal{N}(\boldsymbol{\phi}_k, \boldsymbol{\zeta}_k^2)$ for each topic k over the embedding space of different modalities. In this way, given a UGC matrix $\mathbf{U}$ with topic distribution $\boldsymbol{\theta}$, each embedding $\mathbf{u}_t$ (i.e., $\mathbf{h}_t^{text}$ or $\mathbf{h}_t^{image}$) of our topic model is generated in two steps:

- Choose a topic $\gamma_t \sim$ Multinomial Distribution$(\boldsymbol{\theta})$
- Choose the embedding $\mathbf{u}_t \sim \mathcal{N}(\boldsymbol{\phi}_{\gamma_t}, \boldsymbol{\zeta}_{\gamma_t}^2)$

Then, the probability distribution $p(\mathbf{U}|\boldsymbol{\phi}_{1:K}, \boldsymbol{\zeta}_{1:K})$ for the UGC matrix $\mathbf{U}$ can be computed as follows:

$$p(\mathbf{U}|\boldsymbol{\phi}_{1:K}, \boldsymbol{\zeta}_{1:K}) = \int p(\boldsymbol{\theta}) \prod_{t=1}^{n} \sum_{\gamma_t} p(\mathbf{u}_t|\boldsymbol{\phi}_{\gamma_t}, \zeta_{\gamma_t}) p(\gamma_t|\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} \tag{3}$$

Finally, the loss function for the proposed modality-agnostic topic model is computed as follows:

$$\mathcal{L} = \mathrm{KL}[q(\mathbf{z}|\mathbf{U})||p(\mathbf{z})] - \mathbb{E}_{q(\mathbf{z}|\mathbf{U})}\left[\log p(\mathbf{U}|\boldsymbol{\phi}_{1:K}, \boldsymbol{\zeta}_{1:K})\right] \tag{4}$$

where $p(\mathbf{z})$ is a standard Normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$. In the first part of Eq.(4), we use KL divergence to measure the similarity between the learned distribution $q(\mathbf{z}|\mathbf{U}))$ and true prior distribution $p(\mathbf{z})$. The second part of Eq.(4) represents the likelihood of reconstructing original input via the generative network.

On the basis of the proposed modality-agnostic topic model, we further construct two auxiliary tasks 1) textual topic modeling and 2) visual topic modeling, aiming at capturing the topic information inside texts and images respectively. Concretely, we first take advantage of the mean vector $\boldsymbol{\phi}_k \in \mathbb{R}^M$ of each Gaussian distribution to construct the topic embedding matrix $\boldsymbol{\Phi} = [\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, ..., \boldsymbol{\phi}_K]$ of all topics, since each topic has an independent Gaussian distribution as mentioned above. Then, we leverage the topic embedding matrix to compute the topic representation of each modality (texts or images) according to their corresponding topic distribution $\boldsymbol{\theta}$. More concretely, two auxiliary tasks are formulated as follows.

Auxiliary Task 1: Textual Topic Modeling. The text sequence posted by a user is first encoded into a text matrix $\mathbf{U}^{text}$, which is then fed into a modality-agnostic topic model to obtain the textual topic

	Depressed			***Non-Depressed***		
	# User	**# Text**	**# Text + Image**	**# User**	**# Text**	**# Text + Image**
Train	981	172,672	15,934	981	606,279	45,909
Dev.	140	19,372	1,867	140	95,487	7,806
Test	281	40,851	4,394	281	177,259	10,644
All	1,402	232,895	22,195	1,402	879,025	64,359

Table 1: Statistics of the dataset adopted in our experiments. **# User** denotes the number of users. **# Text** denotes the number of tweets only containing the text. **# Text + Image** denotes the number of tweets containing both text + image pairs.

distribution $\boldsymbol{\theta}_{text} \in \mathbb{R}^K$. Finally, the textual topic representation $\mathbf{q}_{text} \in \mathbb{R}^M$ of all texts is computed as $\mathbf{q}_{text} = \boldsymbol{\Phi}_{text}^{\top} \boldsymbol{\theta}_{text}$, where $\boldsymbol{\Phi}_{text} \in \mathbb{R}^{K \times M}$ denotes the textual topic embedding matrix.

Auxiliary Task 2: Visual Topic Modeling. The image sequence posted by the user is first encoded into a image matrix $\mathbf{U}^{image}$, then fed into another modality-agnostic topic model to obtain the visual topic distribution $\boldsymbol{\theta}_{image} \in \mathbb{R}^K$. Finally, the visual topic representation $\mathbf{q}_{image} \in \mathbb{R}^M$ of all images is computed as $\mathbf{q}_{image} = \boldsymbol{\Phi}_{image}^{\top} \boldsymbol{\theta}_{image}$, where $\boldsymbol{\Phi}_{image} \in \mathbb{R}^{K \times M}$ denotes visual topic embedding matrix.

After obtaining both the textual and visual topic representation, we compute the output representation $\mathbf{h}_{topic} \in \mathbb{R}^{2M}$ of all auxiliary tasks as $\mathbf{h}_{topic} = \mathbf{q}_{text} \oplus \mathbf{q}_{image}$.

3.3 Topic-Enriched Auxiliary Learning

Different from multi-task learning whose goal is to achieve better performance across all tasks, auxiliary learning differs in that better performance is only required for a single primary task, and the role of auxiliary tasks is to assist the performance improvement of this primary task. To this end, we take advantage of two strategies (i.e., adaptive learning and auxiliary training) to combine the primary task with the auxiliary tasks, which are illustrated as follows.

Adaptive Learning. To distinguish the output representation of primary task from that of auxiliary tasks for topic modeling, we utilize an adaptive gate $\mathbf{e} \in \mathbb{R}^{2d}$ to combine representations from both the primary and auxiliary tasks. The final user representation $\mathbf{r} \in \mathbb{R}^{2d}$ is computed as follows:

$$\mathbf{e} = \mathrm{sigmoid}(\mathbf{W}_e(\mathbf{h}_n \oplus \mathbf{h}_{topic}) + \mathbf{b}_e) \tag{5}$$

$$\mathbf{r} = (1 - \mathbf{e}) \odot \mathbf{h}_n + \mathbf{e} \odot (\mathbf{W}_r \mathbf{h}_{topic}) \tag{6}$$

where $\mathbf{h}_n \in \mathbb{R}^{2d}$ is the primary task representation. $\odot$ is the element-wise multiplication. $\mathbf{W}_e \in \mathbb{R}^{2d \times (2d+2M)}$, $\mathbf{b}_e \in \mathbb{R}^{2d}$ and $\mathbf{W}_r \in \mathbb{R}^{2d \times 2M}$ are trainable parameters. Further, vector $\mathbf{r}$ is fed to a softmax layer for depression prediction, i.e., $\mathbf{p}_\Theta = \mathrm{softmax}(\mathbf{W}_p \mathbf{r} + \mathbf{b}_p)$. Here, $\mathbf{W}_p \in \mathbb{R}^{n \times 2d}$, $\mathbf{b}_p \in \mathbb{R}^n$ are trainable parameters. n is category number. $\mathbf{p}_\Theta$ is the probability distribution over two categories.

Auxiliary Training. We employ the joint loss function to optimize all the primary and auxiliary tasks simultaneously. Here, the joint loss consists of two parts. One is the supervised loss for depression detection, and the other contains the unsupervised losses for the two auxiliary tasks of modality-agnostic topic modeling. Specifically, the loss $\mathcal{L}_{primary}$ for the primary depression detection task is computed as:

$$\mathcal{L}_{primary} = -\frac{1}{N} \sum_{i=1}^{N} \log \mathbf{p}_\Theta(y_i | x_i) + \frac{\delta}{2} ||\Theta||_2^2 \tag{7}$$

where N is the number of all twitter users. y_i is the ground-true label for the i-th user x_i. δ is an L_2 regularization weight. Θ denotes all training parameters in the model.

In addition, the loss function for our topic model has been shown in Eq.(4). For clarity, the losses for two auxiliary tasks, i.e., the textual topic modeling and visual topic modeling, are denoted as $\mathcal{L}_{text}$ and $\mathcal{L}_{image}$ respectively. Finally, the joint loss $\mathcal{L}$ is defined as follows:

$$\mathcal{L} = \mathcal{L}_{primary} + \lambda(\mathcal{L}_{text} + \mathcal{L}_{image}) \tag{8}$$

where λ is a weight and fine-tuned to be 0.25 for balancing the losses for primary and auxiliary tasks.

Approach	**Modality**	**Precision (P)**	**Recall (R)**	**F1**	**Acc.**
H-LSTM (Wang et al., 2018)	Text	77.2	77.1	77.1	77.1
BERT + LSTM		79.2	79.2	79.2	79.2
BERT + LSTM + Textual Topic Modeling		81.1	81.2	81.1	82.3
VGG + LSTM	Image	62.3	61.7	62.0	61.7
VGG + LSTM + Visual Topic Modeling		66.8	66.7	66.7	66.7
EF-LSTM (Zadeh et al., 2018)	Text + Image	79.9	79.9	79.9	79.9
CoMemory (Xu et al., 2018)		80.6	80.4	80.5	80.4
CoATT (Zhang et al., 2018)		79.6	80.3	79.9	80.4
Hybrid Attention (Gu et al., 2018)		80.6	80.6	80.6	80.6
CoMMA (Gui et al., 2019b)		78.3	79.4	78.8	79.2
Primary Task		81.4	80.9	81.1	80.9
MTAL		**84.2**	**84.2**	**84.2**	**84.2**

Table 2: Performance comparison of various kinds of approaches with the single-modality (text or image) and the multimodality (text and image) for depression detection.

4 Experimentation

To validate the effectiveness of our approach, we evaluate the performance of the proposed and baseline approaches and show the details in Table 2.

4.1 Experimental Settings

Data Setting. We conduct experiments based on the multimodal depression dataset[1] released by Gui et al. (2019b). Different from Gui et al. (2019b), we adopt a new data setting and split the original dataset into the standard train/development/test sets with the ratio of 7:1:2. The reason why we adopt this different setting is that Gui et al. (2019b) follows the same experimental setting proposed by Shen et al. (2017) for a fair comparison, while this old setting contains no real development set. Instead, Gui et al. (2019b) regard the test set as the development set for training and use the five-fold cross validation results on this development set as final results. We believe this is not well-suited for the training of the iterative neural network based approach because involving the label information of the test set in the training phase may not convincingly evaluate the generalization ability of a iterative neural network based approach. Despite this, for a fair comparison, we re-implement their approach based on our new data setting. Statistics of the new split dataset are shown in Table 1. This dataset retains balanced categories (1,402 *depressed* users and 1,402 *non-depressed* users) like the original dataset (Note that we also evaluate our approach in the imbalance scenario presented in the section of analysis and discussion). To facilitate this corresponding research, the dataset with the new data setting is released as the new benchmark dataset for multimodal depression detection via github[2].

Implementation Details. In our experiments, all hyper-parameters are tuned according to the development set. Specifically, BERT is optimized by the Adam optimizer (Devlin et al., 2019), where $\beta_1 = 0.9$ and the initial learning rate is 10^{-4}. Other parameters of BERT are following (Devlin et al., 2019). For our MTAL approach, we set the dimensions of LSTM hidden states to be 256 and adopt another Adam optimizer (Kingma and Ba, 2015) with an initial learning rate of 10^{-2} and $\beta_1 = 0.9$ for cross-entropy training. The regularization weight is 10^{-5}. The dropout rate is 0.5. For CNN, we set the widths of filters to 3, 4, 5 with 100 features each. The dimension M of topic embeddings in our topic model is 128 and the number of topics (K) is 20. Besides, if a tweet includes no image, the image vector will be initialized as a zero vector.

Evaluation Metrics. The performance is evaluated using standard *Accuracy* (Acc.) and *Macro-F1* (F1) by following Wang et al. (2017). Moreover, t-test is used to evaluate the significance of the performance difference between two approaches by following Yang and Liu (1999).

Baselines. For comparison, we re-implement several approaches as baselines for depression detec-

[1] `https://drive.google.com/file/d/11ye00sHFY5re2NOBRKreg-tVbDNrc7Xd/view`

[2] `https://github.com/MinghuiAn/MTAL`

Approach	P	R	F1	Acc.
Primary Task	81.4	80.9	81.1	80.9
+ Auxiliary Task 1	82.4	82.3	82.3	82.2
+ Auxiliary Task 2	81.6	81.5	81.5	81.5
+ Auxiliary Task 1,2	**84.2**	**84.2**	**84.2**	**84.2**

Table 3: Performances of the primary task (depression detection) with different combinations of auxiliary tasks, i.e., the textual topic modeling and the visual topic modeling.

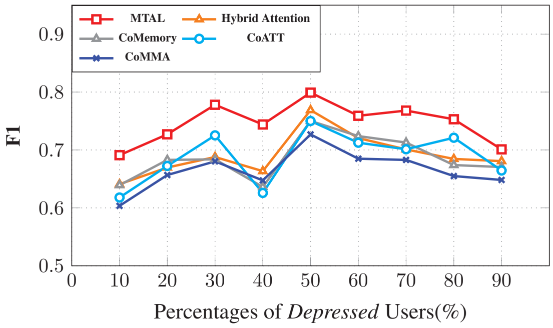


Figure 4: Comparison of various approaches trained on the imbalanced datasets with different percentages of *depressed* users.

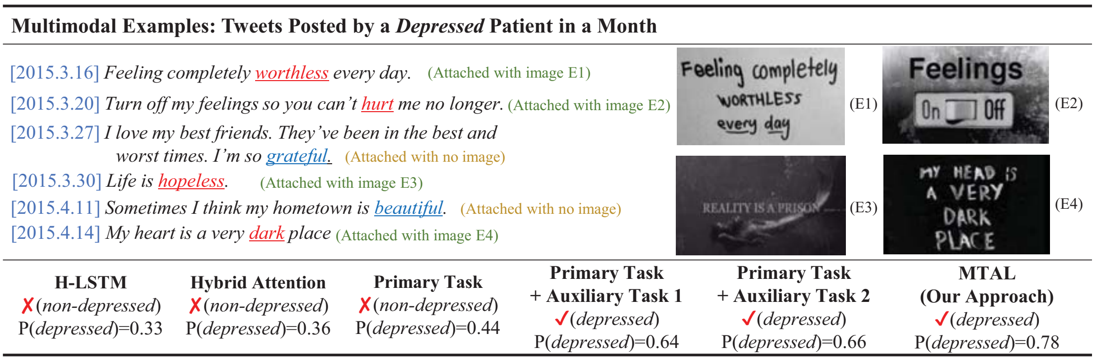
tion. **1) H-LSTM**. This is a hierarchical LSTM approach to aspect sentiment classification. In our implementation, we use it to model word sequence and text sequence for depression classification. **2) BERT+LSTM**. This is a BERT model for encoding each text, followed by an LSTM to encode the text sequence for depression classification. **3) BERT+LSTM+Textual Topic Modeling**. This is an extension of BERT+LSTM with textual topic information. **4) VGG+LSTM**. This is a VGG model for encoding each image, followed by an LSTM to encode the image sequence. **5) VGG+LSTM+Visual Topic Modeling**. This is an extension of VGG + LSTM with visual topic information. **6) EF-LSTM** (Zadeh et al., 2018). This is a state-of-the-art multimodal approach to human communication comprehension task. **7) CoMemory** (Xu et al., 2018). This is a state-of-the-art multimodal approach to the multimodal sentiment analysis task. **8) CoATT** (Zhang et al., 2018). This is a state-of-the-art multimodal approach to the named entity recognition task. **9) Hybrid Attention** (Gu et al., 2018). This is a state-of-the-art multimodal approach using modality attention to learn modality-specific and modality-fusion features for the spoken language classification. Note that the above four multimodal approaches can only encode single text+image pair. In our implementation, we also use an LSTM to encode the final vector sequence of all text+image pairs for depression classification. **10) CoMMA** (Gui et al., 2019b). This is exactly a state-of-the-art multimodal approach to depression detection. In this study, we re-implement it based on our new data setting. **11) Primary Task**. Our approach w/o integrating two auxiliary tasks.

4.2 Experimental Results

Table 2 compares different approaches to the depression detection task. From this table, we can see that:

Single-Modality Performance. When only using the text modality, **1)** The BERT based approach **BERT+LSTM** performs better than **H-LSTM**. This encourages us to use **BERT** as the text encoder for depression detection. **2)** Our approach **BERT+LSTM+Textual Topic Modeling** performs consistently better than **BERT+LSTM**. This encourages us to incorporate the textual topic information for depression detection. When only using the image modality, **1)** The image classification approach **VGG** performs much better than a random performance. This encourages us to consider the image information for depression detection and also indicates the appropriateness of using VGG as the image-region encoder. **2)** Our approach **VGG+LSTM+Visual Topic Modeling** significantly outperforms **VGG+LSTM** (p-value < 0.05). This encourages us to incorporate the visual topic information for depression detection.

Multimodality Performance. When using both the text and image modalities, **CoMemory**, **CoATT** and **Hybrid Attention** perform better than the BERT based single-modal **BERT+LSTM**. This confirms the helpfulness of considering the image information in depression detection. In comparison, our approach **Primary Task** performs consistently better than all the above multimodal approaches in terms of all metrics. This is mainly due to the helpfulness of using BERT as the text encoder. Among all these approaches, our approach **MTAL** performs best and even significantly outperforms (p-value $<$ 0.01) the strong baseline **BERT+LSTM+Textual Topic Modeling** in terms of all metrics. These results encourage us to incorporate both the textual and visual topic information for depression detection.



Multimodal Examples: Tweets Posted by a *Depressed* Patient in a Month

[2015.3.16] *Feeling completely worthless every day.* (Attached with image E1)
[2015.3.20] *Turn off my feelings so you can't hurt me no longer.* (Attached with image E2)
[2015.3.27] *I love my best friends. They've been in the best and worst times. I'm so grateful.* (Attached with no image)
[2015.3.30] *Life is hopeless.* (Attached with image E3)
[2015.4.11] *Sometimes I think my hometown is beautiful.* (Attached with no image)
[2015.4.14] *My heart is a very dark place* (Attached with image E4)

Feeling completely WORTHLESS every day
(E1)
Feelings
On Off
(E2)
REALITY IS A PRISON
(E3)
MY HEAD IS A VERY DARK PLACE
(E4)

H-LSTM	Hybrid Attention	Primary Task	Primary Task + Auxiliary Task 1	Primary Task + Auxiliary Task 2	MTAL (Our Approach)
✗(*non-depressed*)	✗(*non-depressed*)	✗(*non-depressed*)	✓(*depressed*)	✓(*depressed*)	✓(*depressed*)
P(*depressed*)=0.33	P(*depressed*)=0.36	P(*depressed*)=0.44	P(*depressed*)=0.64	P(*depressed*)=0.66	P(*depressed*)=0.78

Figure 5: A *depressed* user example from the test data with output categories and probabilities of the true label *depressed*, predicted by different approaches. ✓(or ✗) denotes that the predicted category is correct (or wrong).

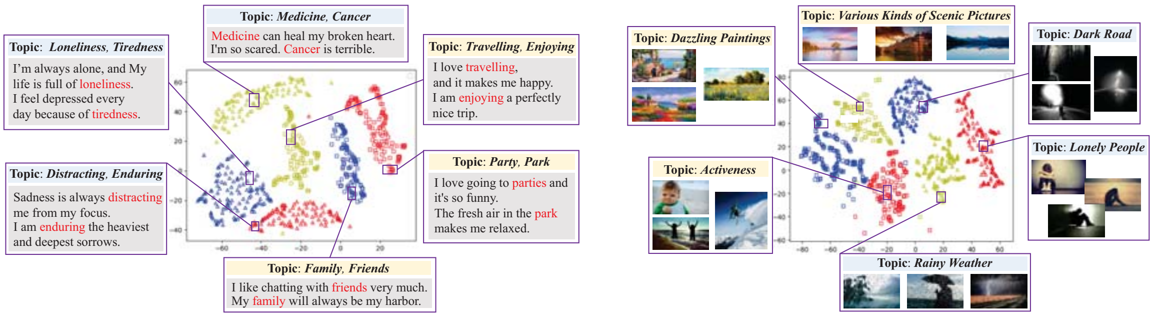
5 Analysis and Discussion

Contribution of Multimodal Topic Information. Table 3 summarizes the results of the primary task integrated with different auxiliary tasks. From this table, we can see that: **1) + Auxiliary Task 1** is superior to **Primary Task** with improving the F1 score by 1.2% (p-value < 0.05). This demonstrates that incorporating the textual topic information is helpful to detect depression. **2) + Auxiliary Task 2** performs slightly better than **Primary task** with the improvement of 0.4% in terms of F1. This demonstrates that incorporating the visual topic information is useful to detect depression. **3) + Auxiliary Task 1,2** performs best and significantly outperforms **Primary Task** (p-value < 0.05) by 3.3% in Acc. and by 3.1% in F1. This indicates that jointly training primary task with two auxiliary tasks can significantly improve the performance and demonstrates that incorporating both the textual and visual topic information can help detect depression.

Analysis of Imbalanced Scenario. In realistic scenarios, only a small proportion of users are *depressed*. Inspired by this, we further evaluate our **MTAL** approach on different percentages of *depressed* users for verifying its robustness. Specifically, we construct 9 different imbalanced training sets where the percentages of *depressed* users are ranging from 10% to 90%, and the total number of users is set to be 1,500 with the train/dev/test setting of 7:1:2. The detail experimental results are shown in Figure 4. From the figure, we can see that our **MTAL** approach can still achieve stable performance even in the case of a very low percentage (10% *depressed* users), and consistently perform better than the state-of-the-art baselines. This can further justify the robustness and effectiveness of our approach.

Qualitative Analysis. Figure 5 shows the multimodal tweets posted by a *depressed* user in a month, together with the predicted categories and probabilities of the ground-true label via different approaches. From this figure, we can see that: **1)** Though this user expresses positive emotions (e.g., "*grateful*" and "*beautiful*") in the time points [2015.3.27] and [2015.4.11], he/she still expresses the negative emotions (e.g., words "*worthless*", "*hurt*" and "*hopeless*" which indicate a *world-weary* topic) in most of the time in a month. Moreover, all images in Figure 5 have the characteristics of low brightness and dark tones, which also indicates the negative emotion of this user. These highlight the importance of capturing the global topic information for depression detection. **2)** Despite that our approach **Primary Task** gives a wrong prediction, it can still obtain higher probability for the true label *depressed* than **Hybrid Attention**. This indicates the appropriateness of using **Primary Task** approach to fuse the texts and images information. Furthermore, when incorporating either the textual or visual topic information, all our approaches including **MTAL** with topic modeling can give correct prediction, i.e., *depressed*, for this user. This again encourages us to incorporate multimodal topic information for depression detection.

User Visualization with Multimodal Topics. We randomly pick 500 users to perform the t-SNE projection according to the textual topic representation $\mathbf{q}_{text}$ or the visual topic representation $\mathbf{q}_{image}$ of each user. Specifically, we randomly pick six topics in each modality (If a topic is not picked, its topic embedding inside topic embedding matrix $\mathbf{\Phi}$ will be set to be zero vector) to compute the textual (or visual) topic representation of each user. Figure 6 (a) and 6 (b) show the user visualization with the



Topic: ***Loneliness, Tiredness***

I'm always alone, and My life is full of loneliness. I feel depressed every day because of tiredness.

Topic: ***Medicine, Cancer***

Medicine can heal my broken heart. I'm so scared. Cancer is terrible.

Topic: ***Travelling, Enjoying***

I love travelling, and it makes me happy. I am enjoying a perfectly nice trip.

Topic: ***Distracting, Enduring***

Sadness is always distracting me from my focus. I am enduring the heaviest and deepest sorrows.

Topic: ***Party, Park***

I love going to parties and it's so funny. The fresh air in the park makes me relaxed.

Topic: ***Family, Friends***

I like chatting with friends very much. My family will always be my harbor.

(a) Textual Topics

(b) Visual Topics

Figure 6: t-SNE visualization of *depressed* users and *non-depressed* users according to six randomly picked topics from each modality respectively. The squares and triangles represent *non-depressed* and *depressed* users respectively. The six topics in Left (a) are generated by the textual topic modeling, while those in Right (b) are generated by the visual topic modeling. Besides, the sample texts and images are randomly selected from tweets posted by the linked users.

textual topics and visual topics respectively. From the two figures, we can see that: 1) *depressed* and *non-depressed* users are clearly divided by both the textual or visual topics, indicating the helpfulness of both the textual and visual topic information for depression detection; 2) textual or visual topics themselves are also clearly divided, indicating the effectiveness of our modality-agnostic topic model in capturing the textual or visual topic information. In summary, these observations suggest us to consider the topic information inside both the texts and images for depression detection.

6 Conclusion

In this paper, we propose a novel multimodal topic-enriched auxiliary learning approach to depression detection. The main idea of the proposed approach is to incorporate not only the textual topic information but also the visual topic information for depression detection. Experimental results demonstrate that the proposed approach significantly outperforms a number of competitive baselines, including the representative textual depression detection approaches and the state-of-the-art multimodal-based approaches.

In our future work, we would like to explore more information, such as the user personal attribute information (such as profession, location and age) and the social attribute information (e.g., timeline, social behavior and social relationships), to assist depression detection. In addition, we would like to apply our approach to other psychological analysis tasks, such as multimodality-based emotion analysis, anxiety detection and personality inference.

Acknowledgments

We thank our anonymous reviewers for their helpful comments, and thank Tao Gui and Liang Zhu for providing datasets, codes and insightful suggestions. This work was supported by three NSFC grants, i.e., No.62006166, No.62076175 and No.61672366. This work was also supported by a Project funded by China Postdoctoral Science Foundation No.2019M661930 and a Project Funded by the Priority Academic Program Development of Jiangsu Higher Education Institutions (PAPD).

References

American Psychiatric Association and others. 2013. *Diagnostic and statistical manual of mental disorders (DSM-5®)*. American Psychiatric Pub.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of Web Science 2013*, pages 47–56.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of ACL-2015*, pages 795–804.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT-2019*, pages 4171–4186.

Adji B. Dieng, Chong Wang, Jianfeng Gao, and John W. Paisley. 2017. Topicrnn: A recurrent neural network with long-range semantic dependency. In *Proceedings of ICLR-2017*.

George Gkotsis, Anika Oellrich, Tim J. P. Hubbard, Richard J. B. Dobson, Maria Liakata, Sumithra Velupillai, and Rina Dutta. 2016. The language of mental health problems in social media. In *Proceedings of CLPsych@NAACL-HLT-2016*, pages 63–73.

Yue Gu, Kangning Yang, Shiyu Fu, Shuhong Chen, Xinyu Li, and Ivan Marsic. 2018. Hybrid attention based multimodal network for spoken language classification. In *Proceedings of COLING-2018*, pages 2379–2390.

Lin Gui, Jia Leng, Gabriele Pergola, Yu Zhou, Ruifeng Xu, and Yulan He. 2019a. Neural topic model with reinforcement learning. In *Proceedings of EMNLP-IJCNLP-2019*, pages 3476–3481.

Tao Gui, Liang Zhu, Qi Zhang, Minlong Peng, Xu Zhou, Keyu Ding, and Zhigang Chen. 2019b. Cooperative multimodal approach to depression detection in twitter. In *Proceedings of AAAI-2019*, pages 110–117.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of SIGIR-1999*, pages 50–57.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR-2015*.

Diederik P. Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *Proceedings of ICLR-2014*.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of ICML-2017*, pages 2410–2419.

Priya Nambisan, Zhihui Luo, Akshat Kapoor, Timothy B. Patrick, and Ron A. Cisler. 2015. Social media, big data, and public health informatics: Ruminating behavior of depression revealed through twitter. In *Proceedings of HICSS-2015*, pages 2906–2913.

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

Philip Resnik, William Armstrong, Leonardo Max Batista Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan L. Boyd-Graber. 2015. Beyond LDA: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of NAACL-2015*, pages 99–107.

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 *Proceedings of IJCAI-2017*, pages 3838–3844.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *Proceedings of ICLR-2015*.

Akash Srivastava and Charles A. Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of ICLR-2017*.

Michel F. Valstar, Jonathan Gratch, Björn W. Schuller, Fabien Ringeval, Denis Lalanne, Mercedes Torres, Stefan Scherer, Giota Stratou, Roddy Cowie, and Maja Pantic. 2016. AVEC 2016: Depression, mood, and emotion recognition workshop and challenge. In *Proceedings of AVEC@MM-2016*, pages 3–10.

Jingjing Wang, Shoushan Li, and Guodong Zhou. 2017. Joint learning on relevant user attributes in micro-blog. In *Proceedings of IJCAI*, pages 4130–4136.

Jingjing Wang, Jie Li, Shoushan Li, Yangyang Kang, Min Zhang, Luo Si, and Guodong Zhou. 2018. Aspect sentiment classification with both word-level and clause-level attention networks. In *Proceedings of IJCAI-2018*, pages 4439–4445.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *Proceedings of AAAI-2020*, pages 9177–9184.

Nan Xu, Wenji Mao, and Guandan Chen. 2018. A co-memory network for multimodal sentiment analysis. In *Proceedings of SIGIR-2018*, pages 929–932.

Yiming Yang and Xin Liu. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-1999*, pages 42–49.

Shi Yin, Cong Liang, Heyan Ding, and Shangfei Wang. 2019. A multi-modal hierarchical recurrent neural network for depression detection. In *Proceedings of AVEC@MM-2019*, pages 65–71.

Amir Zadeh, Paul Pu Liang, Soujanya Poria, Prateek Vij, Erik Cambria, and Louis-Philippe Morency. 2018. Multi-attention recurrent network for human communication comprehension. In *Proceedings of AAAI-2018*, pages 5642–5649.

Qi Zhang, Jinlan Fu, Xiaoyu Liu, and Xuanjing Huang. 2018. Adaptive co-attention network for named entity recognition in tweets. In *Proceedings of AAAI-2018*, pages 5674–5681.