Cooperative Multimodal Approach to Depression Detection in Twitter

Tao Gui,[1*] Liang Zhu,[1*] Qi Zhang,[1] Minlong Peng,[1] Xu Zhou,[1] Keyu Ding,[2] Zhigang Chen[2]
[1]Shanghai Key Laboratory of Intelligent Information Processing, Fudan University
[2]iFlytek Co., Ltd.
[1]{tgui16, liangzhu17, qz, mlpeng16, xuzhou16}@fudan.edu.cn, [2]{kyding, zgchen}@iflytek.com

Abstract

The advent of social media has presented a promising new opportunity for the early detection of depression. To do so effectively, there are two challenges to overcome. The first is that textual and visual information must be jointly considered to make accurate inferences about depression. The second challenge is that due to the variety of content types posted by users, it is difficult to extract many of the relevant indicator texts and images. In this work, we propose the use of a novel cooperative multi-agent model to address these challenges. From the historical posts of users, the proposed method can automatically select related indicator texts and images. Experimental results demonstrate that the proposed method outperforms state-of-the-art methods by a large margin (over 30% error reduction). In several experiments and examples, we also verify that the selected posts can successfully indicate user depression, and our model can obtained a robust performance in realistic scenarios.

Introduction

Depression is a major contributor to the overall global disease burden. According to a recent fact sheet provided by World Health Organization, globally, more than 300 million people of all ages suffer from depression[1]. Although there are known and effective treatments for depression, fewer than half of those affected around the world (in many countries, fewer than 10%) receive treatment (Olfson, Blanco, and Marcus 2016). The traditional diagnosis of depression requires a face-to-face conversation with a medical doctor, which limits the likelihood of the identification of potential patients (Saxena et al. 2007). Unlike the documentation produced by healthcare professionals, social media data captures people's thoughts, feelings, and conversations in their own voices, and these types of data sources are becoming very important for monitoring a number of public health issues including depression (Benton, Coppersmith, and Dredze 2017). In recent years, the concept of detecting depression by harvesting social media data has opened up new possibilities (Shen et al. 2017; Suhara, Xu, and Pentland 2017).

Figure 1: An example of a multimodal tweet. If we consider only the textual content "*Everyone is so happy*," we cannot easily determine the actual feelings of the author. Images posted by users can provide a wealth of information for detecting depression.

Previous deep learning methods have mainly focused on the use of textual information for detecting depression and have achieved great success (Yates, Cohan, and Goharian 2017). None have yet harnessed the wealth of visual data posted on social media. However, many tweets contain both textual content and images. According to a statistical analysis, more than 42% of tweets contain more than one image[2]. Figure 1 shows an example of a multimodal tweet. If we only consider the textual content "*Everyone is so happy*," our understanding of the feelings of the user who posted it may be minimal. In this scenario, images posted by individuals diagnosed with depression can be reliably distinguished from those made by emotionally healthy individuals (Reece and Danforth 2017). Without taking into consideration any visual information, the meanings of certain tweets cannot be easily determined.

What's more, a diagnosis of depression does not mean that the depressive state persists throughout every moment of every day, and to conduct analysis of an individual's entire posting history with respect to a single unit of observation is therefore rather unrepresentative (Reece and Danforth 2017). Because we have no labels at the text and image level,

*Equal contributions

[1]http://www.who.int/mediacentre/factsheets/fs369/en/
[2] https://thenextweb.com/socialmedia/2015/11/03/what-analyzing-1-million-tweets-taught-us/

it becomes difficult to learn which text or image should be selected, and important to determine how to cooperatively select both related indicator texts and images from a user's entire posting history to detect the presence of depression.

In this work, we propose the use of multi-agent reinforcement learning method, which uses two policy gradient agents to simultaneously select texts and images, and evaluate the utility of joint actions based on the classification accuracy. In cooperative settings, joint selection typically generates only global rewards, which make it difficult to determine the contribution of each selector to the group's success (Foerster et al. 2017). To overcome this challenge, the proposed model employs a novel *cooperative misoperation multi-agent* (COMMA) policy gradients. The COMMA takes an *actor-critic* (Konda and Tsitsiklis 2000) approach with differentiated advantages, in which each actor (i.e., text selector or image selector) is trained by following its own unique gradient estimated by a critic. On one hand, we adopt the framework of centralized training with decentralized execution. The critic is only used during learning, whereas only the actor is needed during execution. On the other hand, we use a misoperation to obtain an advantage for each agent. We provide each agent a shaped reward that compares the current global reward to the reward received when the agent takes an opposite action (misoperation). Experimental results show that the proposed method can achieve a much better performance than existing state-of-the-art methods.

The main contributions of this work can be summarized as follows: 1) we study the problem of detecting depression by incorporating textual and visual information; 2) we propose a novel multi-agent reinforcement learning method, COMMA, to achieve this task, in which text and image selectors cooperatively extract indicator content; and 3) experimental results for the depression benchmark show that COMMA can significantly improve performance compared to that of other baseline approaches.

Related Work and Background

Depression Detection Using Social Media

During the last decade, social media have become extremely popular, with billions of users sharing their thoughts and lives on the go. Accordingly, the detection of depression by harvesting social media data is making great progress. Many previous methods have explored a range of features to detect depression, such as sentiment words (Park, Cha, and Cha 2012), social structure and linguistic patterns (Xu and Zhang 2016; De Choudhury, Counts, and Horvitz 2013), photographs (Reece and Danforth 2017), and so on. These works reveal that both textual and visual features are useful in detecting depression. Recently, Yates, Cohan, and Goharian (2017) proposed a model based on convolutional networks to effectively identify depressed users based on textual information. Despite the excellent performance of neural networks, the potential for utilizing image information has not yet been explored. Yang et al. (2017) examined audio and video data from clinical interviews to detect depression, but their dataset contained only 189 examples and is difficult to obtain. In contrast to the above methods, in this study, we combined textual and visual features by applying reinforcement learning to select indicator posts. Our results indicate that the proposed method is strong and stable in its performance in realistic scenarios.

Reinforcement Learning

In this work, we consider a fully cooperative multi-agent task that can be described as an extension of Markov decision processes (MDPs). The MDPs for N agents can be described as a stochastic game G, represented as a tuple $G = \langle N, S, A, \{R_i\}_{i\in N}, \mathcal{T}\rangle$, where S is the set of states, A is the collection of action sets, with a^i being i-th agent's action, $\mathcal{T}$ is the state transition function: $\mathcal{T} : S \times A \to S$, and $\{R_i\}_{i\in N}$ is the set of reward functions. By the joint actions, each agent receives a reward for judging its own action $r_i : S \times A \to \mathbb{R}$, and aims to maximize its total expected return $R_i = \sum_{l=0}^{T} \gamma^l r_{t+l}$, where γ is the discount factor and T is the total time steps.

Q-Learning and Deep Q-Networks. In a particular environment, the Q-learning technique (Watkins and Dayan 1992) estimates the Q-values $Q^\pi(s, a)$, which is a scalar that estimates the expected sum of the gamma-discounted rewards that will accrue by taking action a in state s and following policy π as $Q^\pi(s, a) = \mathbb{E}_{s'}[r(s, a) + \gamma\mathbb{E}_{a'\in\pi}[Q^\pi(s', a')]]$. Deep Q-Networks (Mnih et al. 2013) extend standard Q-learning by using a deep neural network as a Q-value function approximator by minimizing the following loss function:

$$\mathcal{L} = \mathbb{E}_{s,a,r,s'}[(Q^*(s, a|\theta) - y)^2], \tag{1}$$

where $y = r_t + \gamma \max_{a'} \bar{Q}(s', a'|\bar{\theta})$ is the update target given by the target network $\bar{Q}$, whose parameters $\bar{\theta}$ are periodically updated with the most recent θ to stabilize the learning (Mnih et al. 2015).

Policy Gradient and Actor-Critic Algorithms. Policy gradient methods are another type of reinforcement learning technique that rely upon optimizing parametrized policies to maximize the expected return $J(\theta) = \mathbb{E}_\pi[R]$. A classic method is the REINFORCE algorithm (Williams 1992), in which the gradient is as follows

$$\nabla J(\theta) = \mathbb{E}_\pi[\sum_{t=0}^{T} \nabla_\theta[\log \pi_\theta(a_t|s_t; \theta)R]]. \tag{2}$$

Alternatively, using the Q function to serve as a critic for estimating the rewards leads to a class of *actor-critic* algorithms, e.g., that use the *temporal difference* (TD) error $r_t + \gamma V(s_{t+1} - V(s_t))$ (Sutton, Barto, and others 1998).

However, a core issue in multi-agent learning is how to design a reward for each agent, because joint actions typically generate only global rewards. All the agents' actions contribute to that global reward, which may make the gradient of each agent very noisy (Foerster et al. 2017). In the next section, we describe in detail our proposed novel multi-agent reinforcement learning technique for tackling the above problems.

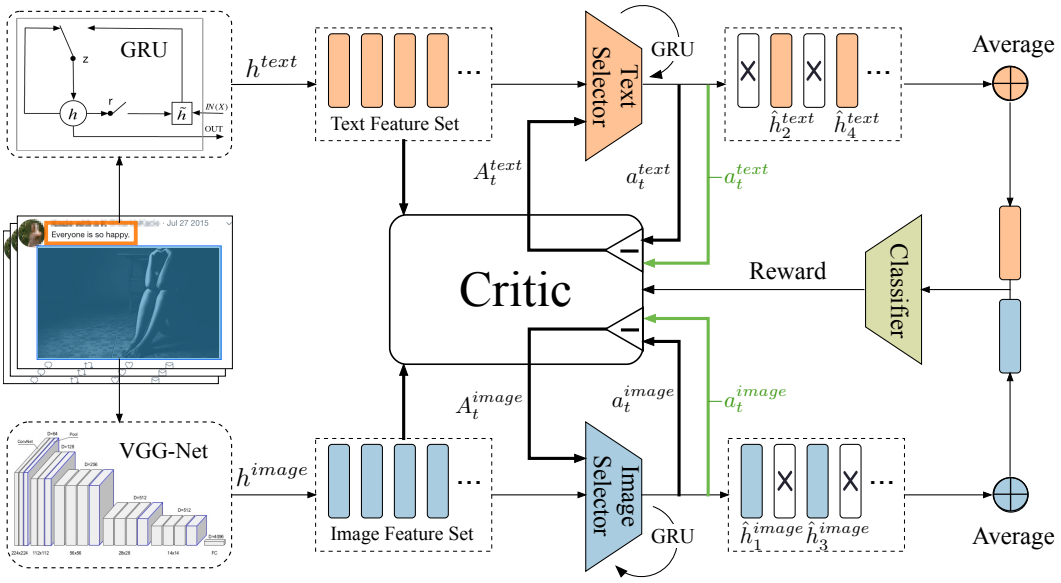

Figure 2: Architecture of the proposed model. At each time step t, the advantage A_t^e of selector e is given by comparing the current global reward to the reward received when that agent's action is replaced with an opposite action $-a_t^e$.

Approach

We propose COMMA policy gradients, which adopt a centralized training framework with decentralized execution by applying a centralized critic and differentiated advantages, as shown in Figure 2. Both the text and image selectors are policy gradient agents, which take the text and image features as inputs and determine whether to select the features. The selectors are trained by following the different gradients estimated by the critic. The differentiated advantages are shaped rewards that compare the current global reward to those received when each agent's action is replaced with an opposite action (misoperation). The text and images features are extracted by GRU and pretrained VGG-Net, respectively, and then the classifier uses the features selected by agents to detect depression.

Feature Extraction

Suppose that we have a set of posts of the u-th user denoted by P_u, which contains T pairs of text and image posted by one user. We use a gated recurrent unit (GRU) (Chung et al. 2014) and convolutional neural networks (CNN) (Simonyan and Zisserman 2014) to extract the textual and visual features, respectively.

Text Feature Extraction. Each text consists of a sequence of words $w_{t_1}, w_{t_2}, \cdots, w_{t_n}$, where $w_{t_i} \in \mathbb{R}^d$ is the d-dimensional word vector for the i-th word in the t-th text and n is the length of the text. Let $\mathcal{V}$ be the vocabulary of words. We used GRU to compute continuous representations of sentences with the following semantic composition:

$$\begin{aligned} z_i &= \sigma(W_z w_{t_i} + U_z h_{i-1}) \\ r_i &= \sigma(W_r w_{t_i} + U_r h_{i-1}) \\ \tilde{h}_i &= \tanh(W_h w_{t_i} + U_h(r_i \odot h_{i-1})) \\ h_i &= (1 - z_i) \odot \tilde{h}_i + z_t \odot h_{i-1}, \end{aligned} \tag{3}$$

where σ is the $sigmoid$ function, $\odot$ is the Hadamard product, and W is the parameters of GRU. We used a bidirectional GRU to encode the text from beginning to end denoted by $\overrightarrow{h}_n$, and from end to beginning denoted by $\overleftarrow{h}_1$. Hence, the feature representation of the t-th text is obtained by concatenation operation: $h_t^{text} = \overrightarrow{h}_{t_n} \oplus \overleftarrow{h}_{t_1}$.

Image Feature Extraction. We extracted image features from a 16-layer pretrained VGGNet (Simonyan and Zisserman 2014). As in previous studies (Gao et al. 2015), we used features identified in the first fully connected layer *fc-4096* that produce a global vector. For the convenience of calculation, we used a multilayer perceptron (MLP) to convert each image vector into a new vector that has the same dimensionality as the textual feature vector:

$$h_t^{image} = tanh(W_v \tilde{h}_t + b_v), \tag{4}$$

where $\tilde{h}_t$ is the output of the first fc-layer, and h_t^{image} is the t-th image feature representation after its transformation by the single-layer perceptron.

Cooperative Misoperation Multi-Agent RL

A diagnosis of depression does not involve a moment-by-moment persistence of the depressive state throughout every day, and using an individual's entire posting history may negatively impact the depression detection outcome (Reece and Danforth 2017). First, we introduce the use of two agents to select indicator texts and images, respectively.

Test Selector and Image Selector. A certain user's historical posts P_u correspond to the sequential inputs of one episode. For the text and image selectors, we set each state at step $t \in T$ as the extracted features h_t^{text} and h_t^{image}, respectively. We denote the action at step $t \in T$ by $a_t \in A = \{0, 1\}$, which indicates whether or not to select the feature.

The process of learning the policy $\pi(a_{1:T})$ must rely on the local action-observation histories during execution. To consider the full history, we used GRU (Chung et al. 2014) networks to model the selectors. Therefore, we formulated the agents' policy $\pi(a_{1:T})$ as follows:

$$\begin{aligned}\pi^e(a_{1:T}) =& \prod_{t=1}^{T} \pi^e(a_t|s_t) \\ g_t^e =& GRU(h_t^e, g_{t-1}^e) \\ \pi^e(a_t|s_t) =& (1-a_t^e) * (1-\sigma(MLP(g_t^e))) \\ &+ a_t^e * \sigma(MLP(g_t^e)),\end{aligned} \tag{5}$$

where $e \in \{text, image\}$. g_t is the hidden state of the GRU and MLP refers to the multilayer perceptron. $\sigma(\cdot)$ is the sigmoid function that transforms g_t into a probability. Next, the selector samples an action to either select the feature ($a_t = 1$) or not ($a_t = 0$). If a certain feature is selected, then the feature h_t^e will be rewritten as $\hat{h}_t^e$, and be appended in H_{indi}^e. The user representation subset H_{indi}^e is then used to predict the depression label.

Depression Classifier. At each step t in one episode, we use the subset containing the selected features to predict depression. Depression classification is thus a universal binary classification problem.

We merged H_{indi}^{text} and H_{indi}^{image} to create a representation of the user. Various merging methods can be applied, such as summation or an attention mechanism. In this work, we used the average operation. As such, we processed this representation by two fully connected layers (i.e., multilayer perceptron) using the dropout operation. The output at the last layer is followed by a sigmoid non-linear layer that predicts the probability distribution over two classes.

$$\begin{aligned}o_t &= MLP(avg(H_{indi}^{text}) \oplus avg(H_{indi}^{image})) \\ Pr(y = \hat{y}_u|o_t; \theta_d) &= \hat{y}_u \sigma(o_t) + (1-\hat{y}_u)(1-\sigma(o_t)),\end{aligned} \tag{6}$$

where $\oplus$ is the concatenation operation and $\hat{y}_u$ represents the prediction probabilities. To encourage the selector to take better actions, we can relate the reward to the likelihood of the ground truth $Pr(y = \hat{y}_u)$. A simple way to do so is to use the actor-critic algorithm, which applies the change in the $Pr(y = \hat{y}_u)$ after updating its sets with the newly chosen examples as the unified TD error, as reported in (Yeung et al. 2017):

$$\begin{aligned}r_t &= Pr(y = \hat{y}_u|o_{t+1}) - Pr(y = \hat{y}_u|o_t) \\ \mathcal{L}_t(\theta_c) &= [r_t + \gamma Q(H_{indi}^{t+1}, \Pi_{t+1}, A_{t+1}) - Q(H_{indi}^t, \Pi_t, A_t)]^2,\end{aligned} \tag{7}$$

where $H_{indi}^t = H_{indi}^{text} \oplus H_{indi}^{image}$, $\Pi = \pi^{text} \oplus \pi^{image}$, and $A = a^{text} \oplus a^{image}$. However, using the same advantages makes it difficult to deduce each selector's own contribution. Hence, differentiating the advantages is a key challenge.

Differentiated Advantages Using Misoperation. In fact, a centralized critic can be used to implement difference rewards in our COMMA setting. Although COMMA learns a centralized critic, which estimates Q-values for joint action $\mathbf{a}$ conditioned on the central state H_{indi}^t, we can provide a particular advantage to each agent by comparing the global reward to the reward when agent takes an opposite action. Intuitively, this mechanism can give a positive reward when the Q-value of gold action subtracts that of opposite action. Formally, for each selector e, we can then compute an advantage function that compares the Q-value for the current action a^e to a misoperation baseline that takes an opposite action $-a^e$, while keeping the other agent's action a^{-e} fixed:

$$A^e(H, \pi, A) = Q(H, \Pi, (a^e, a^{-e})) - Q(H, \Pi, (-a^e, a^{-e})). \tag{8}$$

Hence, $A^e(H, \Pi, A)$ computes a separate baseline for each agent and a centralized critic to consider each advantage. Therefore, the model can be optimized according to the Algorithm 1.

Algorithm 1 COMMA for Depression Detection

1: Randomly initialize critic network $Q(S, \pi, \mathbf{a}|\theta_Q)$ and two selectors $\pi(s|\theta_\pi^e)$ with weights θ_Q and θ_π^e.
2: Initialize target network Q' and π' with weights $\theta_{Q'} \leftarrow \theta_Q$, $\theta_{\pi'}^e \leftarrow \theta_\pi^e$. Initialize replay buffer R
3: **for** episode = 1, M **do**
4: Receive initial observation state h_1^e
5: **for** $t = 1, T$ **do**
6: Select action $a_t^e = \pi(h_t^e|\theta_\pi^e)$ according to the current policy
7: Execute action a_t^e and observe the likelihood of ground truth $\Pr(y = \hat{y}_u|o_t)$ and observe the new state h_{t+1}^e
8: Execute action a_{t+1}^e and observe the likelihood of ground truth $\Pr(y = \hat{y}_u|o_{t+1})$, thereby obtain the reward $r_t = \Pr(y = \hat{y}_u|o_{t+1}) - \Pr(y = \hat{y}_u|o_t)$
9: Store transition $(H_{indi}^t, A_t, r_t, H_{indi}^{t+1})$ in R
10: Sample a random minibatch of N transitions $(H_{indi}^i, A_i, r_i, H_{indi}^{i+1})$ from R
11: Set $z_i = r_i + \gamma Q'(H_{indi}^{i+1}, \Pi_{i+1}, A_{i+1})$
12: Update critic by minimizing the loss: $\mathcal{L}(\theta_Q) = \frac{1}{N}\sum_i [z_i - Q(H_{indi}^i, \Pi_i, A_i|\theta_Q)]^2$
13: Update selectors using differentiated advantages:

$$A^e(H, \Pi, A) = Q(H, \Pi, A) - Q(H, \Pi, (-a^e, a^{-e}))$$

$$\nabla_{\theta_\pi^e} J(\theta_\pi^e) = \nabla_{\theta_\pi^e} \log \pi(a_t^e|h_t^e) A^e(H, \Pi, A)$$

14: Update the target networks:

$$\theta_{Q'} = \tau\theta_Q + (1-\tau)\theta_{Q'}, \theta_{\pi'}^e = \tau\theta_\pi^e + (1-\tau)\theta_{\pi'}^e$$

15: **end for**
16: Update the depression classifier by minimizing the cross entropy loss:

$$J(\theta_C) = -[y_u \log \hat{y}_u + (1-y_u)\log(1-\hat{y}_u)]$$

17: **end for**

Experimental Setup

In this section, we first describe the datasets and then the baseline methods we used, including a number of classic methods and a series of neural networks methods. Finally,

Dataset		# Users	# T	# T + I
$\mathbf{D_1}$	Depressed	1,402	292,564	-
	Non-Depressed	5,160	3,953,183	-
$\mathbf{D_2}$	Depressed	1,402	251,834	40,730
	Non-Depressed	5, 160	3,302,366	650,817

Table 1: Statistical details of the datasets used in our experiments, where **# T** and **# T + I** represent the number of tweets that contain only texts and that contain both text + image pairs, respectively.

we detail the configuration of the proposed model.

Datasets

Textual Depression Dataset $\mathbf{D_1}$. Shen et al. (2017) constructed a textual depression dataset on Twitter. Inspired by the work reported in (Coppersmith, Dredze, and Harman 2014), the authors labeled users as depressed if their anchor tweets satisfied a strict pattern, i.e., "(I'm/ I was/ I am/ I've been) diagnosed depression." Since clinical experience tells us that people should be observed over a period of time, they also obtained all the other tweets published within one month of the anchor tweet. With these, the authors also constructed a non-depression dataset, in which users were labeled as non-depressed if they had never posted any tweet containing the character string "depress."

Multimodal Depression Dataset $\mathbf{D_2}$. Based on the ids of the tweets in D_1, we collected all the images using Twitter APIs. Then, based on these images and tweets, we constructed a new multimodal dataset containing 691 K user-generated image and textual tweet pairs and 3,554 K tweets containing no images. Table 1 shows a statistical summary of the dataset. In our experiments, if a certain tweet contained no image, we used a zero vector to represent the image.

We conducted experiments on the above two datasets, and found that images are important features that facilitate detection of depression.

Comparison Methods

Next, as baselines for comparison, we applied several classic and state-of-the-art methods, including feature-based and neural networks methods.

Feature-based methods. As reported in (Shen et al. 2017), feature-based methods use various features and a wide range of external resources, such as social network features, user profile features, visual features, emotional features, topic-level features, and domain-specific features. As baseline models, we used the Naive Bayes (NB) (Pedregosa et al. 2011), multiple social networking learning (MSNL) (Song et al. 2015), Wasserstein dictionary learning (WDL) (Rolet, Cuturi, and Peyré 2016), and multimodal depressive dictionary learning (MDL) (Shen et al. 2017) methods.

Neural network methods. We compared our model performance with those of some neural network methods that use texts, images, or both to identify depression.

- **Gated recurrent unit (GRU):** We applied GRU to obtain text representations, which we then used for classification.
- **VGG-Net:** We applied VGG-Net to obtain image representations, which we then used for classification.
- **MultiModal Methods:** We applied several recently proposed multimodal methods to the task of depression detection, including the co-attention (Lu et al. 2016), dual-attention (Nam, Ha, and Kim 2017), and modality attention (Moon, Neves, and Carvalho 2018) methods.
- **Critic using unified advantages:** A strong baseline model uses the same text and image selectors, and unified advantages are then generated to optimize the selectors (Egorov 2016).
- **Random sampling:** We also randomly sampled half of the posts from each user to train the baseline model.

Initialization and Hyperparameter

We initialized the word embeddings and other parameters related to the deep learning models by randomly sampling from a standard normal distribution and a uniform distribution in [-0.05, 0.05], respectively. We set the dimensionality of the word embedding to 128. In addition, we used one layer of bidirectional GRU to model the post text, and set the hidden neurons of each GRU to 64. To extract image features, we applied VGG-Net, and mapped the output of the first fully connected layer to a vector with a dimensionality of 128. Each of text and image selectors use a GRU with 50 hidden neurons.

Our model can be trained end-to-end with backpropagation, and we performed gradient-based optimization using the Adam update rule (Kingma and Ba 2014), with a learning rate of 0.001.

Results and Analysis

In this section, we detail the performances of the proposed and baseline models, and present the results of a series of experiments to demonstrate the effectiveness of the proposed model from different aspects.

Method Comparison

To make a fair comparison, we constructed the training and test sets in the same way as reported in (Shen et al. 2017). With 1,402 depressed users in total, we randomly selected 1,402 non-depressed users in D_1 to make the percentage of depressed users equal to 50%, but we did so in a more difficult manner by removing all the anchor tweets. After obtaining the dataset, we trained and tested these methods using five-fold cross validation.

We compared the depression detection performance of the proposed and baseline models in terms of four selected measures, i.e., accuracy, macro-averaged precision, macro-averaged recall, and a macro-averaged F1-measure. Table 2 shows a summary of our results.

In the table, the first four lines list the results of the feature-based methods, in which the visual and textual features are the two main features contributing to the performance of the models. Hence, earlier neural networks models that use only textual information may be limited in their performances.

Methods	Training Data	Accuracy	Precision	Recall	F1
NB (Pedregosa et al. 2011)	Various Features	0.724	0.727	0.728	0.728
MSNL (Song et al. 2015)		0.818	0.818	0.818	0.818
WDL (Rolet, Cuturi, and Peyré 2016)		0.768	0.769	0.768	0.768
MDL (Shen et al. 2017)		0.848	0.848	0.850	0.849
GRU (Chung et al. 2014)	Text	0.824	0.825	0.823	0.824
GRU + Random sampling		0.760	0.760	0.757	0.756
VGG-Net (Simonyan and Zisserman 2014)	Image	0.702	0.703	0.702	0.702
VGG-Net + Random sampling		0.642	0.643	0.642	0.643
GRU + VGG-Net	Text+Image	0.845	0.843	0.847	0.845
GRU + VGG-Net + Random sampling		0.811	0.811	0.810	0.810
Co-Attention (Lu et al. 2016)		0.866	0.871	0.863	0.865
Dual-Attention (Nam, Ha, and Kim 2017)		0.848	0.848	0.848	0.848
Modality Attention (Moon, Neves, and Carvalho 2018)		0.866	0.868	0.862	0.864
GRU + VGG-Net + Unified advantages (Egorov 2016)		0.866	0.866	0.865	0.865
GRU + VGG-Net + COMMA (text + image)		**0.900**	**0.900**	**0.901**	**0.900**

Table 2: Comparison of performances in terms of four selected measures.

The second and third part of Table 2 show the results of models using only textual and visual features, respectively. We can see that the performance when using only texts is better than that when using only images, maybe because the volume of texts is nearly six times greater than that of images. In addition, we also note that although the accuracies of both models are no higher than 83%, the summation of the successfully predicted depressed users covers the 95% of the total depressed users. Hence, textual and visual information must represent different user aspects. Next, we show that combining texts and images to detect depression significantly improves performance.

The fourth part of Table 2 shows the results of multimodal methods that combine both texts and images. We can see that simply incorporating the multimodal features, GRU + VGG (text + image) improves the F1 score by 2.1% from using GRU (text) and by 14.3% from using VGG-Net (image) alone. We also tested three more complex multimodal methods on our dataset, including state-of-the-art methods for visual question answering (co-attention and dual-attention) and named entity recognition tasks (modality attention). We can see that these three methods achieves a slightly better performances than GRU + VGG (text + image). However, the improvement is not intuitively obvious. Through our analysis, we found that the average user generates more than 600 tweets and only a small number of which are relevant to depression. As such, attention-based multimodal methods may find it difficult to capture the indicator information.

We propose the adoption of a reinforcement-learning-based selection strategy to tackle the above challenge. Compared to random sampling models, a model adopting unified advantages achieves a much better performance, with an 86.5% F1 score, which is the same level of performance as state-of-the-art multimodal methods. This proves the significance of the selection strategy. However, as explained above, when using unified advantages it is difficult to optimize each selector. If we use the differentiated advantages proposed in our model to optimize the selectors, the COMMA achieves the best performance, with a value of more than 90% for the F1-measure. This indicates that using differentiated advantages is much more effective than using unified advantages. Next, we demonstrate why our selection strategy is more effective than other strategies.

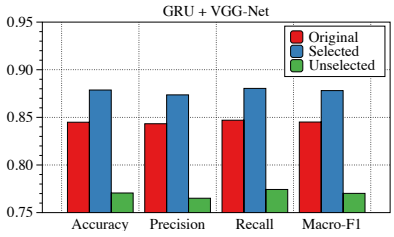

Figure 3: Comparison of models trained on original posts, selected posts, and unselected posts.

Effectiveness of Selected Data

To verify the effectiveness of our proposed method, we applied the COMMA to select depression indicator texts and images from the original dataset, and then compared the baseline model (GRU + BGG-Net) trained on the original dataset, selected dataset, and unselected dataset.

In the results shown in Figure 3, we can see that the model performance benefits from the selected posts, performing almost 3.3% better than those on original dataset, with an error reduction rate greater than 21.7%. Inevitably, the unselected dataset achieved poor performance. These results also indicate that the agent can select depression indicator posts that are more beneficial to depression classification.

Robustness Analysis in Realistic Scenarios

In the previous section, we constructed a training set that in the same way as that reported in (Shen et al. 2017). However, in realistic scenarios, there may be only a small proportion of depressed users. To verify the robustness of the proposed model, we tested their performances on different proportions of depressed users. With a total of 1,402 depressed users, we

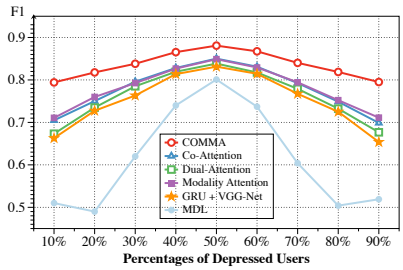

Figure 4: Comparison of the models trained on the datasets with different percentages of depressed users. The total number of users is 1,500.

Dataset	Top words (by frequency)
Selected data of depressed users	bad, cancer, insurance, hate, medical, pain, cost, mental, ...
Unselected data of depressed users	people, online, time, know, life, free, school, weight, work, ...
Original data of non-depressed users	wow, idk, like, party, gotta, funny, ◡̈, honestly, team, :) ...

Table 3: Example words arranged in descending order of word frequency.

fixed the capacity of our dataset to 1,500 and varied the percentage of depressed users from 10% to 90% at increments of 10%, the experimental results of which are shown in Figure 4. From the figure, we can see that our method achieved stable and outstanding performance despite even a very low proportion of users with depression. However, when the percentages of depressed users does not lay at 50%, we observed poor performances of other methods. When we only have ten percent of depressed users in dataset, the proposed model still achieves good performance with a almost 80% F1 score, which performs 10% better than other methods. Therefore, our method is more robust than other methods in realistic scenarios.

Indicator Posts Discovery

To further investigate the kinds of posts that indicate depression and the differences between depressed and non-depressed users, we made a statistical comparison of selected (indicator) posts and other posts.

For images, we computed the pixel-level averages with respect to hue, saturation, and value (HSV), i.e., the three color properties commonly used in image analysis, where lower hue values indicate more red and higher hue values indicate more blue. Saturation refers to the vividness of an image. Value refers to image brightness, where lower brightness scores indicate a darker image. We also used detection software[3] to determine whether or not a photograph contained a human face, and to count the total number of faces in each photo. We then compared the original dataset and that selected by COMMA, as shown in Figure 5, which computes the ratio of values of depressed users to those of

[3]https://pypi.org/project/face_recognition/

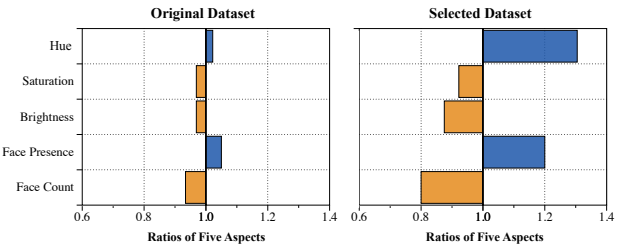

Figure 5: Comparison of original and selected posts. The y-axis values show the five aspects of each image, and the x-axis values are the ratios of these five aspect values of depressed users to those of non-depressed users.

non-depressed users. In the figure, we can see that in the original dataset, the difference between depressed and non-depressed users is marginal, because of the non-persistence of the depressive state in every moment. Hence, using all the data to detect depression is very difficult. In the selected dataset, the depressed users were associated with increased hue, along with decreased brightness and saturation. Also, depressed users were more likely to post photos with faces, but had a lower average face count per photograph than non-depressed users. These results are consistent with those reported in (Reece and Danforth 2017) and verify the effectiveness of selected posts.

For texts, we compute word frequencies in the selected dataset. From the dataset, we removed all stop words, pronouns, numbers, and neutral words (Baccianella, Esuli, and Sebastiani 2010). Finally, we arranged the remaining words in descending order of word frequency, as shown in Table 3. We found that words about medical treatment and negative emotions may indicate depression. Non-depressed users tended to use words with positive emotions and more emojis. We release all the codes and results for further research.

Conclusions

In this study, we investigated the problem of detecting depression based on a combination of texts and images posted by users on social media and proposed a new cooperative multi-agent reinforcement learning method. We verified that the proposed model can cooperatively select indicator texts and images, which are more beneficial to the classifier. Our experimental results demonstrate that the proposed method achieved better performance than existing methods by a large margin (over 30% error reduction). Through several experiments, we also found that our model obtained a strong and stable performance in realistic scenarios. In addition, we investigate the characteristics of the selected texts and images, and found that the selected posts can indicate depression effectively.

Acknowledgments

The authors wish to thank the anonymous reviewers for their helpful comments. This work was partially funded by China National Key R&D Program (No. 2017YFB1002104, 2018YFC0831105), National Natural

Science Foundation of China (No. 61751201, 61532011, 61473092, and 61472088), and STCSM (No. 16JC1420401, 17JC1420200).

References

Baccianella, S.; Esuli, A.; and Sebastiani, F. 2010. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In *LREC*, volume 10, 2200–2204.

Benton, A.; Coppersmith, G.; and Dredze, M. 2017. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, 94–102.

Chung, J.; Gulcehre, C.; Cho, K.; and Bengio, Y. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Coppersmith, G.; Dredze, M.; and Harman, C. 2014. Quantifying mental health signals in twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, 51–60.

De Choudhury, M.; Counts, S.; and Horvitz, E. 2013. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, 47–56. ACM.

Egorov, M. 2016. Multi-agent deep reinforcement learning.

Foerster, J.; Farquhar, G.; Afouras, T.; Nardelli, N.; and Whiteson, S. 2017. Counterfactual multi-agent policy gradients. *arXiv preprint arXiv:1705.08926*.

Gao, H.; Mao, J.; Zhou, J.; Huang, Z.; Wang, L.; and Xu, W. 2015. Are you talking to a machine? dataset and methods for multilingual image question. In *NIPS*, 2296–2304.

Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Konda, V. R., and Tsitsiklis, J. N. 2000. Actor-critic algorithms. In *NIPS*, 1008–1014.

Lu, J.; Yang, J.; Batra, D.; and Parikh, D. 2016. Hierarchical question-image co-attention for visual question answering. In *NIPS*, 289–297.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Graves, A.; Antonoglou, I.; Wierstra, D.; and Riedmiller, M. 2013. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*.

Mnih, V.; Kavukcuoglu, K.; Silver, D.; Rusu, A. A.; Veness, J.; Bellemare, M. G.; Graves, A.; Riedmiller, M.; Fidjeland, A. K.; Ostrovski, G.; et al. 2015. Human-level control through deep reinforcement learning. *Nature* 518(7540):529.

Moon, S.; Neves, L.; and Carvalho, V. 2018. Multimodal named entity recognition for short social media posts. *arXiv preprint arXiv:1802.07862*.

Nam, H.; Ha, J.-W.; and Kim, J. 2017. Dual attention networks for multimodal reasoning and matching. In *CVPR*, 299–307.

Olfson, M.; Blanco, C.; and Marcus, S. C. 2016. Treatment of adult depression in the united states. *JAMA internal medicine* 176(10):1482–1491.

Park, M.; Cha, C.; and Cha, M. 2012. Depressive moods of users portrayed in twitter. In *Proceedings of the ACM SIGKDD Workshop on healthcare informatics (HI-KDD)*, volume 2012, 1–8. ACM New York, NY.

Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research* 12(Oct):2825–2830.

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

Rolet, A.; Cuturi, M.; and Peyré, G. 2016. Fast dictionary learning with a smoothed wasserstein loss. In *Artificial Intelligence and Statistics*, 630–638.

Saxena, S.; Thornicroft, G.; Knapp, M.; and Whiteford, H. 2007. Resources for mental health: scarcity, inequity, and inefficiency. *The lancet* 370(9590):878–889.

Shen, G.; Jia, J.; Nie, L.; Feng, F.; Zhang, C.; Hu, T.; Chua, T.-S.; and Zhu, W. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI-17*, 3838–3844.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Song, X.; Nie, L.; Zhang, L.; Akbari, M.; and Chua, T.-S. 2015. Multiple social network learning and its application in volunteerism tendency prediction. In *SIGIR*, 213–222. ACM.

Suhara, Y.; Xu, Y.; and Pentland, A. 2017. Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks. In *WWW*, 715–724.

Sutton, R. S.; Barto, A. G.; et al. 1998. *Reinforcement learning: An introduction*. MIT press.

Watkins, C. J., and Dayan, P. 1992. Q-learning. *Machine learning* 8(3-4):279–292.

Williams, R. J. 1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning* 8(3-4):229–256.

Xu, R., and Zhang, Q. 2016. Understanding online health groups for depression: social network and linguistic perspectives. *Journal of medical Internet research* 18(3).

Yang, L.; Jiang, D.; Han, W.; and Sahli, H. 2017. Dcnn and dnn based multi-modal depression recognition. In *ACII, 2017 Seventh International Conference on*, 484–489. IEEE.

Yates, A.; Cohan, A.; and Goharian, N. 2017. Depression and self-harm risk assessment in online forums. *arXiv preprint arXiv:1709.01848*.

Yeung, S.; Ramanathan, V.; Russakovsky, O.; Shen, L.; Mori, G.; and Fei-Fei, L. 2017. Learning to learn from noisy web videos. In *2017 CVPR*, 7455–7463. IEEE.