Depression Detection on Social Media with Reinforcement Learning

Tao Gui, Qi Zhang(✉), Liang Zhu, Xu Zhou, Minlong Peng, and Xuanjing Huang

Shanghai Key Laboratory of Intelligent Information Processing, Fudan University, School of Computer Science, Fudan University, 825 Zhangheng Road, Shanghai, China
{tgui16,qz,liangzhu17,xuzhou16,mlpeng16,xjhuang}@fudan.edu.cn

Task
Idea
Results
Critique

Abstract. Depression detection is a significant issue for human well-being. Conventional diagnosis of depression requires a face-to-face conversation with a doctor, which limits the likelihood of the identification of potential patients. We instead explore the potential of using only the textual information to detect depression based on the content users posted on social media sites. Since users may post a variety of different kinds of content, only a small number of posts are relevant to the signs and symptoms of depression. We propose the use of reinforcement learning method to automatically select the indicator posts from the historical posts of users. Our experimental results demonstrate that the proposed method outperforms both feature-based and neural network-based methods (over 14.6% error reduction). In addition, a series of experiments demonstrate that our model can deal with the noise of data effectively and can generalize to more complex situations.

Keywords: Depression · Social media · Reinforcement learning

1 Introduction

Depression is a worldwide prevailing mental disease and a major contributor to the overall global disease burden. A recent fact sheet provided by World Health Organization shows that more than 300 million people of all ages suffer from depression globally[1]. The conventional clinical diagnosis of depression requires a face-to-face conversation between a doctor and patient, which is not available to many potential patients, especially in the early stages. On the other hand, social media is continuously growing and is set to be the communication medium of choice for most people. According to a report published by *The Next Web*, there are over 3 billion social media users around the world[2]. Users post large quantities of content about their daily lives and feelings. Hence, in recent years,

[1] http://www.who.int/mediacentre/factsheets/fs369/en/.
[2] https://thenextweb.com/contributors/2017/08/07/number-social-media-users-passes-3-billion-no-signs-slowing/.

M. Sun et al. (Eds.): CCL 2019, LNAI 11856, pp. 613–624, 2019.
https://doi.org/10.1007/978-3-030-32381-3_49

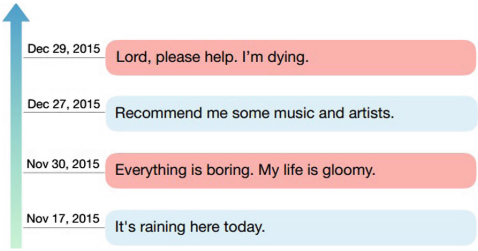

Fig. 1. An example of a user's historical posts. Depression indicator posts are usually sparse on social media. Only the tweets with red highlights may be regarded as the indicators of depression. (Color figure online)

the task of detecting depression via harvesting social media data has received considerable attention [3,12,15].

Previous researchers studied the task of detecting depression via social media using various features, including language, emotion, style and user engagement [3,10]. [12] proposed the use of well-defined discriminative depression-oriented feature groups and a multimodal depressive dictionary learning method to detect depressed users. These methods have proved that the diagnosis of depression through the content published by users on social media is reasonable and feasible. However, most of the existing methods used hand-crafted feature groups to perform the task. In addition to the content users posted on social media, features extracted from user behaviors were also taken into consideration, e.g., hospital attendance [15]. In some cases these user behaviors were hard to be captured [12], which limited the usability of these methods.

In this work, we propose a method to achieve the task using only the historical posts of users. Because the content posted by users on social media are diverse and multi-faceted, depression indicator posts are usually sparse on social media. Figure 1 illustrates an example. From this example, we can observe that there is only two tweets related to the indicators of depression. The other ones are related to the music and weather. If the entire posting history of a user is used as inputs, these content may negatively impact the depression detection. Hence, the model should extract the indicator posts separately from the posting history of a user. However, since it is difficult and a time-consuming task to label each post, most of the benchmark datasets contain only labels at the user level.

To overcome this issue, in this work, we propose a reinforcement learning-based method to achieve the task. Even though we only have the label at the user level, we can evaluate the utility of the selected posts based on the classification accuracy. Our key insight is that the post selection policies can be learned from the utility of the selected posts. Intuitively, a good policy selects posts in a way that allows a classifier trained on these posts to achieve high classification accuracy. Although selecting posts is a non-differentiable action, it can be naturally achieved in a reinforcement learning setting, where actions correspond to the selection of posts and the reward is the effect on the downstream classifier

accuracy. Inspired by the work [1], the proposed method consists of two components: a policy gradient agent, which selects depression indicator posts from the entire posting history of users, and a depression classifier trained using the selected posts. Experimental results show that the proposed method can achieve a much better performance than existing state-of-the-art methods.

2 Related Work

During the last decade, social media have become extremely popular, on which billions of users write about their thoughts and lives on the go. Therefore, researchers began analyzing the online behaviors of users to identify depression. [8] explored the potential benefits of using online social network data for clinical studies on depression. They utilized the real-time moods of users captured on the Twitter social network and explored the use of language in describing depressive moods. In their later work, [9] found that depressed individuals tended to perceive Twitter as a tool for social awareness and emotional interaction. Recently, [18] attempted to explain how web users discuss depression-related issues from the perspective of the social networks and linguistic patterns revealed by the members' conversations. In this work, we studied the problem from a text classifier perspective. Inspired by these works, we also proposed the use of only the textual posts of a user to detect depression.

There is a growing body of research focusing on the use of machine learning to analyze and detect depression via social media. [3] used crowdsourcing to collect gold standard labels and applied an SVM to predict depression of an individual. [10] studied the use of supervised topic models in the analysis of linguistic signal for detecting depression, and provided promising results using several models. Most recently, [12] released a well-labeled depression and non-depression dataset on Twitter, and proposed a multimodal depressive dictionary learning model to detect depressed users on Twitter. [19] proposed a model based on a convolutional network to effectively identify depressed users based on textual information. In contrast to these research, we applied reinforcement learning to select indicator posts, and obtained better results. In addition, the proposed method demonstrated a strong and stable performance in realistic scenarios.

3 Approach

In this work, we propose to study the task of detecting depression based on the content of users posted on the social media. We denote the historical posts of the i-th user as $P^i_{hist} = \{p^i_1, p^i_2, ..., p^i_T\}$, where p^i_t is the text of t-th post. Each user has one label y_i corresponding to whether the user is depressed or not. Based on the description given in the previous section, we know that only a small number of posts may related to signs and symptoms of depression. Hence, we try to select a subset P^i_{indi}, which contains depression indicator posts, from the entire posts of users. The depression classifier is trained based on P^i_{indi}.

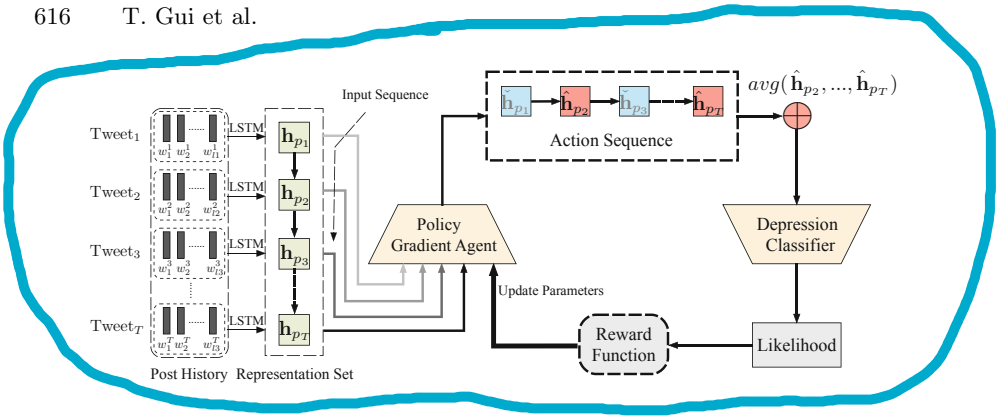

Fig. 2. Architecture of reinforcement learning-based depression detection network. w refers to the word embedding, and $\oplus$ refers to average operation. In the process of training, the policy gradient agent selects post representations in sequence, and then the selected posts (with red highlights) are used to train a better depression classifier. The policy gradient agent computes the rewards based on the likelihood of ground truth to update its parameters. (Color figure online)

The architecture of the proposed method is shown in Fig. 2. It consists of two components: (1) a policy gradient agent [17] that selects depression indicator posts from P_{hist}, and (2) a depression classifier trained using the indicator posts for classification and returning the rewards to the agent. Our goal for training is to optimize the parameters of the depression classifier, which are denoted as θ_d, together with the agent parameters θ_a. The two components should interact with each other to update the parameters during the training process.

3.1 Policy Gradient Agent

We wish to select a subset of depression indicator posts. However, it becomes a key challenge when encountering the diverse content of posts. In addition, the nature of discrete selection decisions makes the loss no longer differentiable. To overcome this problem, we propose the use of reinforcement learning for the task.

The i-th user's historical posts P^i_{hist} correspond to the sequential inputs of one episode. At each step, the agent chooses an action a_t (selecting the current post or not) after observing the state s_t, which is represented by the current post, selected posts and irrelevant posts. When all of the selections are made, the depression classifier will give a delayed reward to update the parameters of agent θ_a.

Next, we will introduce several key points of the agent, including the state representation s_t, the action a_t, and the reward function.

State Representation. We suppose that each post is made up of a sequence of words $p_t = [w_1, w_2, ..., w_l]$, where l is the max length of the post. We use long short-term memory (LSTM) [4] to model each post text. Then, the last hidden state $\mathbf{h}_t$ of LSTM turns into the post representation that will be transferred to the agent, i.e., $\mathbf{h}_t = LSTM(p_t)$. At the step t, the model has obtained t posts

as inputs, which are denoted by $P_{1:t}$. Given $P_{1:t}$, the policy gradient agent could make the following observations: the current post representation $\mathbf{h}_t$, the indicator post set $\mathbf{H}_{indi} = [\hat{\mathbf{h}}_1, \hat{\mathbf{h}}_2, ...]$, and the irrelevant post set $\mathbf{H}_{irre} = [\check{\mathbf{h}}_1, \check{\mathbf{h}}_2, ...]$. The notations $\hat{\mathbf{h}}$ and $\check{\mathbf{h}}$ will be defined in the **action** part. Note that at the initial time, $\mathbf{H}_{indi}$ and $\mathbf{H}_{irre}$ are empty set. We use zero vectors to initialize these two sets. We thereby formulate the agent's state s_t as follows:

$$s_t = [\mathbf{h}_t \otimes avg(\mathbf{H}_{indi}) \otimes avg(\mathbf{H}_{irre})], \tag{1}$$

where avg refers to the average pooling operation, and $\otimes$ is the concatenation operation.

Action. The agent takes an action a_t at step t using policy $a_t \sim \pi(s_t, a_t; \theta_a)$, which is attained by sampling from the multinomial distribution. We define action $a_t \in \{1, 0\}$ to indicate whether the agent will select the current post p_t. Therefore, we could adopt a logistic function to sample the actions from the policy function as follows:

$$\begin{aligned} \pi(a_t|s_t; \theta_a) &= Pr(a_t|s_t) \\ &= a_t * \sigma(MLP(s_t)) + (1 - a_t) * (1 - \sigma(MLP(s_t))), \end{aligned} \tag{2}$$

where MLP represents the multilayer perceptron used to map the state s_t to a scalar, and $\sigma(.)$ is the sigmoid function. If the agent takes an action to select the post ($a_t = 1$), then the hidden state $\mathbf{h}_t$ will be rewritten as $\hat{\mathbf{h}}$ and be appended in $\mathbf{H}_{indi}$. Otherwise, it will be rewritten as $\check{\mathbf{h}}$ and be appended in $\mathbf{H}_{irre}$.

Reward Function. After executing a series of actions, the agent will construct a depression indicator post representation set $\mathbf{H}_{indi}$. The set $\mathbf{H}_{indi}$ is used for classification and will be described in Sect. 3.2. Note that we set the reward to be the likelihood of the ground truth after finishing all the selections of the i-th user. In addition, to encourage the model to delete more posts, we include a regularization to limit the number of selected posts as follows:

$$r_i = Pr(y_i|H_{indi}; \theta_d) - \lambda T'/T, \tag{3}$$

where T' refers to the number of selected posts and λ refers to a hyperparameter to balance the reward. By setting the reward to be the likelihood of the ground truth, we capture the intuition that optimal selections will promote the probability of the ground truth. Therefore, by interacting with the classifier through the rewards, the agent is incentivized to select the optimal posts from P_{hist} for training a good classifier.

3.2 Depression Classifier

Depression classification is a universal binary classification problem. As previously mentioned, at the end of each episode, the post representation subset $\mathbf{H}_{indi}$ is further used to predict the depression label.

We merged $\mathbf{H}_{indi}$ to create a representation of the user's activity across all of the depression related posts. Various merging methods can be applied, such as summation and the attention mechanism, and so on. In this work, we adopted the average operation. This representation is then processed by two fully connected layers (i.e., multilayer perceptron) with the dropout [14] operation. The output at the last layer will be followed by a sigmoid non-linear layer that predicts the probability distribution over two classes.

$$\begin{aligned} o_t &= MLP(avg(\mathbf{H}_{indi})) \\ Pr(\hat{y}_i|\mathbf{H}_{indi};\theta_d) &= \hat{y}_i\sigma(o_t) + (1-\hat{y}_i)(1-\sigma(o_t)), \end{aligned} \tag{4}$$

where $\hat{y}_i$ represents the prediction probabilities, and o_t is the output unit of the fully connected layers.

3.3 Optimization

We train the agent using a standard reinforcement learning algorithm called REINFORCE [17]. The objective of training the agent is maximizing the expected reward under the distribution of the selection policy:

$$J_1(\theta_a) = \mathbb{E}_{\pi(a_{1:T})}[r], \tag{5}$$

where $\pi(a_{1:T}) = \prod_{t=1}^{T} Pr(a_t|s_t;\theta_a)$.

However, the gradient is intractable to obtain because of the discrete actions and high dimensional interaction sequences. Following the REINFORCE algorithm, an approximated gradient can be computed as follows:

$$\begin{aligned} \nabla_{\theta_a} J_1(\theta_a) &= \sum_{t=1}^{T} \mathbb{E}_{\pi(a_{1:T})}[\nabla_{\theta_a} \log(Pr(a_t|s_t;\theta_a)) * r] \\ &\approx \frac{1}{N}\sum_{n=1}^{N}\sum_{t=1}^{T}[\nabla_{\theta_a} \log(Pr(a_t|s_t;\theta_a)) * r^n], \end{aligned} \tag{6}$$

where N denotes the quantity of sampling on one user. In our experiment, $N = 1$ is enough to obtain great performance. By applying the above algorithm, the loss $J_1(\theta_a)$ can be computed by standard backpropagation.

Optimizing the classifier is straightforward, and can be treated as a classification problem. Because the cross entropy loss $J_2(\theta_d)$ is differentiable, we can apply backpropagation to minimize it as follows:

$$J_2(\theta_d) = -[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)], \tag{7}$$

where $\hat{y}_i$ is the output of the classifier. Then, we can get the final objective by minimizing the following function:

$$J(\theta_a, \theta_d) = \frac{1}{M}[\sum_{m=1}^{M}(-J_1(\theta_a) + J_2(\theta_d)], \tag{8}$$

where M denotes the quantity of the minibatch, and the objective function is fully differentiable.

Table 1. Statistical details of the datasets used in our experiments, where **# Users** and **# Tweets** represent the number of users and tweets, respectively.

Dataset		# Users	# Tweets
D_1	Depressed	1,402	292,564
	Non-depressed	5,160	3,953,183
D_2	Candidate	36,993	35,076,667

4 Experimental Setup

In this section, we first describe the datasets used for experiments. Then, we detail describe several baseline methods and the hyperparameters of our model.

4.1 Datasets

We used the depression datasets introduced by [12]. They constructed a well-labeled depression dataset on Twitter. They also constructed an unlabeled depression-candidate dataset. The statistics of these datasets are summarized in Table 1.

Depression Dataset D_1. The depression dataset D_1 was constructed based on the tweets between 2009 and 2016. This dataset contained 1,402 depressed users and 5,160 non-depressed users with 4,245,727 tweets within one month. According to [2], users were labeled as depressed if their anchor tweets satisfied the strict pattern "*(I'm/ I was/ I am/ I've been) diagnosed with depression*". The non-depressed users were labeled if they had never posted any tweet containing the character string "depress".

Depression-Candidate Dataset D_2. The unlabeled depression-candidate dataset D_2 was constructed based on the tweets on December 2016. The users in D_2 were obtained if their anchor tweets loosely contained the character string "depress". By this method, D_2 would contain more depressed users than randomly sampling. Finally, D_2 contained 36,993 depression-candidate users and over 35 million tweets within one month, which will be used for indicator posts discovery.

4.2 Comparison Methods

We applied several classic and state-of-the-art methods for comparison. In addition, we used a series of deep learning methods as baselines for comparison.

Feature-Based Methods. The feature-based methods used various features and a lot of external resources, such as social network features, user profile features, visual features, emotional features, topic-level features, and domain-specific features as shown in [12]. The methods of Naive Bayes (NB), multiple social networking learning (MSNL) [13], Wasserstein dictionary learning

(WDL) [11], and multimodal depressive dictionary learning (MDL) [12] are used as baseline models.

Neural Network Methods. We also made a comparison to a series of neural network methods. These methods just used the context information to identify depression, i.e., the users' posts were the only resource for all the methods.

- **Convolutional neural networks (CNN):** CNN has been widely applied to text classification [5]. We used CNN to model each post of users to obtain the post representations, which would be merged to identify depression [19].
- **Long short-term memory (LSTM):** Similar to CNN, we applied LSTM to obtain the representations of posts, which were then used for classification.
- **SDP-attention and MPSDP-attention:** We introduced two self-attention mechanisms on post level as our baselines. One was defined as $Attention(Q, K, V) = softmax(d_k^{-\frac{1}{2}} QK^T)V$ called Scaled Dot-Product Attention (SDP-attention) [16]. The other one could be achieved by average over all the attention vectors, and then normalizing the resulting weight vector to sum up to 1, i.e., $Attention(Q, K, V) = softmax(avg(d_k^{-\frac{1}{2}} QK^T))V$ [7], denoted by Mean Pooling Scaled Dot-Product Attention (MPSDP-attention).
- **Random sampling:** We also randomly sampled half of posts from each user to train CNN and LSTM model.

4.3 Initialization and Hyperparameter

More difficult than [12], we did not apply emoji processing, stemming, irregular words processing and pretraining word2vec. The word embeddings and other parameters for all the deep learning models were initialized 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 use one layer of the LSTM to model the post text, and set the hidden neurons of LSTM to 200. The policy agent used a two fully connected layers with 100 and 20 units for each layer.

Our model could be trained end-to-end with backpropagation, and gradient-based optimization was performed using the Adam update rule [6], with a learning rate of 0.0001.

5 Results and Analysis

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

5.1 Method Comparison

For a fair comparison, we constructed the training and test set in the same way as reported in [12]. With 1,402 depressed users in total, we randomly selected

Table 2. Comparison of performance in terms of four selected measures. CNN/LSTM+RL refers to the proposed model.

Methods	Accuracy	Precision	Recall	F1
NB	0.724	0.727	0.728	0.728
MSNL [13]	0.818	0.818	0.818	0.818
WDL [11]	0.768	0.769	0.768	0.768
MDL [12]	0.848	0.848	0.850	0.849
CNN [19]	0.843	0.843	0.843	0.844
CNN + Random sampling	0.789	0.789	0.788	0.785
CNN + SDP-attention [16]	0.836	0.836	0.836	0.837
CNN + MPSDP-attention [7]	0.849	0.850	0.849	0.849
CNN + RL	**0.871**	0.871	**0.871**	**0.871**
LSTM	0.828	0.830	0.828	0.828
LSTM + Random sampling	0.760	0.760	0.757	0.756
LSTM + SDP-attention [16]	0.847	0.848	0.847	0.847
LSTM + MPSDP-attention [7]	0.850	0.850	0.850	0.850
LSTM + RL	0.870	**0.872**	0.870	**0.871**

1,402 non-depressed users on D_1 to make the scale of depressed users 50%, but in a more difficult manner by removing all the anchor tweets [2]. After obtaining the dataset, we trained and tested these methods using 5-fold cross validation.

We compared the depression detection performance of the proposed model with the baselines in terms of the four selected measures, i.e., Accuracy, Macro-averaged Precision, Macro-averaged Recall, and Macro-averaged F1-Measure. The comparison results are summarized in Table 2.

In the table, the first four lines list the results of the classic methods reported in [12], which use various features for training. MDL achieved the previous state-of-the-art performance with 0.849 in F1-Measure, indicating that combining the multimodal strategy and dictionary learning strategy is effective in depression detection.

The remaining part of the table lists the results of the neural network methods, which only use the users' posts for training. From the results, we can see that just using the posts, the CNN and LSTM model can achieve the accuracies of greater than 82.8%, which indicates the post text contains valuable information and is reasonable to use for depression detection. To give more attention to depression indicator posts, we evaluated two different self-attention models on the dataset. The results of the attention models showed that post level attention is effective in depression detection in most cases. The performance of the MPSDP-attention model is better than that of the SDP-attention model. We can see that the MPSDP-attention model may be more suitable for a task of this nature. Because there are an average of 396.6 tweets per user in the dataset, attention mechanism may be hard to effectively model the users to obtain an

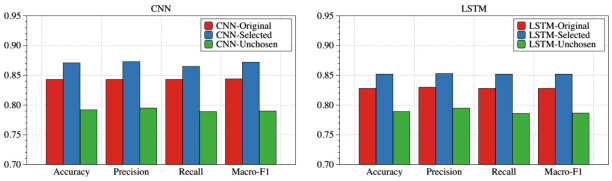

Fig. 3. Comparison between the models trained on original posts, selected posts, and unchosen posts.

obvious improvement. The LSTM + random sampling model has the lowest performance, which shows that the important effect of the post selection strategy. The poor selection strategy may be harmful the model. If we used the RL model to select posts, the CNN/LSTM + RL methods achieve the best performance, with a value of more than 87% for the F1-measure compared with both the CNN-based model and LSTM-based model, indicating that the RL post selection strategy was the most effective in depression detection. Next, we will show why our RL selection strategy was more effective than other methods.

5.2 Utility of Selected Posts

In order to verify the effectiveness of the method, we compared the baseline models trained on the original dataset, selected dataset, and unchosen dataset. We first trained the policy gradient agent to provide depression indicator posts and unchosen posts from the original dataset. Then, these indicator posts and unchosen posts made up the selected dataset and unchosen dataset, respectively. We compared the baseline models with three settings. One setting was training the model on the original dataset, which was denoted as model-original. The other settings denoted as model-selected and model-unchosen were training on the selected dataset and unchosen dataset, respectively.

The comparisons are shown in Fig. 3. From the results, we can observe that both of the models could benefit from the selected posts. The baseline models trained on selected dataset can achieve almost 2.4% better than those on original dataset. The error reduction rate was more than 9%. Inevitably, the unchosen dataset achieves poor performance. The results also indicate that the agent can select depression indicator posts that are more beneficial for depression classification.

5.3 Robustness Analysis in Realistic Scenarios

For a fair comparison, in the Table 2, we constructed the training/test set the same as [12]'s setting, where they made a balanced data set, and made sure 50% of the data contained depressed users. This does not seem a realistic scenario, as the real world data set may only contain a small number of depressed users.

With 1,402 depressed users in total, we fixed the capacity of our dataset to 1,500 and varied the scale of depressed users from 10% to 90% with increment

of 10%. Figure 4 shows the trend of detection performance with different proportions of depressed users. It can be found that our method achieved a stable and outstanding performance even though there is only a very low proportion of users with depression. However, when the depression users' scale does not laid at 50%, we retrieved a seriously decent performance of MDL under imbalanced scales. Therefore, our method is more instructive in detecting the depression than MDL in the realistic scenario.

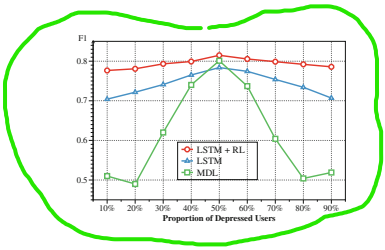

Fig. 4. Comparison between the models trained on the datasets with different scales of depressed users. The total number of users is 1,500.

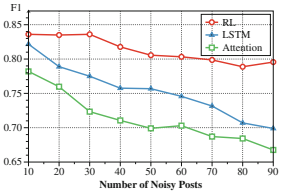

Fig. 5. Effect of different number of noisy data. The average post number of one user is 396.6.

5.4 Analysis of Noisy Data

Because social media are full of noisy data, we also evaluated all the models in a situation where different number of noisy posts were inserted in the dataset. We randomly selected 252,360 posts from the depression-candidate dataset D_2, and added from 10 to 90 posts to each user. We wanted to verify if the RL model can select indicator posts from noisy data.

As shown in Fig. 5, the performance of the models decreased to various degrees. However, as the number of posts increased, the advantage of the proposed model became increasingly obvious, and the RL model remarkably outperformed the other models. The other models suffered more form noisy posts. Espectively, at the 90 point, our model outperforms attention-based model over 13% in F1 score. The results indicated that our proposed model could obtain better performance when encountering the noisy data.

6 Conclusion

In this study, we investigated the problem of detecting depression based on the content users posted on social media, and verified the feasibility of using only the contextual information to detect depression. To overcome the problem of discrete selection, we proposed a reinforcement learning-based method to select indicator posts and remove other posts. Experimental results demonstrated that the proposed method could achieve better performance than previous methods. Through several experiments, we found that other detection models could benefit from the newly selected dataset. The further experiments demonstrated that our model could obtain a strong and stable performance in realistic scenarios.

References

1. Reinforcement learning for relation classification from noisy data. In: AAAI (2018)
2. Coppersmith, G., Dredze, M., Harman, C.: Quantifying mental health signals in Twitter. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 51–60 (2014)
3. De Choudhury, M., Counts, S., Horvitz, E.: Social media as a measurement tool of depression in populations. In: Proceedings of the 5th Annual ACM Web Science Conference, pp. 47–56. ACM (2013)
4. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural Comput. **9**(8), 1735–1780 (1997)
5. Kim, Y.: Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408.5882 (2014)
6. Kingma, D., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
7. Liu, Y., Sun, C., Lin, L., Wang, X.: Learning natural language inference using bidirectional LSTM model and inner-attention. arXiv preprint arXiv:1605.09090 (2016)
8. Park, M., Cha, C., Cha, M.: Depressive moods of users portrayed in Twitter. In: Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD), vol. 2012, pp. 1–8. ACM, New York (2012)
9. Park, M., McDonald, D.W., Cha, M.: Perception differences between the depressed and non-depressed users in twitter. In: ICWSM, vol. 9, pp. 217–226 (2013)
10. Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd-Graber, J.: Beyond LDA: exploring supervised topic modeling for depression-related language in Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, pp. 99–107 (2015)
11. Rolet, A., Cuturi, M., Peyré, G.: Fast dictionary learning with a smoothed wasserstein loss. In: Artificial Intelligence and Statistics, pp. 630–638 (2016)
12. Shen, G., et al.: Depression detection via harvesting social media: a multimodal dictionary learning solution. In: IJCAI, pp. 3838–3844 (2017)
13. Song, X., Nie, L., Zhang, L., Akbari, M., Chua, T.S.: Multiple social network learning and its application in volunteerism tendency prediction. In: SIGIR, pp. 213–222. ACM (2015)
14. Srivastava, N., Hinton, G.E., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(1), 1929–1958 (2014)
15. Suhara, Y., Xu, Y., Pentland, A.: DeepMood: forecasting depressed mood based on self-reported histories via recurrent neural networks. In: WWW, pp. 715–724. International World Wide Web Conferences Steering Committee (2017)
16. Vaswani, A., et al.: Attention is all you need. arXiv preprint arXiv:1706.03762 (2017)
17. Williams, R.J.: Simple statistical gradient-following algorithms for connectionist reinforcement learning. Mach. Learn. **8**(3–4), 229–256 (1992)
18. Xu, R., Zhang, Q.: Understanding online health groups for depression: social network and linguistic perspectives. J. Med. Internet Res. **18**(3), e63 (2016)
19. Yates, A., Cohan, A., Goharian, N.: Depression and self-harm risk assessment in online forums. arXiv preprint arXiv:1709.01848 (2017)