Feature Attention Network: Interpretable Depression Detection from Social Media

Hoyun Song* **Jinseon You*** **Jin-Woo Chung** **Jong C. Park†**
School of Computing
Korea Advanced Institute of Science and Technology
{hysong, jsyou, jwchung, park}@nlp.kaist.ac.kr

Abstract

Although depression is one of the most common mental disorders, the depressed individuals may not be aware of their symptoms at all so that they sometimes miss the appropriate time for treatment. In order to prevent this problem, many researchers looked into social media to figure out depressed individuals by analyzing the differences in language use. While they have recently achieved reasonable performance in detecting depression, especially using deep learning methods, such methods still do not provide a clear way to explain why certain individuals have been detected as depressed. To address this issue, we propose Feature Attention Network (FAN), inspired by the process of diagnosing depression by an expert who has background knowledge about depression. We evaluate the performance of our model on a large scale general forum (Reddit Self-reported Depression Diagnosis) dataset. Experimental results demonstrate that FAN shows good performance with high interpretability despite a smaller number of posts in training data. We investigate different aspects of posts by depressed users through four feature networks built upon psychological studies, which will help researchers to investigate social media posts to find useful evidence for depressive symptoms.

1 Introduction

Depression is one of the most common mental disorders which not only hampers an individual's daily function, but also causes a variety of serious social problems such as suicide. The depressed individuals may not be aware of their depressed symptoms at all so that they sometimes miss the appropriate time for taking care of depression. To prevent this problem, many researchers used social media to identify depressed users by analyzing the differences in language use.

The researchers reported good performance in detecting depression by analyzing a variety of features such as lexical features including symptom lexicons (De Choudhury et al., 2013a; De Choudhury et al., 2013b; Coppersmith et al., 2014), syntactic features (Nambisan et al., 2015; Gkotsis et al., 2016), sentiment analysis (Wang et al., 2013; Preoţiuc-Pietro et al., 2015), or topic modeling (Resnik et al., 2015; Preoţiuc-Pietro et al., 2015). However, these approaches are mainly based on labor-intensive engineering with handcrafted features, which show limited performance for classification. To further enhance the performance of depression detection, researches have recently proposed deep learning methods (Kshirsagar et al., 2017; Yates et al., 2017; Orabi et al., 2018). However, despite the improved performance, they still do not present a clear way to explain why certain individuals have been detected as depressed, which is critical for a follow-up study on specific diagnosis and prevention.

We propose a model that can interpret the process of classifying social media users when it derives the result of depression detection based on their posts. For such an interpretable model, we assume that the initial setup of a model imitates the process of diagnosing depression by a clinician who has the neces-

* Equal contribution (alphabetically ordered)
† Corresponding author

sary background knowledge about depression. The intuition behind this model is two-fold. First, depressive symptoms are not always clearly exhibited in social media, but may be manifested through a small number of posts, even for highly depressed individuals. To address this issue, we used a post-level attention, which finds salient posts that play an important role in detecting depressed individuals from social media. Second, if the model is built upon a psychological theory for depression which accounts for how depressed individuals use social media, it would be possible to explain relevant factors involved in the detection process by associating it to an established theory. To implement such an explainable model, we propose Feature Attention Network (FAN), which effectively incorporates evidence from psychological theories.

In this paper, we evaluate the performance of our model on a large scale general forum dataset presented by Yates et al. (2017). We carry out experiments on the dataset to validate our model with respect to the performance in detection and the interpretability of each extracted feature. We show that our proposed model achieves good performance with high interpretability in spite of a limited use of computing power. We investigate different aspects of posts by depressed users through four feature networks, which will help researchers to effectively screen social media posts to find useful evidence for depressive symptoms with respect to psychological theories.

2 Related Work

2.1 Depression Detection in Social Media

Language in social media activities is known to represent the current state of writers including their mental health. By analyzing the language use in social media, many researchers have discovered a way to identify depressed individuals, where they focused on differences in word usage between depressed user groups and control groups (De Choudhury et al., 2013a; De Choudhury et al., 2013b; Coppersmith et al., 2014) or on syntactic differences such as the number of part-of-speech types or the depth of parse trees (Nambisan et al., 2015; Gkotsis et al., 2016; Al-Mosaiwi and Johnstone, 2018). Some studies attempted to predict depression by comparing the differences between subjects with depression groups and control groups (Resnik et al., 2015; Shen et al., 2017), or to use sentiment analysis techniques based on the assumption that depressed individuals reveal negative emotions much more frequently than those without depression (Wang et al., 2013; Zucco et al., 2017; Shen et al., 2017).

However, previous work has some limitations. In particular, most studies relying on feature engineering conducted their experiments only on small datasets, making it difficult to achieve high classification performance. Only a few studies adopted a neural network approach (Yates et al., 2017; Orabi et al., 2018), achieving reasonable performance. However, most previous studies could not explain the detection results adequately with respect to relevant theories in the field, making it difficult to conduct a more detailed analysis for further processes such as diagnosis and prevention.

2.2 Interpretable Research

Various studies have been conducted to explain the classification results from neural network to analyze relevant factors behind its performance or to further improve it. Many studies in vision have used neural visualization with the representation learned from subsequent layers to generate human interpretable information (Simonyan et al., 2014; Zeiler and Fergus, 2014; Girshick et al., 2014). Some studies have applied interpretable methods to NLP, focusing mainly on interpreting vector-based models for several tasks (Murphy et al., 2012; Li et al., 2016; Palangi et al., 2018). For example, Palangi et al. (2018) interpreted lexical-semantic meanings and grammatical roles for each word by interpreting learned internal representation, which is in contrast to looking at the input patterns to interpret activated internal neurons. However, interpreting a result by analyzing attention or neurons has a limitation that it cannot provide a rich explanation.

Kshirsagar et al. (2017) attempted to generate explanations about detected results for suicidal posts by using representation learning, but they conducted the attention mechanism only on the words in a post, with accompanying limitations.

We applied the attention mechanism to the posts of a user, which is quite challenging because the proportion of posts with depression indicators is not

high enough, so that most of the posts do not contain information useful enough for depression detection. By leveraging the concepts mentioned above, we interpret the feature representations related to various depression factors learned from feature networks, so as to understand which features are activated significantly during depression detection.

3 RSDD: Reddit Self-reported Depression Diagnosis

We use the Reddit Self-reported Depression Diagnosis (RSDD) dataset provided by Yates et al. (2017) to train and evaluate our model. It contains posts of Reddit users who claim that they have been diagnosed with depression, and categorizes them into a depression group. In addition, three annotators manually validated whether such categorized posts actually contain the claims that the users have been diagnosed with depression. Also, users who were randomly selected and whose posts do not contain any depression-related keywords are categorized into a control group. The RSDD dataset contains 9,210 users in the depression group and 107,274 users in the control group, with an average of 969 posts for each user and a median of 646. The dataset is divided into three categories (training, validation, and testing), each of which contains 3,070 diagnosed users and almost 35,000 control users. The task is to classify users into one of the two groups, given a collection of posts they created.

4 Method

The overall architecture of Feature Attention Network is shown in Figure 1. It contains four feature networks, each of which analyzes posts based on an established theory related to depression and a post-level attention on top of the networks. We describe the details of each network and the post-level attention in the following sections with the intuition behind why they are designed as such and how they are implemented.

4.1 Feature Networks

The feature networks are inspired by the depression detection process performed by domain experts with background knowledge about depression, where they examine social media posts of users

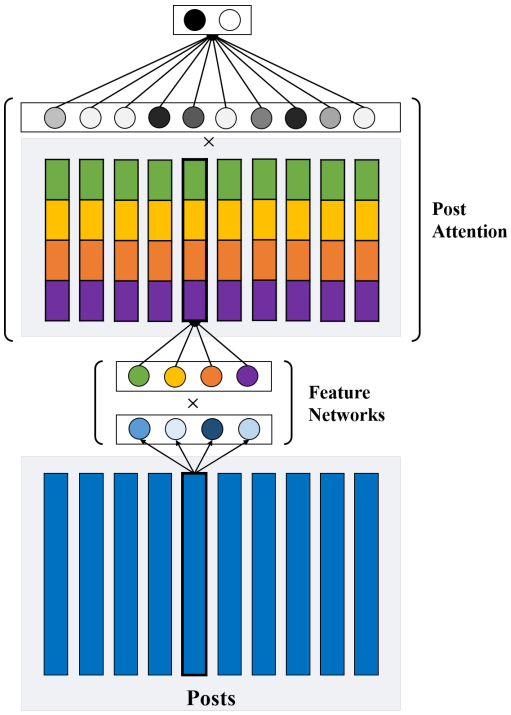

Figure 1: An overview of Feature Attention Network (FAN). Each colored circle in the feature networks indicates a different feature.

with potential depression to find relevant indicators based on domain knowledge. With this motivation, we used four types of strong indicators of depression taken from psychological studies, and developed four neural networks that are tailored for each indicator. In the following descriptions, a simple symbol (ex. p) denotes a non-vector, a bold-faced lower-case symbol (ex. $\mathbf{p}$) denotes a vector, and a bold-faced upper-case symbol (ex. $\mathbf{P}$) denotes a sequence of vectors, namely a matrix.

- **Depressive Symptoms (F1):** Posting comments directly related to a particular symptom is apparently the most distinct behavioral pattern of people with depression. Based on this observation, we propose a feature network for locating mentions related to depressive symptoms in posts. To identify which symptom is related to depression, we built a dictionary of evidence keywords taken from a lexicon of nine groups of depressive symptoms in *Diagnostic and Statistical Manual of Mental Disorders (DSM-V)* (Whooley and Owen, 2014)

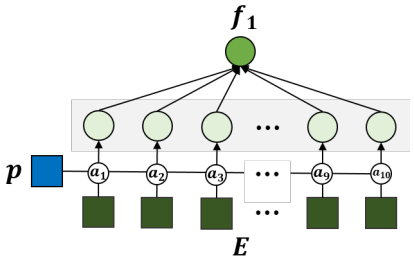

Figure 2: A schema of feature network for depressive symptom.

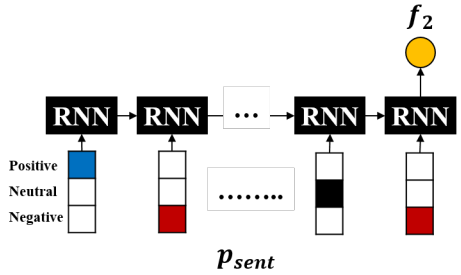

Figure 3: A schema of feature network for sentiment.

and from a list of the synonyms for antidepressants from a Wikipedia page on an antidepressant. Our dictionary contains 69 keywords about nine groups of symptoms and 11 antidepressant names.

To capture each piece of evidence from posts for the 9 symptoms and an antidepressant, we compute the similarity between a given post and tokens of the dictionary. First, we transform the word vectors for each symptom category to a single vector by element-wise multiplication. As a result, we generate a symptom matrix containing representative vectors for each category. Then, we obtain a matrix showing the similarity between an encoding vector of posts and the matrix. Last, we project the matrix to the feature vector ($\mathbf{f}_1$) through Multi-Layer Perceptron (MLP) (Figure 2).

$$\mathbf{a}_i = \mathbf{pWE}_i \quad (i = 1, ..., 10)$$
$$\mathbf{A} = softmax([\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_{10}])$$
$$\mathbf{f}_1 = tanh(f(\mathbf{A}))$$

- **Sentiments (F2):** According to the cognitive theory (Beck, 1983), depressed people tend to show negative thinking patterns and negative emotions. Hence, we assume that depressed people on social media also tend to express negative polarity on their posts more frequently than others.

  Based on the assumption, we propose the second feature network that captures such a tendency by considering sentiments in posts. To this end, we calculate sentiment scores of each word by using SentiWordNet (Baccianella et al., 2010). Based on the scores, we convert all the words in a post into one of the three categories (positive, neutral, and negative) produced by SentiWordNet, and then encode the one-hot vectors to a feature vector ($\mathbf{f}_2$) using Recurrent Neural Network (RNN) (Figure 3).

$$\mathbf{f}_2 = RNN(\mathbf{p}_{sent})$$

- **Ruminative Thinking (F3):** It is known that the ruminative response style can be expressed as repetitive thoughts and behavior (Nolen-Hoeksema, 1991). People with depression tend to express their feelings or negative experiences repeatedly so that sentences in relevant topics may also repeatedly appear on their posts.

  Based on this theory, we implement a network identifying how frequently certain stories of relevant topics are repeated. To measure the degree of relevance between a given post and others, we calculate two vectors using dot production and obtain the degree of relevance for all posts. Then, we project the degree to a feature vector ($\mathbf{f}_3$) through MLP (Figure 4).

$$\mathbf{a} = softmax(\mathbf{p} \cdot \mathbf{P})$$
$$\mathbf{f}_3 = tanh(f(\mathbf{a}))$$

- **Writing Style (F4):** According to some research, people with depression exhibit differences with respect to linguistic styles such as the distribution of nouns, verbs and adverbs and the complexity of sentences, which are conceptualized unconsciously (Gkotsis et al., 2016).

  The fourth feature network is based on the study above and is designed to capture different writing styles. We focus on the order of

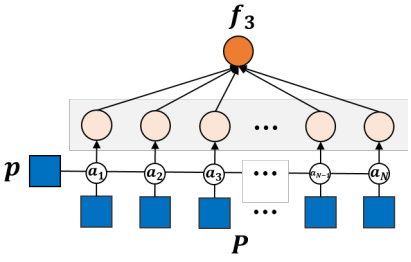

Figure 4: A schema of feature network for ruminative thinking.

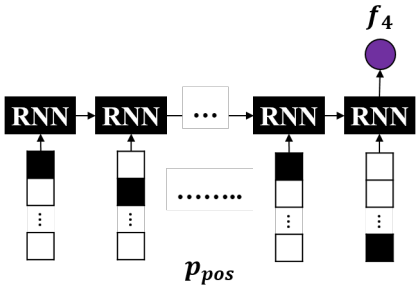

Figure 5: A schema of feature network for writing style.

words among other styles and use the distribution of part-of-speech tags. Thus, we input a sequence of part-of-speech tags consisting of a post to the network. The network then converts the sequence to a one-hot vector with the number of part-of-speech dimensions and encodes the one-hot vector to a feature vector ($\mathbf{f}_4$) using RNN (Figure 5).

$$\mathbf{f}_4 = RNN(\mathbf{p}_{pos})$$

We note that it is necessary to consider feature weights prior to combining the feature networks because each post shows a different level of depressive characteristics. Thus, we generate a vector with weights indicating which feature is most representative in analyzing the post for classifying the user. Then, we multiply the weights of the feature networks and generate a post vector considering all the features ($\mathbf{p}'$) by merging the weighted feature vector with a vector by the summation of elements.

$$\mathbf{w} = softmax(f(\mathbf{p}))$$
$$\mathbf{p}' = \sum_i \mathbf{w}_i \cdot \mathbf{f}_i \quad (i = 0, ..., 4)$$

Since the weights indicate how much each feature contributes to classifying the post, helping to explain how and why depression is developed, we can interpret the behavior of the network by changing the weights.

4.2 Post-level Attention

We note that even an individual who has depression does not always express depressive moods through their social media posts. For this reason, when the experts predict depression, they focus on some particular posts which clearly show depressive characteristics distinct from other posts. Thus, we believe that it is necessary to appropriately select and process such posts according to their importance.

To this end, we apply attention mechanism to the posts in a way similar to the hierarchical attention method (Yang et al., 2016). We introduce a post level context vector ($\mathbf{v}$) and project the vector to the post vector ($\mathbf{p}'$) to measure the importance of the posts:

$$\mathbf{a} = softmax(\mathbf{p}'\mathbf{W}\mathbf{v})$$
$$\mathbf{o} = \sum_i \mathbf{a}_i \cdot \mathbf{p}'_i \quad (i = 1, ..., M)$$

where M is the number of posts. We project the output vector ($\mathbf{o}$) as classifying depression using MLP.

5 Experiment

5.1 Baselines

To simulate the diagnosis by experts and interpret a deep neural network, we built the feature networks. The approach is also based on the assumption that the merged vector representation generated by the neural network is similar to one of the feature networks. In order to see the difference between the conventional network and the feature network built upon human intuition, we build a baseline model by replacing the feature network of FAN with RNN. In other words, each post vector is generated by bi-directional RNN.

Method	Precision	Recall	F1
BoW†	0.44	0.31	0.36
BoW‡	0.72	0.29	0.42
Feature-rich†	0.69	0.32	0.44
Feature-rich‡	0.71	0.31	0.44
FastText	0.37	**0.70**	0.49
CNN-E	0.59	0.45	0.51
CNN-R	**0.75**	0.57	**0.65**
Baseline	0.52	0.58	0.54
FAN	0.61	0.52	0.56

Table 1: Results of evaluation on the test set. † and ‡ indicate MNB and SVM classifiers, respectively. While CNN-E uses 400 latest posts for each user as inputs, CNN-R uses 1,500 random posts.

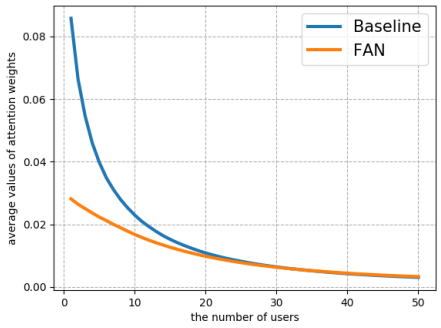

Figure 6: Change of average values of attention weights over the users.

As another baseline, we compare the models proposed by Yates et al. (2017). They present five models based on feature engineering and two models based on neural networks. The neural models encode posts to vectors using convolutional neural network (CNN) and then merge the post vectors to a single vector. By projecting the vector, the models classify users. We note that one of the networks showing the best performance is similar to our models with respect to the process of classifying the posts (i.e., processing each post, merging them, and classifying users).

5.2 Setup

We trained our model on the RSDD training set and optimized it on the RSDD validation set. We split each post into a sequence of tokens and performed part-of-speech tagging using Stanford CoreNLP (Manning et al., 2014). We discarded posts whose number of tokens is either smaller than 5 or bigger than 100. We then randomly selected 500 posts from the whole posts for each user and used them for training.

We used GloVe (Pennington et al., 2014) for embedding word vectors and GRU (Cho et al., 2014), one of the RNN variants, for encoding the sequence. In order to improve generalization, we used dropout and an L2 regularization. We set a learning rate and an L2 regularization rate to 0.001 and 0.0001, respectively. We separately set a dropout rate according to the models, where we set it 0.3 and 0.2 for the baseline and our model, respectively. In building the vocabulary, we only retained words that appear more than 5 times and replaced the others with <UNK> tokens. We used the Adam optimizer (Kingma and Ba, 2014) and weighted loss for imbalance between depressed and control groups.

5.3 Result and analysis

Table 1 shows the performance of identifying depressed users on the RSDD test set. It is shown that our model outperforms all the other models except for CNN-R, a state-of-the-art neural model. One possible reason for this is the difference in the amount of posts in training data; we used only 500 posts for each user as training data, which are three times smaller than the data used by CNN-R (i.e., 1,500 posts) due to the limitation of the computing power in our experimental environment. However, our model performs better than CNN-E which is essentially the same as CNN-R except that it uses only 400 posts for training, only 100 fewer than ours (i.e., 500 posts). This may be due to the attention mechanism we used, which is an advanced technique that helps a model to focus more on important posts for classification. As a result, our model shows good performance despite relying on much fewer posts than the state-of-the-art model, with high interpretability to be discussed in Section 6. We expect that more posts in training data will further boost the performance of our model, as seen in the positive correlation between the number of posts and performance (Yates et al., 2017).

Class	F1	F2	F3	F4
TP-High	0.33	0.43	0.13	0.11
TP-Low	0.86	0.01	0.08	0.05
TN-High	0.42	0.24	0.19	0.15
TN-Low	0.84	0.02	0.09	0.06

Table 2: The average feature weights on each class. TP and TN indicate true-positive and true-negative, respectively. High and Low indicate the highest and the lowest attention posts, respectively.

As shown in Table 1, FAN and our baseline model show similar F1 scores, but with different performance balance: While FAN shows a higher precision than a recall, the baseline shows a reversed tendency. In order to find the reason for this, we examined post-level attention weights from both models to analyze interpretable and important factors for classifying users. From each of the nearly 1,000 depressed users who are classified as depressed by both models, we selected the top 50 posts (10%) with the highest attention weights obtained for each model. We found that only an average of 16 posts are the same for both the baseline and FAN. This means that, on a majority of posts, two models produce different attention weights, leading to different detection results and performance.

Furthermore, to analyze the change of the effect of posts with high attention produced by the two models, we present attention weights of top 50 posts in the order of weights, which are averaged over almost 1,000 users (Figure 6). Interestingly, the highest attention weight of the baseline is marginally bigger than that of FAN. From this observation, we conjecture that the baseline classifies users based only on a small number of posts with high attention weights, while FAN considers many posts more evenly when classifying users. This means that, if such a small number of posts are contaminated, the probability that the baseline misclassifies becomes higher. In contrast, FAN is less likely to fall into this danger, but is also likely to ignore a few important posts. This also suggests why the two models show different precision and recall scores.

Word	Frequency	Polarity
anxiety	2233	neg
meds	1229	neu
medication	934	pos
disorder	698	neg
psychiatrist	382	neu
adderall	320	pos
suicidal	316	neg
disability	145	neg
abusive	136	neg
insomnia	131	neg

Table 3: A set of words that appear only in the high F2 weighted group compared to the low F2 weighted group and their polarities derived by SentiWordNet.

6 Interpretation

Using FAN, we can interpret the detection results by analyzing learned representations.[1] We selected a set of almost 1,500 users who are detected as a true-positive for depression detection. Then we sampled the top 100 posts with the highest attention and bottom 100 posts with the lowest attention from each user. We also selected a set of almost 63,000 users who are detected as a true-negative and sampled both the top and bottom 100 posts from each user in the same way. Table 2 shows average feature weights for each of the four classes.

We looked into the instances in each class in Table 2 to validate that FAN yields interpretable results enough to meet our expectations. We observed two notable trends in the results. First, as for the second feature weight (F2), we found that the higher this weight is in a post, the higher attention scores the post shows. This suggests that sentiment information plays an important role in detecting depressed users. Table 3 shows the most frequent words with their polarity found in a group of posts with high F2 weights. For example, the word 'anxiety', which has negative polarity, does not appear in the group of low F2 weighted posts from users in TP-High and TP-Low classes. In contrast, in the group of the high F2 weighted posts, it appears 2,233 times in 29,360 posts, which means that 7.6% of the high F2

[1] In this paper, we do not present any post-level examples from data because of possible privacy and ethical issues.

F1	F3	Phrase
0.58	0.16	I am closed
0.57	0.16	I thought
0.53	0.15	I feel like
0.38	0.17	I hate seeing people

Table 4: Example phrases from posts with high F1 and F3 weights.

Tag	TP-High	TP-Low
VB	1.83	1.67
VBD	1.04	0.95
VBG	0.83	0.76
VBN	0.59	0.54
VBP	1.76	1.61
VBZ	1.18	1.08

Table 5: The proportion of verb tags to the number of posts in each class.

weighted posts contain this word. Considering the most frequent general words such as 'like' (appearing 6,181 times, 21.1%) and 'would' (5,181 times, 17.6%), the frequency of 'anxiety' is relatively high in this group of posts. In addition, the table shows that most of the top-ranked frequent words in the high F2 weighted posts have negative polarity, creating negative and depressive mood in a post. This means that the feature network based on sentiment information plays a significant role in distinguishing depressive posts from others in depressed users.

Second, there is a trend regarding to F1 (depressive symptoms) and F3 (ruminative thinking). The posts with high F1 weights are mostly distributed in the TP-Low class. Moreover, we found that the words in Table 3 also appear frequently in the low F1 weighted posts. We look into the reason why the depressive symptoms are negatively correlated to the attention. If the F1 weight of a post is low, most posts do not seem to be related to depressive mentions, because the keywords we selected for depressive symptoms rarely appear throughout the entire posts. In other words, a post without features we established shows biased feature weights toward F1.

However, some of the posts in the TP-High class show high F1 and F3 weights. Table 4 shows example phrases taken from them. We see that many of the posts with such phrases are related to so-called "self-attention", where users repeatedly mention how they feel or what they go through. For further analysis of this trend, we compared the frequency of 'I' in all posts from two classes (TP-High and TP-Low) and in some posts with F1 weights higher than 0.50 and F3 weights higher than 0.15. The average occurrence of 'I' is 1.25 in all posts, and 1.35 in the high F1 and F3 weighted posts. The proportion of posts with high F1 and F5 weights ($F1 > 0.50, F3 > 0.15$) is 14.8% in the TP-High class while it is 0.4% in the TP-Low class. This confirms that people with mental health problems show high self-attentional disposition (De Choudhury et al., 2016).

Compared to other feature networks, F4 (writing style) has little effect on detecting depression. However, we found that when the F4 weight increases, the number of verb phrases is also increased. As shown in Table 5, when the frequency of verbs is increased, attention is also increased accordingly. This suggests that people with mental health problems show different sentence complexity with respect to their language (Gkotsis et al., 2016).

7 Conclusion

We proposed Feature Attention Network which simulates the process of detecting depressed people from social media texts by domain experts who have background knowledge about depression. Through this process our model can focus more on depression-relevant sentences by post-level attention, which fits well into the real-world situation where only a few posts are relevant to depression even for depressed users. It also enables interpretation of why a particular post is relevant to depression in terms of features taken from psychological studies, which is important for further clinical analysis of depressive symptoms. However, our model uses smaller training data as input due to limited computing power, showing lower performance than the state-of-the-art model that uses three times more data than ours and is trained in a less interpretable way. With more computing power, we anticipate that our model will show competitive performance against the state-of-the-art model. FAN currently consists only of the four features that are based on psychological studies on depression. Since

our model takes advantage of high-dimensional representations of neural networks and at the same time allows other high-level features to be readily incorporated, if we add other useful features to the model, it will be possible to obtain more reasonable and diverse explanations for different aspects of depression. If we can generate appropriate feature networks for other mental disorders (such as dementia, schizophrenia, and bipolar disorder), it will be possible to simulate the process of diagnosing them in a similar way.

Acknowledgments

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (No. NRF2017R1A2B4012788).

References

Al-Mosaiwi, M. and Johnstone, T. (2018). In an Absolute State: Elevated Use of Absolutist Words Is a Marker Specific to Anxiety, Depression, and Suicidal Ideation. *Clinical Psychological Science*, 6(4):529–542.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC)*, pages 2200–2204.

Beck, A. T. (1983). Cognitive therapy of depression: New perspectives. *Treatment of depression: Old controversies and new approaches, New York: Raven Press*, pages 265–284.

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734.

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 (CLPsych)*, pages 51–60.

De Choudhury, M., Counts, S., and Horvitz, E. (2013a). Social Media as a Measurement Tool of Depression in Populations. In *Proceedings of the 5th Annual ACM Web Science Conference (WebSci)*, pages 47–56. ACM.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013b). Predicting Depression via Social Media. *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 128–137.

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., and Kumar, M. (2016). Discovering Shifts to Suicidal Ideation from Mental Health Content in Social Media. In *Proceedings of the 2016 CHI conference on human factors in computing systems (CHI)*, pages 2098–2110. ACM.

Girshick, R., Donahue, J., Darrell, T., and Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 580–587.

Gkotsis, G., Oellrich, A., Hubbard, T., Dobson, R., Liakata, M., Velupillai, S., and Dutta, R. (2016). The language of mental health problems in social media. In *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 63–73.

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

Kshirsagar, R., Morris, R., and Bowman, S. (2017). Detecting and Explaining Crisis. *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 66–73.

Li, J., Chen, X., Hovy, E., and Jurafsky, D. (2016). Visualizing and Understanding Neural Models in NLP. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 681–691.

Manning, C., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 55–60.

Murphy, B., Talukdar, P., and Mitchell, T. (2012). Learning Effective and Interpretable Semantic Models Using Non-Negative Sparse Embedding. In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, pages 1933–1950.

Nambisan, P., Luo, Z., Kapoor, A., Patrick, T. B., and Cisler, R. A. (2015). Social Media, Big Data, and Public Health Informatics: Ruminating behavior of depression revealed through Twitter. In *Proceedings of the 48th Hawaii International Conference on System Sciences (HICSS)*, pages 2906–2913.

Nolen-Hoeksema, S. (1991). Responses to depression and their effects on the duration of depres-

sive episodes. *Journal of abnormal psychology*, 100(4):569.

Orabi, A. H., Buddhitha, P., Orabi, M. H., and Inkpen, D. (2018). Deep Learning for Depression Detection of Twitter Users. In *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology (CLPSych)*, pages 88–97.

Palangi, H., Smolensky, P., He, X., and Deng, L. (2018). Question-Answering with Grammatically-Interpretable Representations. In *Proceedings of the 32nd Association for the Advancement of Artificial Intelligence (AAAI)*, pages 5350–5357.

Pennington, J., Socher, R., and Manning, C. (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Preoţiuc-Pietro, D., Sap, M., Schwartz, H. A., and Ungar, L. (2015). Mental Illness Detection at the World Well-Being Project for the CLPsych 2015 Shared Task. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 40–45.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., and Boyd-Graber, J. (2015). Beyond LDA: Exploring Supervised Topic Modeling for Depression-Related Language in Twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych)*, pages 99–107.

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 *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3838–3844.

Simonyan, K., Vedaldi, A., and Zisserman, A. (2014). Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. *arXiv preprint arXiv:1312.6034*.

Wang, X., Zhang, C., Ji, Y., Sun, L., Wu, L., and Bao, Z. (2013). A Depression Detection Model Based on Sentiment Analysis in Micro-blog Social Network. In *Proceedings of the 17th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 201–213. Springer.

Whooley and Owen (2014). *Diagnostic and Statistical Manual of Mental Disorders (DSM-5®)*. John Wiley & Sons, Ltd.

Yang, Z., Yang, D., Dyer, C., He, X., Smola, A., and Hovy, E. (2016). Hierarchical Attention Networks for Document Classification. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1480–1489.

Yates, A., Cohan, A., and Goharian, N. (2017). Depression and Self-Harm Risk Assessment in Online Forums. *arXiv preprint arXiv:1709.01848*.

Zeiler, M. D. and Fergus, R. (2014). Visualizing and Understanding Convolutional Networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 818–833. Springer.

Zucco, C., Calabrese, B., and Cannataro, M. (2017). Sentiment Analysis and Affective Computing for depression monitoring. In *Proceedings of the International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1988–1995.