Towards Explainability in Using Deep Learning for the Detection of Anorexia in Social Media

Hessam Amini(✉) and Leila Kosseim

Computational Linguistics at Concordia (CLaC) Laboratory,
Department of Computer Science and Software Engineering,
Concordia University, Montréal, QC H3G 2W1, Canada
{hessam.amini,leila.kosseim}@concordia.ca

Abstract. Explainability of deep learning models has become increasingly important as neural-based approaches are now prevalent in natural language processing. Explainability is particularly important when dealing with a sensitive domain application such as clinical psychology. This paper focuses on the quantitative assessment of user-level attention mechanism in the task of detecting signs of anorexia in social media users from their posts. The assessment is done through monitoring the performance measures of a neural classifier, with and without user-level attention, when only a limited number of highly-weighted posts are provided. Results show that the weights assigned by the user-level attention strongly correlate with the amount of information that posts provide in showing if their author is at risk of anorexia or not, and hence can be used to explain the decision of the neural classifier.

Keywords: Explainability · Deep learning · Attention mechanism · Anorexia · Social media

1 Introduction

Social media is a rich source of information for the assessment of mental health, as its users often feel they can express their thoughts and emotions more freely, and describe their everyday lives [12]. This is why the use of natural language processing (NLP) techniques to extract information about the mental health of social media users has become an important research question in the last few years [18,27].

One of the main challenges of developing tools for the automatic detection of mental health issues from social media is providing justification for the decisions. Mental health issues are still often stigmatised and labelling a user as a victim of a mental health illness without a proper justification is not socially responsible. As a result, to be applicable in a real-life setting, automatic systems should not only be accurate, but their decisions need to be explained.

In the past decade, deep learning algorithms have become the state of the art in many NLP applications. By automatically learning the representation of useful

E. Métais et al. (Eds.): NLDB 2020, LNCS 12089, pp. 225–235, 2020.
https://doi.org/10.1007/978-3-030-51310-8_21

linguistic features for the tasks they are performing, deep learning approaches have lead to impressive improvements in most NLP tasks [4,7]. This also applies to the domain of NLP for mental health assessment, where recent deep learning models have led to state-of-the-art results in the field [19,21,22]. However, despite achieving high performance, one of the most important drawbacks of these models is their *black box* nature, where the reasoning behind their decision is difficult to interpret and explain to the end users. This constitutes a serious setback to their adoption by health professionals [10].

The focus of this paper is to assess the usefulness of user-level attention mechanism [22] as a means to help explain neural classifiers in mental health. Although the experiments were performed on the detection of anorexia in social media, the methdology is not domain-dependent, hence can be applied to other tasks involved in the detection of mental health issues of social media users, based on their online posts.

The paper is organized as follows: Sect. 2 explains the two levels where the attention mechanism can be used (i.e. intra-document and inter-document), and describes the related work in validating explainability using attention mechanism. Section 3 explains our experiments to validate the interpretability of user-level attention, whose results are then presented in Sect. 4. Section 5 provides additional observations in terms of how the attention mechanism has worked. Finally, Sect. 6 concludes the paper and provides future directions for the current work.

2 Related Work

Attention mechanism [1] has become an essential part of many deep learning architectures used in NLP, as it allows the model to learn which segments of text should be focused on to arrive at a more accurate decision. In text classification applications, such as the detection of mental health issues, attention mechanisms can be applied both at the intra and the inter-document levels [20].

At the intra-document level, the attention mechanism learns to find informative segments of each document, and assigns higher weights to these segments when creating a representation of the whole document. The success of the intra-document attention mechanism has made it an essential part of transformers [25], which are now the building block of several powerful NLP models, such as BERT [5].

On the other hand, the inter-document attention mechanism tries to identify entire documents that are more informative from a collection, and assign higher weights to these when computing the representation of the whole collection. The inter-document attention mechanism is generally used when the classification pertains to the entire collection, as opposed to individual documents. Previous work in NLP for clinical psychology has typically used this type of attention mechanism to create a representation of social media users: a collection of online posts from each user is fed to the model and the inter-document attention (also referred to as *user-level* attention) creates a representation of the user through a weighted average of the representations of their online posts, with the most informative posts are assigned higher weights. While Mohammadi et al. [21] and Matero et al. [19] have used inter-document attention for the task of suicide risk assessment, Maupome et al. [20] and Mohammadi et al. [22] have utilized it for the detection of depression and anorexia, respectively.

To explicitly provide explainability in deep NLP models, several methods have been proposed. Wang et al. [26], Lee et al. [11], Lin et al. [13], and Ghaeini et al. [6] have used attention visualization based on attention heat maps. These heat maps graphically show which parts of the texts have been given higher or lower attention weights.

In NLP for clinical psychology, the data is usually sensitive and standard attention visualization are not ideal. Hence, other methods have been developed to show the validity of the attention explainability. For example, Ive et al. [8] provided paraphrased sentences from the dataset, alongside their assigned attention weights.

Jain and Wallace [9] and Serrano and Smith [24] proposed quantitative approaches to validate the explainability of intra-document attention mechanism. While Jain and Wallace's method was focused on randomly shuffling, and also generating adversarial attention weights [9], Serrano and Smith analyzed attention explainability by zeroing out the attention weights.

In this paper, we propose a quantitative approach, specifically focused on the user-level (inter-document) attention mechanism in a binary classification task of detection of a specific mental health issue, anorexia.

3 Experiments

Our approach is based on monitoring the performance measures of a neural classifier, with and without user-level attention, when only a limited number of highly-weighted posts are provided.

The neural classifier used is the *CNN-ELMo* model from Mohammadi et al. [22]. This model was chosen because it achieved comparable results to the best performing model at the recent eRisk shared task [17,22], and is based on an end-to-end architecture, which makes the reasoning behind its decision more easily explainable.

The trained model was first run on the testing data, and for each user, her/his posts were ranked from the highest attention weights to the lowest. We then ran the following two experiments:

1) We tested the model by feeding it only the n top-weighted posts by each user. We gradually increased values of n from 1 to 1000, and monitored the performance of the system as n changes. The purpose of this experiment was to compare the performance of the model when all the posts are available, with when only the top-ranking posts (based on the attention weights) are available to the system.
2) We replaced the user-level attention with a simple average pooling and re-ran experiment 1. The aim of this experiment was to evaluate the contribution of the user-level attention by ablating it from the model.

3.1 Model Architecture

The architecture of the *CNN-ELMo* model is shown in Fig. 1. For each user, her/his posts are first tokenized and then fed to an embedder, to extract a dense representation for each token. For the embedder, the original 1024d version of ELMo [23], pretrained on the 1 Billion Word Language Model Benchmark [2] was used.

For each post, 300 unigram and 50 bigram convolution filters were applied on the token embeddings. The output of the convolution filters were then fed to a Concatenated Rectified Linear Unit (CReLU), and max pooling was applied to the output of the CReLUs. The output of the two max pooling layers were then concatenated and used as the representation for each post.

The final user representation of a user was calculated by averaging (experiment 2) or weighted averaging (experiment 1) the representations of the available posts by that user. In order to calculate the weights, a single fully connected layer was applied to the representation of each post, mapping the post representation to a scalar. A softmax activation function was then applied over the scalars, which resulted in the weights corresponding to each post.

The last layer of the model was comprised of a single fully-connected layer, mapping the user representation to a vector of size two. Finally, by applying a softmax activation function over this vector, the probability for each user belonging to the anorexic/non-anorexic class was calculated.

3.2 Dataset

The dataset used is from the first sub-task of the eRisk 2019 shared task [17], whose focus is the early risk detection of anorexia. The dataset consists of a collection of posts from the Reddit social media, and is annotated at the user-level, indicating whether a user is anorexic or not. For this work, we have focused on the detection of anorexia, without considering the earliness of the detection as the shared task does.

Table 1 shows statistics of the training, validation, and testing datasets. As the table shows, the data contains posts from 152 users for training, 320 users for validation, and 815 users for testing, with an average of 300 to 400 posts per user.

As indicated in Losada et al. [17], the dataset was collected following the extraction and annotation method, proposed by Coppersmith et al. [3]. The anorexic users were self-identified by explicitly stating being diagnosed with anorexia on Reddit, while the non-anorexic users were randomly crawled from the same social media. From the set of anorexic users, these specific posts which discussed being diagnosed with anorexia were removed from the dataset.

4 Results

The results from the experiments are shown graphically in Fig. 2, and selected results are provided in Table 2.

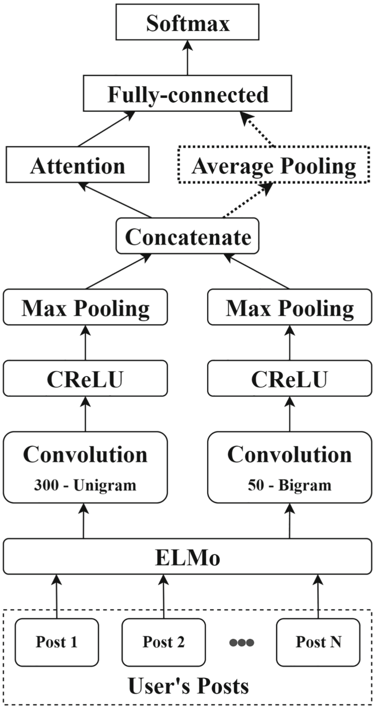


Fig. 1. The architecture of the model used for the experiment. In the ablated model, the *Attention* is replaced by the *Average Pooling* (shown in the dotted box).

As the solid lines in Fig. 2 show, by increasing the maximum number of available posts per user, the performance of the model with user-level attention (experiment 1) generally improves in terms of accuracy, precision, and F1, while the recall drops. It can also be observed that, the changes in performance measures decreases as the number of available posts increases, and the performance gradually converges to the final ones when all the posts are available (see Table 2). We believe that the gradual improvement in the precision and drop in recall is because, in general, the posts that have been highly weighted by the user-level attention mechanism, include signals that the user is anorexic (rather than signals that the user is not).

The dotted lines in Fig. 2a show that, by increasing the maximum number of available posts from 1 to 10, the performance of model with the user-level average pooling (experiment 2) also improves in terms of accuracy, precision, and F1, but deteriorates in terms of recall. This shows that, the first 10 highly-weighted posts included information necessary for the system to make a prediction about the user. This has even led the model with average pooling to have a higher F1 score than the model with user-level attention, as the former has a tendency to get less biased towards specific posts.

Table 1. Statistics of the dataset.

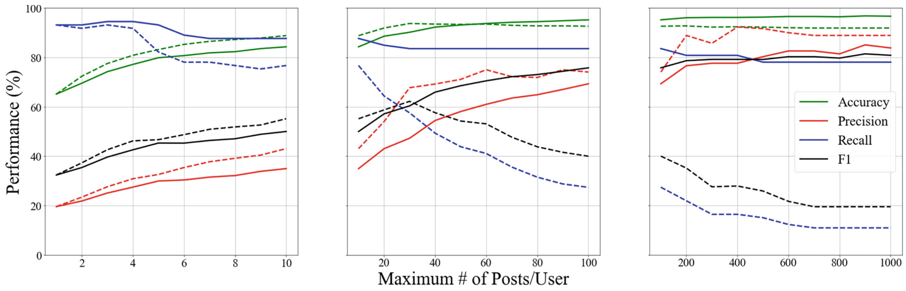
Dataset	# of users		# of posts per user			
	Anorexic	Non-anorexic	Min	Max	Ave	Med
Train	20	132	9	1999	558	330
Validation	41	279	9	1999	527	318
Test	73	742	10	2000	700	478

Table 2 and Fig. 2b show that, the F1 and the accuracy of the model with the user-level average pooling starts to drop from 30 and 60 posts, respectively. As a result, the model with user-level attention overtakes the one with average pooling in terms of F1 and accuracy, after more than 30 and 50 posts are available, respectively. This shows the higher capability of the model with the user-level attention over the other in handling the higher number of posts.

Figure 2 also shows that increasing the maximum number of available posts leads to a rapid drop in the recall of the model with user-level average pooling. This shows that, the higher the number of available posts to the model with average pooling, the more this model loses the capability on observing the patterns that are useful in detecting anorexia. This can also support the hypothesis that the user-level attention mechanism generally assigns higher weights to the posts that are more signalling of anorexia.

Table 2. Performance of the system (in percentage) in terms of the maximum number of highly-weighted posts from each user. The columns labelled as *with Avg Pool* refer to the model in which the user-level attention mechanism is ablated. The last row refers to the case when all the posts from each user are provided to the system.

Max # of posts/user	With attention (experiment 1)				With avg pool (experiment 2)			
	A	P	R	F1	A	P	R	F1
1	65.15	19.60	93.15	32.38	65.15	19.60	93.15	32.38
2	69.57	21.86	93.15	35.42	72.39	23.43	91.78	37.33
5	79.88	29.96	93.15	45.33	83.19	32.61	82.19	46.69
10	84.29	34.97	87.67	50.00	88.83	43.08	76.71	55.17
20	88.59	43.06	84.93	57.14	91.90	54.02	64.38	58.75
30	90.18	47.29	83.56	60.40	93.74	67.74	57.53	62.22
40	92.27	54.46	83.56	65.95	93.50	69.23	49.31	57.60
50	93.13	58.09	83.56	68.54	93.37	71.11	43.84	54.24
60	93.74	61.00	83.56	70.52	93.50	75.00	41.10	53.10
70	94.23	63.54	83.56	72.19	93.01	72.22	35.62	47.71
80	94.48	64.89	83.56	73.05	92.76	71.87	31.51	43.81
90	94.85	67.03	83.56	74.39	92.76	75.00	28.77	41.58
100	95.21	69.32	83.56	75.78	92.64	74.07	27.40	40.00
200	96.07	76.62	80.82	78.67	92.76	88.89	21.92	35.16
500	96.32	80.28	78.08	79.17	92.27	91.67	15.07	25.88
1000	96.69	83.82	78.08	80.85	91.90	88.88	10.96	19.51
2000 (all)	96.93	86.36	78.08	82.01	91.90	88.89	10.96	19.51



(a) Top 10 posts in steps of 1

(b) Top 100 posts in steps of 10

(c) Top 1000 post in steps of 100

Fig. 2. Performance of the system in terms of the maximum number of highly-weighted posts from each user. The solid lines correspond to the model with user-level attention (experiment 1), while the dotted lines correspond to the model with user-level average pooling (experiment 2).

5 Discussion

In order to further analyze the behavior of the user-level attention mechanism, the highest weights assigned by the attention mechanism were studied across users. In addition, we also calculated the average of the n-th highest weights assigned to the posts by the users, with n ranging from 1 to 10. We compared these values for two types of users: labelled by the model as anorexic (i.e. true-positive and false positive users) and labelled by the model as non-anorexic (i.e. true-negative and false-negative users). As Table 3 shows, on average, the attention mechanism has assigned 6.96 higher weights to the most highly weighted posts in users detected as anorexic, compared to users detected as non-anorexic. The value of this ratio drops in the lower-ranked posts. This seems to indicate that, generally when the attention mechanism assigns a high weight to a post, the system is more likely to label its author as positive. It is similar to when humans observe a piece of evidence, and tend to heavily base their decision upon it. This also seems to support the hypothesis that the attention mechanism assigns weights based mostly on how signalling their authors were anorexic, as opposed to signalling not having anorexia.

As opposed to Jain and Wallace [9] and Serrano and Smith [24], who reported that attention is not a means to explainability, our findings are generally in favor of explainability in the user-level attention mechanism. This may be due to the following two reasons:

1. The approach by Jain and Wallace [9] was only focused explainability of attention mechanism, when applied on the output of a recurrent encoder. We argue that, in such a case, each sample (contextual word representation, in their case) already has part of the information from the other samples in the context. As a result, finding the source of information is difficult in such

Table 3. Average weights assigned by the user-level attention mechanism to the n^{th} highest weighted posts for users detected by the system as anorexic (W_p) or non-anorexic (W_n)

Rank	W_p	W_n	$\mathrm{W}_p/\mathrm{W}_n$
1st	0.388	0.056	6.96
2nd	0.124	0.034	3.66
3rd	0.067	0.026	2.62
4th	0.047	0.021	2.18
5th	0.035	0.019	1.86
6th	0.029	0.017	1.70
7th	0.024	0.015	1.59
8th	0.019	0.014	1.31
9th	0.016	0.013	1.23
10th	0.015	0.012	1.20

a case. Serrano and Smith [24] also with using attention over non-encoded samples, and they showed that the level of explainability in this case is significantly higher than when the input to the attention is encoded (using an RNN or CNN). However, they mainly focused their report on the cases where the attention input is encoded. Our work was fully focused on non-encoded attention inputs.

2. The difference in the nature of the task we are performing is generally different from Jain and Wallace [9] and Serrano and Smith [24], as our approach focuses on the user-level (inter-document) attention mechanism, while their experiments were focused on intra-document attention. In a task involving the detection of a mental health problem, such as anorexia, the number relevant and informative posts is quite rare [14–17], while even in a similar task, there may be several ways of inferring information from a particular document.

Finally, in order to achieve stronger evidence that an inter-document attention is explainable, we believe that our approach would benefit from being used in conjunction with the experiments proposed by Jain and Wallace [9] and Serrano and Smith [24], as their experiments can also be applied to the inter-document attention mechanism.

6 Conclusion

In this work, we proposed a quantitative approach to validate the explainability of the user-level attention mechanism for the task of the detection of anorexia in social media users based on their online posts. Our results show that, the user-level attention mechanism has assigned higher weights to the posts from a user based on how much they were signalling the user is at risk of anorexia.

Two directions for the future work can be proposed: As indicated in Sect. 5, the first direction is to complement the current experiments with the ones proposed by Jain and Wallace [9] and Serrano and Smith [24], in order to see if the findings from the current experiments are in line with theirs. The second direction is to expand the current set of experiments to other mental health binary classification tasks (such as detection of depression, PTSD, or suicide risk), and later to multi-class or multi-label classification tasks in the field of NLP for clinical psychology.

Acknowledgements. The authors would like to thank the anonymous reviewers for their feedback on a previous version of this paper. This work was financially supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Proceedings of the 3rd International Conference for Learning Representations (ICLR 2015), San Diego, California, USA, May 2015
2. Chelba, C., et al.: One billion word benchmark for measuring progress in statistical language modeling. In: 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014), Singapore, September 2014
3. 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, Baltimore, Maryland, USA, pp. 51–60. Association for Computational Linguistics, June 2014. https://doi.org/10.3115/v1/W14-3207, http://aclweb.org/anthology/W14-3207
4. Zhao, X., Li, C.: Deep learning in social computing. In: Deng, L., Liu, Y. (eds.) Deep Learning in Natural Language Processing, pp. 255–288. Springer, Singapore (2018). https://doi.org/10.1007/978-981-10-5209-5_9
5. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2019), Minneapolis, Minnesota, USA, June 2019
6. Ghaeini, R., Fern, X., Tadepalli, P.: Interpreting recurrent and attention-based neural models: a case study on natural language inference. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018), Brussels, Belgium, pp. 4952–4957, October-November 2018
7. Goldberg, Y., Hirst, G.: Neural Network Methods in Natural Language Processing. Morgan & Claypool Publishers (2017)
8. Ive, J., Gkotsis, G., Dutta, R., Stewart, R., Velupillai, S.: Hierarchical neural model with attention mechanisms for the classification of social media text related to mental health. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPsych 2018), New Orleans, Louisiana, USA, pp. 69–77. Association for Computational Linguistics, June 2018
9. Jain, S., Wallace, B.C.: Attention is not explanation. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL 2019), Minneapolis, Minnesota, USA, pp. 3543–3556. Association for Computational Linguistics, June 2019

10. Kwak, G.H.J., Hui, P.: DeepHealth: deep learning for health informatics. Computing Research Repository arXiv:1909.00384 (2019)
11. Lee, J., Shin, J.H., Kim, J.S.: Interactive visualization and manipulation of attention-based neural machine translation. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Copenhagen, Denmark, pp. 121–126. Association for Computational Linguistics, September 2017
12. Lin, H., Tov, W., Qiu, L.: Emotional disclosure on social networking sites: the role of network structure and psychological needs. Comput. Hum. Behav. **41**, 342–350 (2014)
13. Lin, Z., et al.: A structured self-attentive sentence embedding. In: Proceedings of the 5th International Conference on Learning Representations (ICLR 2017), Toulon, France, April 2017
14. Losada, D.E., Crestani, F.: A test collection for research on depression and language use. In: Fuhr, N., et al. (eds.) CLEF 2016. LNCS, vol. 9822, pp. 28–39. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-44564-9_3
15. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF lab on early risk prediction on the Internet: experimental foundations. In: Jones, G.J.F., et al. (eds.) CLEF 2017. LNCS, vol. 10456, pp. 346–360. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65813-1_30
16. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2018: early risk prediction on the Internet (extended lab overview). In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 2018
17. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk 2019: early risk prediction on the Internet. In: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 2019
18. Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., Schwartz, H.A.: CLPsych 2018 shared task: predicting current and future psychological health from childhood essays. In: Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPsych 2018), New Orleans, Louisiana, USA, pp. 37–46. Association for Computational Linguistics, June 2018
19. Matero, M., et al.: Suicide risk assessment with multi-level dual-context language and BERT. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019), Minneapolis, Minnesota, USA, pp. 39–44. Association for Computational Linguistics, June 2019
20. Maupomé, D., Queudot, M., Meurs, M.J.: Inter and intra document attention for depression risk assessment. In: Proceedings of the 2019 Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, Canada, pp. 333–341, May 2019
21. Mohammadi, E., Amini, H., Kosseim, L.: CLaC at CLPsych 2019: fusion of neural features and predicted class probabilities for suicide risk assessment based on online posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019), Minneapolis, Minnesota, USA, pp. 34–38. Association for Computational Linguistics, June 2019
22. Mohammadi, E., Amini, H., Kosseim, L.: Quick and (maybe not so) easy detection of anorexia in social media posts. In: Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 2019

23. Peters, M., et al.: Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2018), New Orleans, Louisiana, USA, pp. 2227–2237. Association for Computational Linguistics, June 2018
24. Serrano, S., Smith, N.A.: Is attention interpretable? In: Proceedings of 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019), Florence, Italy, vol. abs/1906.03731. Association for Computational Linguistics, July 2019
25. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems (NIPS 2017), Long Beach, California, USA, vol. 30, pp. 5998–6008, January 2017
26. Wang, Y., Huang, M., Zhu, X., Zhao, L.: Attention-based LSTM for aspect-level sentiment classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016), Austin, Texas, USA, pp. 606–615. Association for Computational Linguistics, November 2016
27. Zirikly, A., Resnik, P., Uzuner, Ö., Hollingshead, K.: CLPsych 2019 shared task: predicting the degree of suicide risk in Reddit posts. In: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2019), Minneapolis, Minnesota, USA. pp. 24–33. Association for Computational Linguistics, June 2019