Published in "Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes", 2019, 9-12 September, Lugano, Switzerland, which should be cited to refer to this work.

A Data-Driven Approach for Measuring the Severity of the Signs of Depression using Reddit Posts

Paul van Rijen[1,2], Douglas Teodoro[1,2], Nona Naderi[1,2,3], Luc Mottin[1,2], Julien Knafou[1,2], Matt Jeffryes[1,2], Patrick Ruch[1,2]

[1] BiTeM group, HES-SO / HEG Geneva, Information Sciences, Geneva, Switzerland
[2] SIB Text Mining, Swiss Institute of Bioinformatics, Geneva, Switzerland
[3] University of Toronto, Toronto, Canada
contact: paul.vanrijen@hesge.ch

Abstract. In response to the CLEF eRisk 2019 shared task on measuring the severity of the signs of depression from threads of user submissions on social media, our team has developed a data-driven, ensemble model approach. Our system leverages word polarities, token extraction via mutual information, keyword expansion and semantic similarities for classifying Reddit posts according to the Beck's Depression Inventory (BDI). Individual models were combined at the post level by majority voting. The approach achieved a baseline performance for the assessed metrics, including Average Hit Rate and Depression Category Hit Rate, being equivalent to the median system in the limit of one standard deviation.

Keywords: Depression severity assessment, Social networks, Natural language processing, Machine learning.

1 Introduction

Depression is increasingly recognized as a major burden in public healthcare worldwide [1, 2]. In 2015 the World Health Organization (WHO) estimated that the total number of people living with depression was 322 million [2]. Depression is ranked as the single largest contributing factor to non-fatal health loss worldwide [1]. Major depressive disorder is associated with increased morbidity, disability and costs, increased mortality due to other co-occurring medical conditions including cardiovascular and pulmonary diseases and is a leading cause of suicide [2–4]. In addition to the high burden of disease, the majority of patients (50% globally) do not receive appropriate care [2]. Barriers to proper diagnoses and treatment include social stigmas and a low detection rate in primary care [5, 6]. Accurate and early detection of depression can help to lower these barriers and thus mitigate the associated health risks.

Social media networks, such as Facebook, Twitter and Reddit, enable people to share their opinions and sentiments about a wide range of topics online [7]. In recent years various studies have explored the potential of data from social media networks for detecting signs of depression [8, 9]. In addition, the scientific community has put forward

various shared tasks such as CLPsych [10] and CLEF eRisk [11, 12]. In CLEF eRisk 2018, the objective was to predict whether a user was depressed or not given a set of posts in a chronological order. Trotzek et al. [13] achieved the top F1-score using a bag of words ensemble method.

For 2019, the CLEF eRisk includes a task aimed towards measuring the severity of the signs of depression from threads of user submissions on social media [14, 15]. The eRisk task 3 involves filling in a Beck's Depression Inventory (BDI) questionnaire [16], which assesses the presence of feelings like sadness, pessimism, loss of energy, etc., in an individual, using a set of social media posts. Hence, the task changed from a standard classification task, as in tasks Early Detection of Signs of Anorexia (task 1) and Self-harm (task 2) of CLEF eRisk 2019 [17], to a combination of an information retrieval and an interactive dialogue task, where the system should simulate how a user would answer/fill in the questionnaire [18]. In response to this challenge our team developed a data-driven, multi-model approach based on word-polarities, mutual information and semantic similarities.

2 Methods

2.1 Beck's Depression Inventory

The BDI questionnaire has 21 questions in the following categories: sadness, pessimism, past failure, loss of pleasure, guilty feelings, punishment feelings, self-dislike, self-criticalness, suicidal thoughts or wishes, crying, agitation, loss of Interest, indecisiveness, worthlessness, loss of energy, changes in sleeping pattern, irritability, changes in appetite, concentration difficulty, tiredness or fatigue, and loss of interest in sex. The answers vary in a [0-3] scale, where 0 means the absence of the feeling and 1 to 3 the presence from a milder (1) to a stronger (3) form.

2.2 Task data

As shown in Table 1, the dataset for eRisk 2019 consists of Reddit posts from 20 users and contains the user identifier, the post timestamp, title and post content. The data was annotated at the user level with the depression severity according to Beck's Depression Inventory. The dataset was shared with the participants without the labels for the system development and was also used for the model evaluation during the test phase.

Table 1. Statics on the eRisk 2019 T3 dataset.

# users	20
# posts	10491
Median # posts per user	327
Sd dev # posts per user	446
Min # posts per user	29
Max # posts per user	1510

2.3 BDI questionnaire answering models

In this section, we describe the models used to automatically fill in the BDI questionnaire using the user's Reddit posts.

Model 1 - Word polarity. In this model, we aim to leverage word polarities for first classifying Reddit posts as depressive and next associate posts to relevant BDI dimensions.

Resources. For this model, we made use of the Multi Perspective Question Answering (MPQA) subjectivity lexicon [19][20] of over 8000 cues that can be used to express private states including emotions, evaluations and stances. In this lexicon, cues are annotated with positive, negative, or neutral polarity. In addition, this lexicon provides information regarding the part-of-speech of the cues and whether they are stemmed or not. In our model, we only considered single-word cues to determine whether a post is depressive or not.

For analyzing the posts with the BDI dimensions, we created a lexicon that provides cues for each dimension by first randomly selecting three subjects' writings (subject2341, subject5897 and subject9694). The MPQA single-word cues that appeared in these writings were used as cues for the BDI dimensions lexicon. Next we expanded the list of cues with the following resources: WordNet [21] to find synonyms, a sexual desires vocabulary [22] and the F.E.A.S.T.'s Eating Disorders Glossary [23]. The annotation process of assigning BDI dimensions to each of the cues was done by three team members. In the final version of lexicon, we included only the annotations that were agreed upon by at least two annotators. Some cues can be associated with multiple BDI dimensions. For instance, the cue 'hate' was associated to both the 'agitation' and 'self-dislike' dimensions, as illustrated by the following examples: Post 1: 'Hate when people do that.', Post 2: 'My life is already disintegrating, and I hate my grades.'. The final lexicon contained 668 words in total for all BDI dimensions. On average, the lexicon included 30 terms for each dimension. The majority of cues (583) were annotated with only a single dimension.

Classifier. First, we tagged the words in each post according to their polarity using the MPQA subjectivity lexicon. Since no training data was available during the official phase, we empirically set a threshold of 0.1 for the ratio of negative to positive words for classifying the posts as depressive. Only considering the depressive posts, we then tagged words with BDI dimensions using the developed lexicon. Finally, we calculated questionnaire responses by normalizing the tag counts for each BDI dimension into a [0-3] score.

Model 2 - Mutual information. In this model, we attempt to create a training dataset from Reddit to classify posts as depressive or not. We used the mutual information measure to extract relevant tokens from depressive posts [24]. Kraskov et al. proposes a model to estimate the mutual information M(X,Y) from samples of random points

distributed according to some joint probability density μ(x,y) based on entropy estimates from k-nearest neighbor distances.

Data. Two subreddit collections, containing 107,129 posts, were extracted as candidates for providing positive and negative depression tokens. The positive collection included 12 mental health related subreddits, such as *Anxiety*, *depression*, *eating_disorders*, *self-harm*, *social anxiety*, and *SuicideWatch*. The negative collection included 32 general subreddits, such as *all*, *AskReddit*, *explainlikeimfive*, *funny*, *movies*, and *worldnews*.

Training collection. Each post of the positive and negative collections was tokenized, stopword-removed, and stemmed, and unigram tokens were extracted and associated to the respective subreddit. Using the mutual information criteria, the 200 most informative tokens from each collection were used to tag each post. If a post from the positive collection contained more positive tokens, it was deemed as positive. Similarly, if a post from the negative collection contained more negative tokens, it was deemed as negative. The training set was then created with the positive and negative posts tagged with the 200 most informative tokens extracted from both collections. The final training collection contains 3,318 positive and 58,328 negative posts.

Classifier. A logistic regression classifier was trained to categorize posts into depressive or not using the positive and negative posts. Then, keywords from the BDI categories were expanded using WordNet and used to tag the positively classified posts. This model did not take into account the nuances of the positive answers for a BDI category, i.e., it considered the task as binary, assigning answers as 0 (negative) or 2 (positive) for a post.

Model 3 - Semantic similarity. Word embeddings have shown to capture semantic similarities and in recent years, various models have been proposed to generate these embeddings, such as word2vec [25], GloVe [26], and BERT [27]. Here, we propose to find the most semantically similar user posts to the questionnaire responses in order to estimate how a user may respond to the questionnaire. Given word embeddings, we generate the representation of each user post by averaging over the embeddings of words in the post. We use a similar approach to represent the questionnaire response vectors, i.e., we average the embeddings of words in each questionnaire response. We will then compute the similarity of a user post and a questionnaire response using cosine similarity. We use pre-trained GloVe word embeddings[1] [26] (trained on 2 billion tweets and has 200 dimensions) to represent the words. In order to filter out the irrelevant posts, in the first step, we remove the posts that are not similar to the questionnaire responses using only the noun and verb vectors and a threshold that was chosen empirically (0.8). For the remaining posts, we compute the vector-based distance of each post and questionnaire responses and choose the most similar response for that post. Treating each post as the average of word embeddings does not consider word orders and

1 https://nlp.stanford.edu/projects/glove/

not likely produce a good representation for the longer posts, but it has shown to provide a relatively good baseline. This post-level representation can be improved by leveraging state-of-the-art sentence embedding models [28].

Model 4 – Ensemble. Two ensemble approaches were tested - micro and macro-voting using models 1 to 3. In micro-voting, models were combined at the post level. If more than two models classified a post as positive for depression, and if two (majority) or three (strict) models classified the post as positive for a category, the category was deemed as positive for that user. In macro voting, results were combined using the average category prediction from the three models. The official run (BiTeM run 0) was generated using the strict micro-voting ensemble.

2.4 Tuning individual models

For each of the individual models we applied a threshold k, which determines the minimal number of positive posts (i.e., categorized as depressed) the system would need to consider a user as depressed. Only then, responses would be given to questions in the 21 BDI categories. Hence, a positive post could be considered as a proxy for a depressive episode. As there was no training data for this task, we used an empirical k=5, that is, five depressive episodes would be needed to regard a user as depressed. Then, for each category, if multiple answers (0 to 4) were retrieved for a deemed positive user, then the system assigned the response with the highest value for the category.

3 Results

System effectiveness metrics considered for this task are Average Hit Rate (AHR), Average Closeness Rate (ACR), Average Difference between Overall Depression Levels (ADODL) and Depression Category Hit Rate (DCHR). The AHR measures the ratio of cases where the computed questionnaire produces exactly the same answer as the real questionnaire. The ACR measures the averaged absolute distance on an ordinal scale between the automated answer and the real answer. ADODL assesses the system's performance by first calculating the overall depression score (sum of all answers) and, next, the absolute difference (ad_overall) between the automated score and the real overall depression score. Depression levels are normalized as follows; DODL = (63 - ad_overall) / 63. DCHR measures the fraction of cases in which the automated questionnaire resulted in same depression severity categorization as the real questionnaire. Table 2 shows the four depth-of-depression categories and the associated depression levels used in this task[15].

Table 2. Depth-of-depression categories

Depression category	Depression levels
Minimal	0-9
Mild	10-18
Moderate	19-29
Severe	30-63

3.1 Official results

Our team submitted one official run to the Task 3 of the eRisk challenge. This run combined the results of the three models described above using voting. The voting was performed in a micro-average fashion, i.e., the results for each model were combined at post level. Table 3 shows the results of our model. Overall, it has a baseline performance for all the metrics, being equivalent to the median model of the participants within the limit of one standard deviation.

Table 3. Official evaluation results. BiTeM results in comparison with the overall task systems

Run	AHR (%)	ACR (%)	ADODL (%)	DCHR (%)
BiTeM run 0	32.14	62.62	72.62	25.00
Median	35.71	66.51	74.81	25.00
Std dev	5.91	5.44	4.04	9.93
Min	22.38	56.19	66.19	5.00
Max	41.43	71.27	81.03	45.00

3.2 Unofficial evaluation results

After the official phase our team conducted some further experiments. In addition to the micro and macro ensemble models, we evaluated the performance on the three models separately. For each of these models we applied a k=5 threshold, i.e., if the model classified at least 5 posts as positive for a category then the category was deemed as positive for that user. Table 4 describes the results of the various models. Overall, both ensemble micro models outperform the individual models and the ensemble macro model. The ensemble micro majority model outperforms the model used in the official phase on the ACR and ADODL metrics but at significant penalty in DCHR.

Table 4. Unofficial evaluation results. Individual models and micro and macro ensembles. *Model used during the official phase.

Run	Voting	AHR (%)	ACR (%)	ADODL (%)	DCHR (%)
Model 1		29.76	56.75	71.98	**30.00**
Model 2		26.42	60.00	73.81	25.00
Model 3		19.29	47.70	59.44	20.00
Ensemble micro	majority	25.71	**66.59**	**77.38**	10.00
	*strict	**32.14**	62.62	72.62	25.00
Ensemble macro	average	25.00	63.73	74.37	20.00

3.3 Binary classifier

We repeated the unofficial experiments under the assumption that the questionnaire is binary, i.e., the user expressed any feelings of depression or not. Table 5 shows the model results under this assumption. Similar to the non-binary results, the ensemble micro majority model outperforms the model used in the official phase but again at a significant penalty in the DCHR metric.

Table 5. Evaluation considering the questionnaire as binary. *Model used during the official phase. **Baseline.

Run	Voting	AHR (%)	ACR (%)	ADODL (%)	DCHR (%)
Model 1		41.91	80.63	84.44	35.00
Model 2		48.81	82.94	88.33	**45.00**
Model 3		50.00	83.33	87.46	25.00
Ensemble micro	majority	56.90	**85.63**	**89.60**	15.00
	*strict	43.81	81.27	84.76	35.00
Ensemble macro	average	48.33	82.78	87.38	25.00
**All 0's		34.52	78.17	78.17	30.00
**All 1's		**65.48**	65.48	88.49	30.00

3.4 Impact of k on the individual model's performance

Fig 1. shows how the ADODL metric varies in function of k, i.e., the number of positive posts necessary to confirm a category as positive. Models 1 and 2 presents their highest ADODL around k=5 whereas model 3 in k=3. As there was no training set during the official submission phase, we set these values empirically to k=5 for all the individual models based on a manual analysis of some results. We suspected that k=1 would create too many false positives. A similar pattern is also seen for the other metrics (not shown here for brevity).

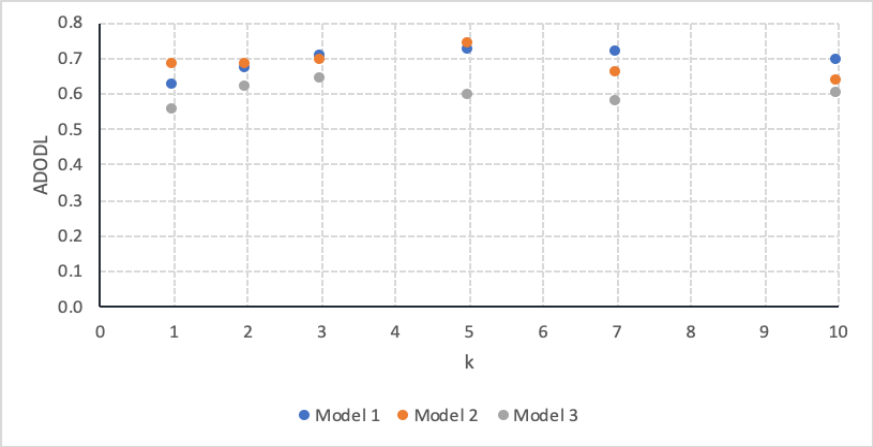

Fig. 1. Variation of the model's performance in function of k.

4 Discussion

Social media can provide valuable resources for assessing individual's mental health that could be useful for early detection and consequent healthcare provision. We developed a simple model for measuring the severity of the signs of depression from Reddit posts based on word polarities, mutual information and semantic similarities. The ensemble model used in the official phase achieved modest results. This could be explained by the significant negative effect of weak individual models during the construction of the ensemble model.

Nevertheless, both micro ensemble models significantly improved upon the individual models' results for all the metrics apart from DCHR. Indeed, for the DCHR metric, model 1 presented the best performance in the standard questionnaire answer, being able to predict the correct depression severity category for 30% of the users. The ensemble macro model did not improve upon the ensemble micro models. One possible cause can be the relatively small set of candidate models from which the results were averaged to calculate the ensemble category predictions.

Answering the BDI questionnaire without training data proved to be a challenging task. Indeed, even when considering the questionnaire as binary, the participant models were outperformed by a naïve all-positive answer baseline on some of the metrics. Model 2 performed the best in the DCHR and correctly predicted depression in 45% of the cases when considering the questionnaire as binary. This remarkable improvement over the 20% performance in DCHR in the standard questionnaire could be explained by the fact that this model, in contrast to model 1 and 3, considered the task as binary already in its conception, not taking the nuances in positive answers into account.

Finally, as expected, training the models for some parameters would significantly improve their performance. Indeed, most of the individual and ensemble model parameters, such as cut-off, k, and voting weight, were set empirically during the official phase and the results reported here do not try to tune them based on the gold standard

answers. As shown in Fig. 1, tuning only k, for example, would result on an average improvement of up to 13% for the ADODL metric if we consider k=1 as the baseline. This effect is also seen for the AHR, ACR and DCHR metrics, which can have an average relative performance increase of up to 32%, 14% and 33%, respectively, with tuning.

5 Conclusion

The task T3 of CLEF eRisk 2019 aimed to measure the severity of the signs of depression using user threads available in social media. The organizers provided a dataset containing Reddit posts from 20 users and the goal was to automatically fill the 21 questions of Beck's Depression Inventory for each of the users. Our team developed a data-driven, ensemble model combining sentiment lexicons, mutual information and embedding similarities in order to overcome the lack of training samples. The model achieved a baseline performance, being equivalent to the median system from the overall challenge. Nevertheless, answering the BDI questionnaire without training data showed to be a challenging task, with an average hit rate of less than 42% for the top 1 system (32% in our case). Indeed, for some metrics, our system was outperformed by a naïve all-positive answer baseline in a binary classification. As next steps, we aim to leverage the post evidences created during this task to improve the performance of our classification model.

References

1. Marcus, M., Yasamy, M.T., Van Ommeren, M., Chisholm, D., Saxena, S.: Depression: A global public health concern. World Health Organization Paper on Depression. 6–8 (2012).
2. World Health Organization: Depression and other common mental disorders: global health estimates. World Health Organization (2017).
3. Forte, A., Baldessarini, R.J., Tondo, L., Vázquez, G.H., Pompili, M., Girardi, P.: Long-term morbidity in bipolar-I, bipolar-II, and unipolar major depressive disorders. Journal of Affective Disorders. 178, 71–78 (2015). https://doi.org/10.1016/j.jad.2015.02.011.
4. Kessler, R.C., Berglund, P., Demler, O., Jin, R., Koretz, D., Merikangas, K.R., Rush, A.J., Walters, E.E., Wang, P.S.: The Epidemiology of Major Depressive Disorder: Results From the National Comorbidity Survey Replication (NCS-R). JAMA. 289, 3095–3105 (2003). https://doi.org/10.1001/jama.289.23.3095.
5. Rodrigues, S., Bokhour, B., Mueller, N., Dell, N., Osei-Bonsu, P.E., Zhao, S., Glickman, M., Eisen, S.V., Elwy, A.R.: Impact of Stigma on Veteran Treatment Seeking for Depression. American Journal of Psychiatric Rehabilitation. 17, 128–146 (2014). https://doi.org/10.1080/15487768.2014.903875.
6. Vermani, M., Marcus, M., Katzman, M.A.: Rates of Detection of Mood and Anxiety Disorders in Primary Care: A Descriptive, Cross-Sectional Study. Prim Care Companion CNS Disord. 13, (2011). https://doi.org/10.4088/PCC.10m01013.
7. Gramlich, J.: 5 Facts about Americans and Facebook. Pew Research Center. 10, (5).

8. Choudhury, M.D., Gamon, M., Counts, S., Horvitz, E.: Predicting Depression via Social Media. In: Seventh International AAAI Conference on Weblogs and Social Media (2013).
9. Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: an integrative review. Current Opinion in Behavioral Sciences. 18, 43–49 (2017). https://doi.org/10.1016/j.cobeha.2017.07.005.
10. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., Mitchell, M.: CLPsych 2015 shared task: Depression and PTSD on Twitter. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality. pp. 31–39 (2015).
11. Losada, D.E., Crestani, F., Parapar, J.: eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 346–360. Springer (2017).
12. Losada, D.E., Crestani, F., Parapar, J.: Overview of eRisk: Early Risk Prediction on the Internet. In: International Conference of the Cross-Language Evaluation Forum for European Languages. pp. 343–361. Springer (2018).
13. Trotzek, M., Koitka, S., Friedrich, C.M.: Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia.
14. Losada, D.E., Crestani, F., Parapar, J.: Early Detection of Risks on the Internet: An Exploratory Campaign. In: European Conference on Information Retrieval. pp. 259–266. Springer (2019).
15. CLEF eRisk: Early risk prediction on the Internet | CLEF 2019 workshop, https://early.irlab.org/.
16. Beck, A.T., Steer, R.A., Carbin, M.G.: Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clinical psychology review. 8, 77–100 (1988).
17. Nona Naderi, Julien Gobeill, Douglas Teodoro, Emilie Pasche, Patrick Ruch: A Baseline Approach for Early Detection of Signs of Anorexia and Self-harm in Reddit Posts. In: Proceedings of the CLEF 2019 Workshop.
18. Sutcliffe, R.F., Peñas, A., Hovy, E.H., Forner, P., Rodrigo, Á., Forascu, C., Benajiba, Y., Osenova, P.: Overview of QA4MRE Main Task at CLEF 2013. In: CLEF (Working Notes) (2013).
19. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing (2005).
20. MPQA Resources, http://mpqa.cs.pitt.edu/#subj_lexicon.
21. Fellbaum, C.: WordNet: An electronic lexical database Cambridge. MA: MIT Press. (1998).
22. feeling sexual excitement or desire - synonyms and related words | Macmillan Dictionary, https://www.macmillandictionary.com/thesaurus-category/british/feeling-sexual-excitement-or-desire.
23. Eating Disorders Glossary, http://glossary.feast-ed.org/.
24. Kraskov, A., Stögbauer, H., Grassberger, P.: Estimating mutual information. Phys. Rev. E. 69, 066138 (2004). https://doi.org/10.1103/PhysRevE.69.066138.
25. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient Estimation of Word Representations in Vector Space. (2013).

26. Pennington, J., Socher, R., Manning, C.: Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP). pp. 1532–1543 (2014).
27. Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. (2018).
28. Ethayarajh, K.: Unsupervised random walk sentence embeddings: A strong but simple baseline. In: Proceedings of The Third Workshop on Representation Learning for NLP. pp. 91–100 (2018).