Feature Engineering for Depression Detection in Social Media

Maxim Stankevich, Vadim Isakov, Dmitry Devyatkin and Ivan Smirnov
Institute for Systems Analysis, Federal Research Center "Computer Science and Control" of RAS, Moscow, Russian Federation

Keywords: Depresion Detection, Social Media, Classification, Machine Learning.

Abstract: This research is based on the CLEF/eRisk 2017 pilot task which is focused on early risk detection of depression. The CLEF/eRsik 2017 dataset consists of text examples collected from messages of 887 Reddit users. The main idea of the task is to classify users into two groups: risk case of depression and non-risk case. This paper considers different feature sets for depression detection task among Reddit users by text messages processing. We examine our bag-of-words, embedding and bigram models using the CLEF/eRisk 2017 dataset and evaluate the applicability of stylometric and morphological features. We also perform a comparison of our results with the CLEF/eRisk 2017 task report.

1 INTRODUCTION

It is a known fact that mental health is a crucial component of public health. Depression is one of the leading causes of disability worldwide. According to the results from the World Health Surveys, millions of people around the world are suffering from depression and this number is growing (Moussavi et al., 2007). This mental disorder is often hidden by their carriers, which can make the early detection task of depression very difficult.

Popular social networks can serve as a tool for dealing with this problem. Text messages published in this networks contain a lot of hidden information about their authors. In this paper, we consider the classification task of Reddit users by processing their text messages in order to detect depression. We evaluate various baseline features (*tf-idf*, embeddings, bigrams) as well as more complex features like morphology and stylometric and compare our results with the performance of CLEF/eRsik 2017 participants. We present word embeddings model, which has not previously tested on the depression task before.

2 RELATED WORK

It is worth noting several works related to the problem of depression detection in social media.

One of the research related to the described task (De Choudhury et al., 2013) investigated the potential of using Twitter as a tool for depression detection and diagnosing. Their gold standard contains text messages of Twitter users who were reported being diagnosed with clinical depression, using a standard psychometric instrument. They revealed features that indicate individuals with depression: lowered social activity, greater negative emotion, high self-attentional focus, increased relational and medicinal concerns, greater expression of religious involvement, and high level of social interaction with other depressed users.

Another related work (Wang et al., 2013) proposed a depression detection method based on the subject-dependent sentiment analysis of the microblog. They argue that negative emotions and a lack of positive emotions are important symptoms that indicate the presence of depression. They used 10 features of depressed users derived from psychological research: quantity of emoticons, interaction features (how users interact with each other), behavior features (frequencies of the posting, active period).

We also should note that there are several interesting approaches, presented by teams who report thier classification models for CLEF/eRisk 2017 task (Losada et al., 2017).

A lot of works (Malam et al., 2017; Trotzek et al., 2017; Sadeque et al., 2017; Almeida et al., 2017) use lexicon features that indicate usage of specific words like emotional words, sentiment words, and specific terms related for medication or diagnosis. This words usually are taken from pre-defined dictionaries.

Stankevich, M., Isakov, V., Devyatkin, D. and Smirnov, I.
Feature Engineering for Depression Detection in Social Media.
DOI: 10.5220/0006598604260431
In *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2018)*, pages 426-431
ISBN: 978-989-758-276-9

Several works present features based on bag-of-words and n-grams (Trotzek et al., 2017; Almeida et al., 2017; Farıas-Anzaldúa et al., 2017). In terms of F1-score, such approaches show best results in the shared task.

In some works (Trotzek et al., 2017; Almeida et al., 2017), authors used features based on part-of-speech tags: the number of nouns, pronouns, and other parts of speech.

There are two works (Trotzek et al., 2017; Sadeque et al., 2017) that use recurent neural network and one work (Villatoro-Tello et al., 2017) that presents graph-based model for CLEF/eRisk 2017 task.

The goal of our work is to examine bag-of-words, bigram and embedding models on CLEF/eRisk 2017 dataset, evaluate the applicability of morphological and stylometric features, compare our classification results with CLEF/eRisk 2017 participants. We want to note that approach based on embeddings was previously untested on depression detection task.

3 CLEF/ERISK 2017 SHARED TASK

This work is based on the CLEF/eRisk 2017 (Conference and Labs of the Evaluation Forum: early risk prediction on the Internet)[1] dataset, which was provided as the part of the pilot task of early risk detection of depression. The dataset consists of 887 Reddit users examples, 135 of them have been identified as belonging to a risk case of depression (Losada and Crestani, 2016). Each example of the dataset contains all text messages of the user for a certain period of time. The time interval between first and last message for each instance is different, as well as the total number of messages. The smallest of the examples contains only 10 messages, and the largest one is 2000 messages. Overall, the dataset consists of 15% of positive examples (risk case of depression) and 85% of negative examples (non-risk case, taken from Reddit users randomly). CLEF/eRsik 2017 task dataset consists of 486 train examples (83 positive samples and 403 negative samples) and 401 test examples (52 positive samples and 349 negative samples).

It is important to note, that we can compare our results with other models reported on CLEF/eRisk 2017 task only with some restrictions. Originally, CLEF/eRisk 2017 organizers sent 10% of all messages (10 chunks for all data) for every test example each week and allowed participants to send their decision on every example (depressed class / non-depressed class) before all 100% of the data were submitted. But CLEF/eRisk 2017 organizers noted in their task review (Losada et al., 2017) that the most of the models made decisions only on the last chunks and some models made decisions only when all chunks were submitted.

[1] http://early.irlab.org/

4 FEATURE SETS

Classification of the Reddit users proceeds to the classification of their texts, represented as messages. We examine various sets of features and use 3 of them as different baselines. The first one relies on term frequencies of all words used by examples. The second approach is based on averaging of users word embeddings. The last feature set contains bigram features. We also investigate stylometric and morphological features as additional feature sets.

We used FreeLing (Padró and Stanilovsky, 2012) for word tokenization and part-of-speech tagging.

4.1 TF-IDF

The first feature set is based on bag-of-words (BoW), which is used to set *tf-idf* values for each word in the dataset, except those that appear in the whole documents collection only once. After this, *tf-idf* vectors are used as features for classification.

The idea behind this approach consists in the fact that some words occur in different groups (positive examples / negative examples) with different frequencies. For example, the word *depression* occurs 1,686 times in depressive texts and 1,233 times in non-depressive texts. On average, 9.04 usages per depressed user and 0.92 usages per non-depressed user (see Figure 1).

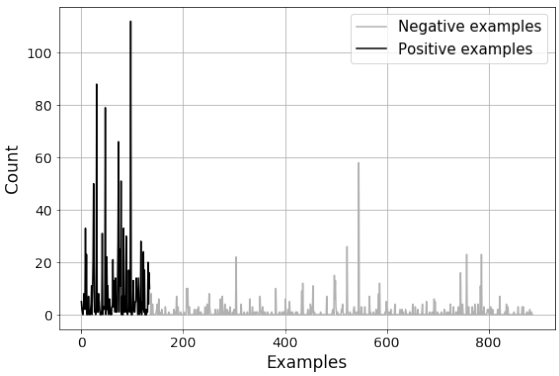

Figure 1: Usage count of word *depression* user.

4.2 Word Embeddings

The second feature set contains 100-dimensional word embeddings. The embeddings were trained on Twitter messages (Pennington et al., 2014) and are suitable for the classification of social media texts. To produce feature vector, we averaged embeddings of 850 most informative (sorted by *tf-idf*) unique words for every user. Each word embedding before averaging was multiplied by the number of usage of this word.

It is also possible to consider another count of words for averaging. We examined different counts of words (from 50 to 1000) and evaluated classification results on training part of the dataset by cross-validation. Best result achieved with 850 words. We used the technique called 't-SNE' (t-distributed Stochastic Neighbor Embedding) (Maaten and Hinton, 2008), which helps to visualize high-dimensional data by giving each datapoint location in the two-dimensional map. Results of t-SNE on the data illustrated on Figure 2.

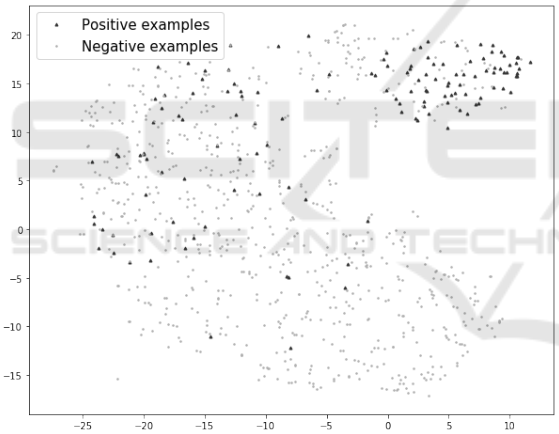

Figure 2: 850 words embeddings converted by t-SNE into 2-dimensional map.

4.3 Bigrams

The last feature set consists of n-grams. N-gram is a contiguous sequence of items. The items can be phonemes, syllables, letters, words or base pairs. N-gram models are widely used in statistical natural language processing for language modeling, sentiment analysis (Pak and Paroubek, 2010), authorship attribution (Kešelj et al., 2003), question answering, relationship extraction, and much more. In this study, we used bigrams as elements consisting of two symbols.

4.4 Additional Features

During the study, it was noted that positive and negative examples have a difference in an average number of words and sentences per each text message. It seems that persons in depressed class usually write longer sentences than non-depressed ones (see Table 1). We considered to use them as additional features for classification. The tests showed that Reddit users from depressed class use slightly less unique words in their speech than another, so we used lexicon size as a feature too.

We also noted that depressed and non-depressed Redditors use different parts of speech in different proportions. The average percentage of various parts of speech (nouns, verbs, adjectives, etc.) are showed in Table 2. We applied usage proportion as the features for classification.

Another assumption was examined that depressed class has a tendency to post their messages at the night time more often than non-depressed class, but Figure 3 shows that there is no significant contrast between classes.

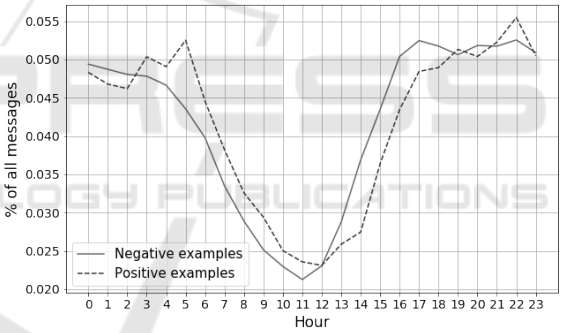

Figure 3: % of messages in different hours.

In this paper, we examined how *tf-idf*, embedding and bigram features work with additional ones. The following labels are used for different feature sets (see Table 3). Stylometric features include lexicon volume, users averaged numbers of words per message, sentences per message and words per sentence. Morphology features contain parts of speech usage proportions.

5 RESULTS OF EXPIREMENTS

We tested various models with *tf-idf*, embeddings and bigrams approaches and decided to use Support vector machine algorithm (SVM) for classification. Random forest also showed decent performance with our

Table 1: Averaged values for stylometric features.

Averaged values	Positive examples	Negative examples
Words in message	41.39 ± 19.94	27.77 ± 13.98
Sentences in message	2.89 ± 1.20	2.01 ± 0.68
Words in sentence	14.77 ± 3.82	13.09 ± 3.36

Table 2: Usages of different parts of speech in %.

Part of speech	PE,%	NE,%
Nouns	19.36	25.18
Verbs	22.04	20.26
Pronouns	13.55	10.00
Adjectives	6.70	7.81
Articles	7.82	8.80
Adverbs	13.69	12.12
Conjunctions	3.60	2.93

Table 3: Feature sets.

Feature set	*Label*
Tf-idf values	tf-idf
Bigrams	bigram
Embeddings	emb
Stylometric features	stl
Morphology features	mrph

feature sets so we included it in the report. Parameters and hyperparameters for classification were set up by grid-search and 5-fold cross-validation. In this work, we used sklearn realization of classification algorithms (Pedregosa et al., 2011).

It is worth mentioning that we used the same examples for train and test parts as it was managed in the CLEF/eRisk 2017 task. We also calculated recall, precision and F1-score with respect to the minority class (Positive examples). It is the same way as it was computed in CLEF/eRisk 2017 task (Losada and Crestani, 2016). Table 4 demonstrates the results of the models.

Overall, SVM model with the *tf-idf+stl+mrph* feature set achieves the best F1-score (63.36%) in our experiments. SVM also shows highest recall score (84.61%), and decent result for F1-score (61.53%) on embeddings with stylometric features. *Tf-idf* with random forest reaches best precision on the test data (79.41%). Bigram sets obtain worse results than other models but still show comparable F1-score (58.71%). It is also notable that random forest achieves best results with morphological features for all three sets.

Our *tf-idf* model with morphological features shows high precision on the test data (79%). The best precision score reported on the CLEF/eRisk 2017 task is 69% (Losada et al., 2017). The embedding model shows high recall score (84.61%) which is not best recall score on the task but this model also has decent F1-score (61.53%). The best CLEF/eRsik 2017 F1-score is 64% and highest F1-score in our experiments is 63%.

6 CONCLUSION

In this paper, we explored different sets of features for the task of depression detection in social media. In order to classify Reddit users, we worked with the text of their messages and investigated how bag-of-words, word embeddings and bigrams work with CLEF/eRisk 2017 data. We considered some additional sets of features (stylometric and morphology) and evaluated their applicability. Our research showed that in the most cases additional features improve classification results. *Tf-idf* with morphology set achieves best results on the test data with 63% F1-score. Embedding features always obtain better recall score than *tf-idf* but accuracy and precision score is less. We assume that high recall with embedding features might be consequence of the fact that embeddings contain additional information about word co-occurrences (Pennington et al., 2014). We achieved decent results in comparison with the CLEF/eRisk 2017 task report.

In the future work, we want to apply semantic role labeling to explore additional features and test different classification models to obtain better results. We are also going to apply our research of the problem for the messages in different languages.

ACKNOWLEDGEMENTS

The reported study was funded by RFBR according to the research project 17-29-02225.

Table 4: Classification results.

SVM				
Sets	*Recall score, %*	*Precision score, %*	*F_1 score, %*	*Accuracy score, %*
tf-idf	57.69	63.83	60.61	90.27
tf-idf + stl	63.46	61.11	62.26	90.02
tf-idf + mrph	63.46	56.89	60.02	89.02
tf-idf + stl + mrph	61.53	**65.31**	**63.36**	**90.77**
emb	84.61	47.31	60.68	85.78
emb + stl	84.61	**48.35**	**61.53**	**86.28**
emb + mrph	80.76	46.66	59.15	85.53
emb + stl + mrph	78.84	43.15	55.78	83.79
bigram	**65.87**	53.80	**57.38**	**86.30**
bigram + stl	61.62	51.40	53.30	84.66
bigram + mrph	54.12	**61.34**	50.42	84.44
bigram + stl + mrph	52.23	59.06	47.80	83.76
Random forest				
Sets	*Recall score, %*	*Precision score, %*	*F_1 score, %*	*Accuracy score, %*
tf-idf	51.92	72.97	60.67	91.27
tf-idf + stl	51.92	77.14	62.06	91.77
tf-idf + mrph	51.92	**79.41**	**62.79**	**92.01**
tf-idf + stl + mrph	51.92	77.14	62.06	91.77
emb	55.76	53.70	54.71	88.02
emb + stl	57.69	53.57	55.55	88.02
emb + mrph	**61.53**	**56.14**	**58.71**	**88.77**
emb + stl + mrph	57.69	54.54	56.07	88.27
bigram	64.42	**45.50**	53.30	**85.36**
bigram + stl	63.88	45.32	52.99	85.30
bigram + mrph	**69.54**	44.07	**53.92**	84.58
bigram + stl + mrph	68.35	43.72	53.31	84.46

REFERENCES

Almeida, H., Briand, A., and Meurs, M.-J. (2017). Detecting early risk of depression from social media user-generated content.

De Choudhury, M., Gamon, M., Counts, S., and Horvitz, E. (2013). Predicting depression via social media. In *ICWSM*, page 2.

Farıas-Anzaldúa, A. A., Montes-y Gómez, M., López-Monroy, A. P., and González-Gurrola, L. C. (2017). Uach-inaoe participation at erisk2017.

Kešelj, V., Peng, F., Cercone, N., and Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, volume 3, pages 255–264.

Losada, D. E. and Crestani, F. (2016). A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

Losada, D. E., Crestani, F., and Parapar, J. (2017). Clef 2017 erisk overview: Early risk prediction on the internet: Experimental foundations.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(Nov):2579–2605.

Malam, I. A., Arziki, M., Bellazrak, M. N., Benamara, F., El Kaidi, A., Es-Saghir, B., He, Z., Housni, M., Moriceau, V., Mothe, J., et al. (2017). Irit at e-risk.

Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., and Ustun, B. (2007). Depression, chronic diseases, and decrements in health: results from the world health surveys. *The Lancet*, 370(9590):851–858.

Padró, L. and Stanilovsky, E. (2012). Freeling 3.0: Towards wider multilinguality. In *LREC2012*.

Pak, A. and Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10.

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

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In

Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543.

Sadeque, F., Xu, D., and Bethard, S. (2017). Uarizona at the clef erisk 2017 pilot task: Linear and recurrent models for early depression detection.

Trotzek, M., Koitka, S., and Friedrich, C. M. (2017). Linguistic metadata augmented classifiers at the clef 2017 task for early detection of depression.

Villatoro-Tello, E., Ramırez-de-la Rosa, G., and Jiménez-Salazar, H. (2017). Uams participation at clef erisk 2017 task: Towards modelling depressed bloggers.

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 *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 201–213. Springer.