Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference “Dialogue 2019”

Moscow, May 29—June 1, 2019

PREDICTING DEPRESSION FROM ESSAYS IN RUSSIAN

Stankevich M. A. (stankevich@isa.ru)

Artificial Intelligence Research Institute, FRC CSC RAS, Moscow, Russia

Smirnov I. V. (ivs@isa.ru)

Artificial Intelligence Research Institute, FRC CSC RAS; Peoples’ Friendship University of Russia (RUDN University), Moscow, Russia

Kuznetsova Y. M. (kuzjum@yandex.ru)

Artificial Intelligence Research Institute, FRC CSC RAS, Moscow, Russia

Kiselnikova N. V. (nv.pirao@gmail.com)

Psychological Institute of Russian Academy of Education, Moscow, Russia

Enikolopov S. N. (enikolopov@mail.ru)

Department of Medical Psychology, Mental Health Research Centre, Moscow, Russia

The study is focused on the detection of depression by processing and classification of short essays written by 316 volunteers. The set of 93 essays was provided by two different teams of psychologists who asked patients with clinically confirmed depression to write short essays on the neutral topic. The other 223 essays on the same topic were written by volunteers who completed questionnaires, which are designed to reveal depression status and did not demonstrate any signs of mental illnesses. The study describes psycholinguistic and classic text features which were calculated by utilizing natural language processing tools and were used to perform on the classification task. The machine learning classification models achieved up to 73% of f1-score for the task of revealing essays written by people with depression.

Key words: depression detection, text classification, psycholinguistic features, natural language processing

ВЫЯВЛЕНИЕ ПРИЗНАКОВ ДЕПРЕССИИ У АВТОРОВ ЭССЕ НА РУССКОМ ЯЗЫКЕ

Станкевич М. А. (stankevich@isa.ru)

Институт проблем искусственного интеллекта
ФИЦ ИУ РАН, Москва, Россия

Смирнов И. В. (ivs@isa.ru)

Институт проблем искусственного интеллекта ФИЦ ИУ РАН;
Университет дружбы народов (РУДН), Москва, Россия

Кузнецова Ю. М. (kuzjum@yandex.ru)

Институт проблем искусственного интеллекта
ФИЦ ИУ РАН, Москва, Россия

Кисельникова Н. В. (nv.pirao@gmail.com)

Психологический институт РАО, Москва, Россия

Ениколопов С. Н. (enikolopov@mail.ru)

Научный центр психического здоровья
НЦПЗ РАН, Москва, Россия

Данная работа направлена на задачу выявления депрессии при помощи обработки и классификации 316 эссе. Коллекция из 93 эссе была предоставлена двумя коллективами психологов, которые попросили пациентов с клинически подтвержденной депрессией написать эссе на нейтральную тему. Остальные 223 эссе на аналогичную тему были написаны добровольцами, которые прошли стандартный опросник на выявление депрессии и не показали признаков наличия ментальных заболеваний. Исследование описывает различные психолингвистические и стандартные текстовые признаки, полученные при помощи инструментов обработки естественного языка и использованные для задачи классификации. Основанные на машинном обучении классификационные модели продемонстрировали до 73% f1-меры в задаче обнаружения эссе, написанных людьми с депрессией.

Ключевые слова: обнаружение депрессии, классификация текста, психолингвистические признаки, обработка естественного языка

1. Introduction

It is a known fact that depression is one of the leading causes of disability worldwide and it affects millions of people around the world [11]. Depression can make a significant impact on the daily lifestyle and behavior of people. At the same time, a considerable number of depression cases stay untreated or undetected [21]. It is also known that severe types of depression affect the way human thinks and, therefore, influence human ability to express thoughts in oral speech and writings [2]. The psycholinguistic investigating this impact of depression and other mental diseases on human linguistics and propose some valuable methodology on it. But manual psycholinguistic analysis requires a lot of effort and time. Development of natural language processing tools allows to partially solve this problem [20]. At the same time, machine learning methods present a lot of opportunities to reveal human psychological attributes when applied on text data, for example in social media [17]. We currently aimed to develop such methods for the Russian social media and users' writings. But for the Russian language, we are lacking background knowledge about relations between psychological attributes of the human and his text.

The main idea of the study consists in applying machine learning and natural language processing tools to perform on the task of depression detection in essay writings on the Russian language. We formed two collections of essays: 93 essays written by people with clinically diagnosed depression, and 223 essays written by volunteers who completed a psychological questionnaire to confirm they did not demonstrate any depression signs. Thus, we focused on binary classification in order to evaluate the ability of machine learning approach to detect if a text belongs to depressed or healthy subject. We present features retrieved from essays including classical text features, psycholinguistic features, n-grams, and sentiment. It is important to note that currently there are no studies devoted to the depression detection task among Russian language and the psycholinguistic features proposed in the study are not previously tested on similar tasks.

2. Related work

Linguistic Inquiry and Word Count (LIWC) is one of the most frequently used tools for automatic text analysis for researches related to psychology and psycholinguistics [13]. The main idea embodied in this tool is that the author's psychological characteristics are related to the text's quantitative parameters: the frequency of punctuation marks, words of a certain part of speech (prepositions, conjunctions, pronouns, adverbs), words of a certain lexic-semantic group (negative or positive emotions, describing cognitive processes).

The task of depression detection mostly focused on social media data. There is a lot of studies that consider the task of detection depression by analyzing social media messages. The work presented in [6] describes the classification of social media messages written by depressed and healthy users. The authors achieved 74% of accuracy with SVM classifier using the following features: social media activity, time, N-grams, postags, and features based on LIWC.

Another work observes depression detection problem as a task of detecting vocabulary related to 9 depression symptoms [24]. Authors processed messages from Twitter to indicate the presence of these symptoms in users' writings. This approach is based on the observation that users of social networks frequently write about their mental state [3]. The experiments were focused on multi-label classification and comparison of semi-supervised topic modeling over time model (ssToT) with supervised SVM and Multinomial Naive Bayes approaches based on bag-of-words features. The ssToT model yielded 68% averaged accuracy which is competitive with a fully supervised approach presented in the study.

It is important to note studies presented by CLPsych 2015 shared task competitors [4]. The shared task provided dataset which consists of text messages samples from Twitter that belongs to users with depression and PTSD. The best average precision (roughly 87%) on the depression vs control task was achieved by the method based on lexical features with tf-idf weighting [15]. Another team with a good result (86% averaged precision) proposed the method of terms clustering and formed the feature set using clusters of terms as N-grams [14].

CLPsych 2018 [9] focused on the two tasks: predicting 11 years old child's current psychological health from essays and predicting future psychological health form the same essays. Participant's submissions on the regression tasks were compared by the Pearson Product-Moment Correlation Coefficient. The best approach for both tasks used regularized linear regression with character and word-level n-gram features [5]. The second place on the first task was achieved by the team that utilized tf-idf and sentiment features with ensembles of different methods: ridge regression, SVMs, boosting, CNNs, RNNs, and feed-forward neural networks [26].

Clef/eRisk 2017 Shared Task [8] provided noisy dataset which consists of 887 Reddit user's messages collections, where 135 of the persons were identified as belonging to a risk case of depression. The best submitted classification model yielded 64% of F1-score by the team who applied an ensemble of tf-idf based classifiers on the data [22]. It is important to note that the same team reworked their models after Clef/eRisk 2017 Shared Task completion and reported 73% F1-score on the same train/test data by utilizing sophisticated linguistic metadata features (including LIWC) with logistic regression [23].

The analysis of related works reveals that it is hard to strictly compare the results of our research with others. The data and experiments design presented in these works differ from study to study and covering only the English-speaking population. The social media based datasets contain much more textual information for each person, but at the same time it much noisier than essays. But we can observe the methods and approaches that yield promising results on depression detection task. For both CLPsych 2018 and Clef/eRisk 2017 shared tasks the classic well-tuned n-gram and tf-idf based models outperform neural networks models. The specific attributes of depression mental state usually revealed through the depression related dictionaries, sentiment, and LIWC features. Although LIWC is a very popular and effective tool, there is no appropriate adaptation for the Russian language. It is also missing some psycholinguistic characteristics of the text.

3. Dataset

The dataset for the research contains two classes of texts: 93 essays written by people with depression (depression group) and 223 essays written by healthy people (control group). Depression essays were collected with the collaboration of two different teams of the psychologist who asked patients with depression disorder to write a short essay with a minimum length of 1800 characters. The control group essays were written mostly by students from different universities and different education programs (psychology, sociology, journalism, and information technology). Volunteers completed Russian adaptation of Beck Depression Inventory [1] to reveal their depression status and only essays written by persons who did not demonstrate depression signs were included in the control group (score less than 14 on the 0–63 Beck Depression Inventory scale). Thus, volunteers from the depression group did not complete the same questionnaire because their depression status was revealed by clinical experts with face to face diagnostic, which is superior to questionnaires. In the other hand, this fact forbids us from investigating this task as regression analysis. The topic of the essays is similar for both depression and control groups. We can generalize this topic as "Me and my relations with others and the world around me". Minimal age of volunteers for both groups is 18.

We should highlight two assumptions that we made around the depression group to perform classification on the dataset. First, the depressive disorder can be divided into many subtypes, and each form of depression has a different severity. Secondly, the part of the depression essays was written by patients who have already taken medications, but another part was written by untreated persons. In another hand, it is a very difficult task to collect a big number of essays from people with clinically confirmed depression. Utilizing machine learning tools require a sufficient number of training examples, which forced us to generalize all of the depression types as one and ignore the fact of medication use by depressed patients.

We present general statistics on the dataset in **Table 1**. It can be noted that generally, people with depression tend to write shorter texts. The mean age of the depression group is insufficiently higher than the control group. The gender distribution in the depression group is 59% females and 41% males. Gender distribution in the control group is 69% females and 31% males.

Table 1. Dataset statistics. Mean values and standard deviation

Value	Depression group	Control group
Number of essays	93	223
Age	28.05±10.67	22.82±10.01
Gender	55 Female, 38 Male	153 Female, 70 Male
Mean characters count	1883 ± 895	1994 ± 207
Mean words count	294.83 ± 145.89	317.75 ± 35.6
Mean sentence count	22.6 ± 11.28	23.11 ± 6.93

4. Methods

4.1. N-grams

Tf-idf and n-grams features are common natural language processing approach which performed well on depression detection task. Thus, we included 2 tf-idf based feature sets: unigrams and bigrams. N-grams that appeared less than in 1% of essays and more than in 90% of essays were removed from the feature sets. The result of t-SNE [10] on the bigrams data demonstrated in **Figure 1**.

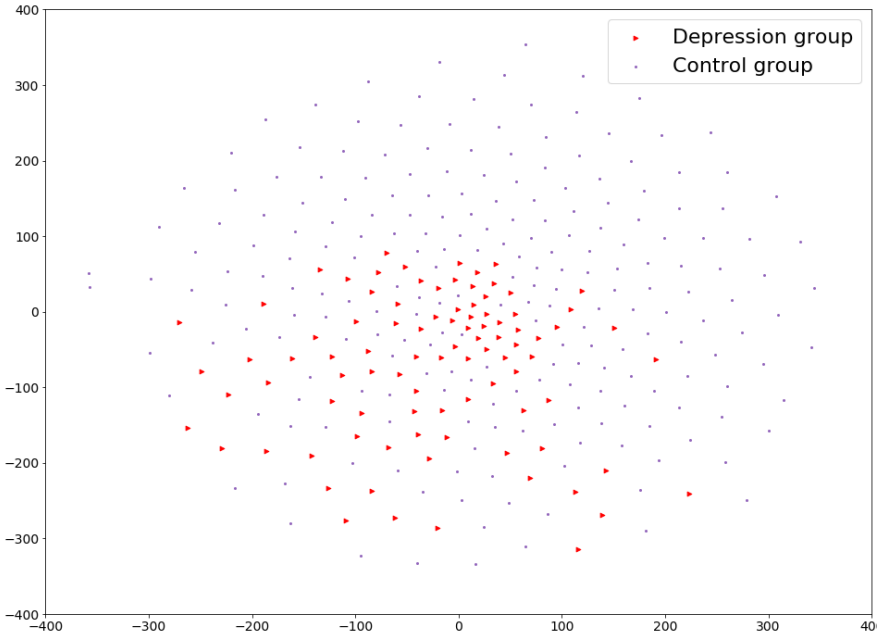

Fig. 1. Results of t-SNE applied on bigrams features

4.2. Depression markers

We annotated following feature set as Depression Markers (DM). It is containing classical text features (mean word/sentence length, POS-tags ratio) and psycholinguistic features. The part of psycholinguistic features described in following works [16], [18], [19], [25], and another part were proposed during the manual analysis of essays by psychologists and linguists. An important point is that most of these features were not previously tested on depression detection task.

To retrieve features from the essays we applied tokenization, lemmatization, and morphological analysis with MyStem. Statistics for DM presented in Table 2 (excluding POS-tags ratio). By observation of morphology related features, it can be noted that verbs and pronouns in various forms yield a lot of differences between groups. The smaller mean depth of syntax tree and mean number of words per sentence demonstrate a tendency of depression group to express themselves with shorter sentences. The *(N verbs) / (N adjectives)* ratio, which is also known as Trager coefficient, is usually differ from 1 among people with higher mental stress, which is also differ in our report among depression group [19].

Table 2. Mean+std for depression markers in depression and control group

Description	Depression group	Control group
Mean number of words per sentence	12.66±3.63	14.6±3.81
Mean number of characters per word	5.01±0.32	5.06±0.29
Lexicon: (N unique words) / (N words)	0.56±0.07	0.53±0.05
Average syntax tree depth	4.96±1.24	5.41±1.28
(N verbs) / (N adjectives)	1.36±0.45	1.11±0.30
(N verbs) / (N nouns)	0.5±0.12	0.5±0.08
(N participles) / (N sentences)	0.11±0.08	0.16±0.11
(N conjunctions + N prepositions) / (N sentences)	2.65±0.89	3.15±1.01
(N infinitives) / (N verbs)	0.23±0.07	0.28±0.08
(N singular first person past tense verbs) / (N verbs)	0.19±0.10	0.13±0.10
(N past tense verbs) / (N verbs)	0.69±0.09	0.62±0.09
(N first person verbs) / (N verbs)	0.2±0.11	0.15±0.10
(N third person verbs) / (N verbs)	0.18±0.08	0.25±0.10
(N first person pronouns) / (N pronouns)	0.56±0.14	0.45±0.15
(N singular first person pronouns) / (N pronouns)	0.53±0.15	0.35±0.19
(N plural first person pronouns) / (N pronouns)	0.01±0.02	0.08±0.08
(N words with wrong spelling) / (N sentences)	0.11±0.13	0.09±0.12
Sentiment rate	−1.31±6.36	3.91±6.62

4.3. Sentiment

Another valuable feature was computed with the help of Linis Crowd word sentiments dictionary [7]. The dictionary provides information about words estimated values of positive (1, 2), negative (−1, −2) or neutral (0) sentiment. We calculated the sentiment rate of each document by matching words from the essay with dictionary values and then summing it up. As demonstrated

in Figure 2, sentiment rates for each group greatly vary and have been included in the feature set. This value was included in the DM set.

5. Result of experiments

To perform classification, we utilized scikit-learn [12] implementation of random forest and SVM algorithms. Overall, we evaluate 5 different sets of features: depression markers (*DM*) that was described in **section 4.2**, tf-idf model computed on unigrams, tf-idf model computed on bigrams, and combination of n-gram models with depression markers. The classification report represented as averaged result of 5-fold cross-validation on the data (**Table 3**). Recall, precision, and F1-score calculated for the class of depression.

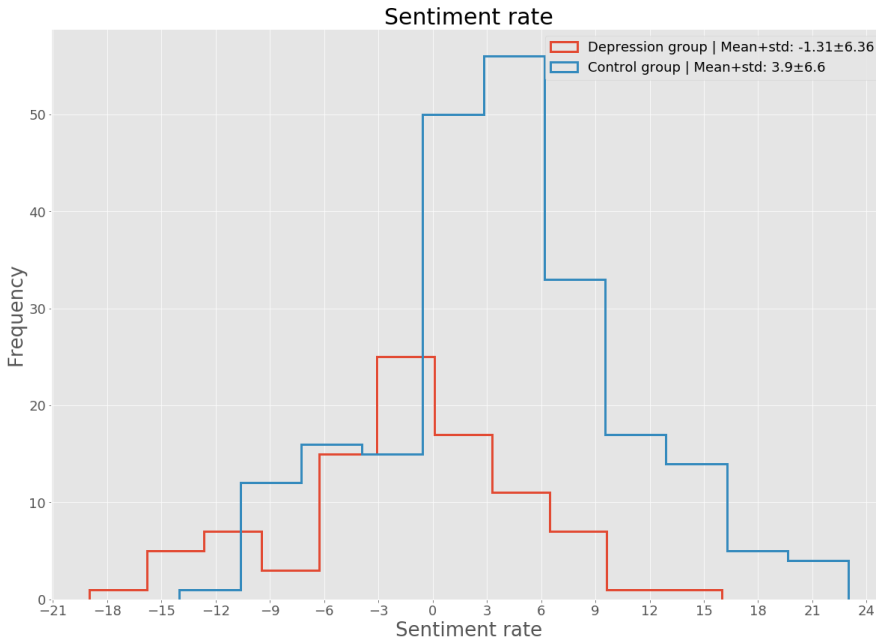

Fig. 2. Sentiment rate for depression and control groups

Depression markers model achieved best score in terms of overall classification performance on the data with 84% accuracy. N-grams based model also performed well with a F1-score around 70%. The combination of all models yielded best result for the task of depression essays classification with 73% F1-score. We relate the high values of standard deviation to the small number of samples in the data. The SVM based models are also yield best performance with bigrams features.

As it was mentioned before, it is hard to provide strict comparison with similar works, since the data format and language is different than in other studies. In terms of experiments design, studies related to the Clef/eRisk 2017 is the closest ones. The best reported F1-score for depression class on Clef/eRisk 2017 data is also 73% [23], which is close to F1-score in our experiments.

Table 3. Classification results

Feature set	Recall, %	Precision, %	F1, %	Accuracy, %
Random Forest				
DM	65.53 ± 8.31	**77.52** ± 4.91	70.65 ± 5.39	**84.16** ± 2.15
Unigrams	69.83 ± 9.36	70.87 ± 7.31	69.69 ± 4.51	82.27 ± 2.40
Unigrams + DM	72.01 ± 10.03	70.7 ± 10.18	70.48 ± 5.36	82.28 ± 3.40
Bigrams	**76.26** ± 7.37	70.42 ± 5.69	72.72 ± 2.44	83.21 ± 1.77
Bigrams + DM	74.18 ± 3.11	72.12 ± 4.16	**73.01** ± 2.11	83.85 ± 1.46
SVM				
DM	78.66 ± 14.27	52.78 ± 5.63	62.96 ± 8.12	73.11 ± 5.44
Unigrams	49.59 ± 16.11	69.04 ± 5.80	56.60 ± 12.67	78.81 ± 4.29
Unigrams + DM	72.28 ± 15.01	61.00 ± 11.09	66.05 ± 12.51	78.21 ± 8.17
Bigrams	64.67 ± 11.74	72.42 ± 8.89	68.01 ± 9.38	82.29 ± 5.04
Bigrams + DM	65.76 ± 11.54	69.16 ± 7.05	67.20 ± 8.76	81.35 ± 4.62

6. Conclusion

The study evaluates the ability of machine learning models and several types of feature sets to perform classification on essays in Russian written by depressed and healthy peoples. The depression markers that was described in the paper, as well as standard NLP approaches like unigrams and bigrams, demonstrated good performance on the data. The Bigrams+DM feature set achieved the best results for the task of revealing depression essays with 73% F1-score. It was discovered that applying word sentiment dictionaries as Linis Crowd is suitable for the depression detection task. We considering this study as a first step in the machine learning based depression detection from texts in Russian.

We currently looking forward to investigating the ability of word embeddings and neural networks models to identify depression in human writings. The dataset possibly will become public-available for research purposes in the fully anonymized format. As a general idea for future work, we planning to apply depression detection methods on Russian-speaking social networks.

Acknowledgments

The publication has been prepared with the support of the “RUDN University Program 5-100” and funded by RFBR according to the research projects №17-29-02225 and №17-29-02305.

References

1. *Beck, A. T., Steer, R. A., & Carbin, M. G.* (1988). Psychometric properties of the Beck Depression Inventory: Twenty-five years of evaluation. Clinical psychology review, 8(1), 77–100.
2. *Bucci, W., & Freedman, N.* (1981). The language of depression. Bulletin of the Menninger Clinic, 45(4), 334.
3. *Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K.* (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 1–10).
4. *Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M.* (2015). 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).
5. *Çöltekin, Ç., & Rama, T.* (2018). Tubingen-Oslo system: Linear regression works the best at Predicting Current and Future Psychological Health from Childhood Essays in the CLPsych 2018 Shared Task. arXiv preprint arXiv:1809.04838.
6. *De Choudhury, M., Counts, S., & Horvitz, E.* (2013, May). Social media as a measurement tool of depression in populations. In Proceedings of the 5th Annual ACM Web Science Conference (pp. 47–56). ACM.
7. *Koltsova, O. Y., Alexeeva, S., & Kolcov, S.* (2016). An opinion word lexicon and a training dataset for russian sentiment analysis of social media. Computational Linguistics and Intellectual Technologies: Materials of DIALOGUE 2016 (Moscow), 277–287.
8. *Losada, D. E., & Crestani, F.* (2016, September). A test collection for research on depression and language use. In International Conference of the Cross-Language Evaluation Forum for European Languages (pp. 28–39). Springer, Cham.
9. *Lynn, V., Goodman, A., Niederhoffer, K., Loveys, K., Resnik, P., & Schwartz, H. A.* (2018). 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 (pp. 37–46).
10. *Maaten, L. V. D., & Hinton, G.* (2008). Visualizing data using t-SNE. Journal of machine learning research, 9(Nov), 2579–2605.
11. *Moussavi, S., Chatterji, S., Verdes, E., Tandon, A., Patel, V., & Ustun, B.* (2007). Depression, chronic diseases, and decrements in health: results from the World Health Surveys. The Lancet, 370(9590), 851–858.
12. *Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J.* (2011). Scikit-learn: Machine learning in Python. Journal of machine learning research, 12(Oct), 2825–2830.
13. *Pennebaker, J. W., Francis, M. E., & Booth, R. J.* (2001). Linguistic inquiry and word count: LIWC 2001. Mahway: Lawrence Erlbaum Associates, 71(2001), 2001.
14. *Preoţiuc-Pietro, D., Sap, M., Schwartz, H. A., & 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: From Linguistic Signal to Clinical Reality (pp. 40–45).

15. *Resnik, P., Armstrong, W., Claudino, L., & Nguyen, T.* (2015). The University of Maryland CLPsych 2015 shared task system. In Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality (pp. 54–60).
16. *Samokhvalov V. P.* (2002), Psychiatry [Psihiatriya], Phoenix, Rostov-on-Don.
17. *Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Dziurzynski, L., Ramones, S. M., Agrawal, M., ... & Ungar, L. H.* (2013). Personality, gender, and age in the language of social media: The open-vocabulary approach. PloS one, 8(9), e73791.
18. *Shvedovsky E. F., Zvereva N. V.* (2015). Studies of speech disorders in schizophrenia. History and state of-the-art [Исследование речевых нарушений при шизофрении. История и современное состояние проблемы]. Psychological Science and Education, 20(2), 78–92.
19. *Smirnova, D. A.* (2010). Clinical and psycholinguistic characteristics of mild depression [Клинические и психолингвистические характеристики легких депрессий] (Doctoral dissertation, Moscow Research Institute of Psychiatry).
20. *Tausczik, Y. R., & Pennebaker, J. W.* (2010). The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24–54.
21. *Thornicroft, G., Chatterji, S., Evans-Lacko, S., Gruber, M., Sampson, N., Aguilar-Gaxiola, S., ... & Bruffaerts, R.* (2017). Undertreatment of people with major depressive disorder in 21 countries. The British Journal of Psychiatry, 210(2), 119–124.
22. *Trotzek, M., Koitka, S., & Friedrich, C. M.* (2017, September). Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression. In CLEF (Working Notes).
23. *Trotzek, M., Koitka, S., & Friedrich, C. M.* (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. IEEE Transactions on Knowledge and Data Engineering.
24. *Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., ... & Sheth, A.* (2017, July). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017 (pp. 1191–1198). ACM.
25. *Zagorovskaya, O. V., Litvinova, O. A., & Litvinova, T. A.* (2016). Identify the tendency of the individual to suicidal behavior on the basis of a quantitative analysis of speech production [Выявление склонности личности к суицидальному поведению на основе количественного анализа ее речевой продукции.]. Studia Humanitatis, (1).
26. *Zaporojets, K., Sterckx, L., Deleu, J., Demeester, T., & Develder, C.* (2018). Predicting Psychological Health from Childhood Essays. The UGent-IDLab CLPsych 2018 Shared Task System. In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (pp. 119–125).