UniOR NLP at eRisk 2021: Assessing the Severity of Depression with Part of Speech and Syntactic Features

Raffaele Manna[1], Johanna Monti[1]

[1] "L' Orientale" University of Naples - UNIOR NLP Research Group, Naples, Italy

Abstract

This paper describes the participation of the UniOR NLP Research Group team in task 3 (T3) within the CLEF eRisk 2021 lab. We report the approaches used to address eRisk 2021 T3, which aims to measure the severity of the signs of depression in social media users. This year's eRisk T3 consists of exploring methods for automatically filling out a 21-question depression questionnaire, namely Beck's Depression Inventory (BDI). We explored and tried different combinations of text pre-processing and feature extraction steps in order to grasp self-referential pieces of text and two main methods for representing the text features as input data for traditional machine learning classifiers.

Keywords

Natural Language Processing, Machine Learning, Topic Modeling, Sentence Embeddings, Mental Health Risk Assessment

1. Introduction

Recent developments in Natural Language Processing (NLP) have led to the use of textual data from social media as sources for early monitoring and identification of risky behaviors associated with mental health problems. With the advent and construction of sub-communities around specific health issues on different social media platforms, users often come together to talk and pour out freely about their inner moods and feelings [1, 2]. Among these social communities, the attention of researchers in NLP and digital psychiatry [3] has focused on social media sub-communities, in which users succeed and interact about their mental and physical health status. Research on mental health status turned to mine specific sub-communities associated with some mental health conditions. As an example of this, the social media Reddit[1] involves several sub-communities (subreddits) associated with different mental health conditions, in which users offer or seek support and describe their behaviors and moods [4, 5]. Data from these communities could represent an opportunity to improve technologies related to health and well-being. Specifically, much of the research focused on the automatic processing and monitoring of social media data in order to recognize signals associated with depressive status and to assess related symptomatology, often underdiagnosed and undertreated [6].

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

rmanna@unior.it (R. Manna); jmonti@unior.it (J. Monti)

[1] https://www.reddit.com/

Since 2017, the CLEF eRisk lab[2] organizes a series of shared tasks focusing on the exploration and evaluation of NLP applications concerning health and safety issues of users on social media, thus offering a forum where scholars can compare different methods on the same data in relation to the issue under investigation [7].

Since then, eRisk lab focused around specific mental health issues and risky behavior of users on social media. In 2017, an exploratory task on the early identification of depression risk was launched. In 2018, besides this latter task, a further task was added on the identification and early risk prediction of signs of anorexia; and, during the 2019 edition, researchers investigated automatic methods to recognize early signs and linguistic traces for the prediction of risk associated with self-harm in social media texts. Starting with the 2019 edition, a task was proposed consisting in the exploration of methodologies for estimating and measuring the level of severity of depression from a thread of user texts.

Considering the issues investigated during those lab editions, eRisk lab represents an opportunity to experiment with reliable tools and applications to help clinicians in their decisions and improve health intervention and prevention strategies for the living conditions of struggling people.

This paper describes the participation of the UniOR NLP team in the third task organized by CLEF eRisk 2021,namely the measuring of the severity of signs of depression. In this task, the team's system has to fill in a questionnaire on depression, that is the BDI, starting from the entire history of social media posts for each user considered in the dataset. We submitted four variants of two models, in which we experimented with a series of pre-processing steps (lemmatization, Part-of-Speech tagging and dependency parsing), two methods for representing textual data (topic modeling and pre-trained models for sentence embeddings) and standard machine learning algorithms. The paper is organized as follows: in Section 2 we summarize the related work, in Section 3 we present the T3 dataset provided by the shared task organizers, in Section 4 we describe the methodology used and then the results obtained are presented in Section 5. Finally, in Section 6 we outline the conclusions.

2. Related Work

The last decade has seen an increase in research around mental health issues attempting to understand and model mental states by exploring data and the language used on social media [8, 9]. In this research field, there have been numerous studies that explore the possibility of using NLP techniques and artificial intelligence (AI) applications for the screening and early prediction of mental health problems [10, 11, 12]. Moreover, research in this field promises many benefits related to the monitoring of patients suffering from mental disorders and the timely intervention in dangerous behaviors such as self-harm, suicidal thoughts and suicide attempts [13, 14].

For these reasons, the research in NLP focused on the development of language technologies capable of identifying linguistic and behavioral cues associated with various mental disorders among which depression, schizophrenia, eating disorders including anorexia, risk of self-harm and suicidal thoughts using as a source of information data from social media communities.

[2] https://erisk.irlab.org/

Among the social media platforms, Twitter, Facebook and Reddit are the most leveraged platforms to derive large scale textual data in this research field.

Thus exploiting these platforms, several researches in NLP focused on the construction of datasets and the consequent analysis of linguistic phenomena and emotional states related to depressive states in tasks of identifying signs of depression. In these researches, the textual production on social media of users with this mental illness was compared both with texts written and posted by control users and also compared with users suffering from other mental disorders. Regardless of the social media platform investigated and with reference to the problem of identifying linguistic signs in depressive episodes and status, the majority of research experiences using NLP techniques highlighted linguistic characteristics and contents in relation to the expression of negative emotions and topics related to the affective and sentimental sphere.

Specifically for linguistic use and style, researches highlighted an increase in the use of the first person personal pronoun [15, 16, 17, 18] and linguistic factors related to the readability and coherence of the texts [19]. Regarding the linguistic contents grouped into topics, NLP studies revealed a focus on topics related to personal concerns, interpersonal relationships and work using topic modeling techniques and LIWC categories [20, 21, 22]. While other studies used sentiment lexicons such as ANEW, Vader [23, 24] and domain-specific lexicons [25, 26, 27] to explore the feelings, emotions and moods involved in users' writing histories related to depression.

3. T3 - Measuring the severity of the signs of depression

This task was launched during the 2019 edition of the eRisk lab[3]. T3 consists in the exploration of possible automatic methods for estimating the level of depression severity in relation to the symptoms identified in Beck's Depression Inventory Questionnaire (BDI) [28, 29]. Considering the writing history of a user who has compiled the BDI as ground truth, this task expects these symptoms to be automatically recognized and scored by a system. The BDI questionnaire includes and assess 21 symptoms associated with the level of depression. The symptoms can manifest themselves both in the user's mood and feelings such as sadness, pessimism, guilty feelings and in symptoms that can alter some behaviors such as crying, loss of energy and changes in sleeping pattern. In the BDI questionnaire, these symptoms correspond to questions associated with answers with numerical values ranging from 0 to 3, with the exception of two questions - Changes in Sleeping Pattern and Changes in Appetite - which are associated with the following values: 0, 1a, 1b, 2a, 2b, 3a or 3b.

3.1. Dataset Description

The training dataset distributed during T3 at eRisk 2021 [30] is composed of the writing histories and the filled golden truth questionnaires from the data used during the 2019 and 2020 editions of this task [31, 32]. This dataset consists of 90 .xml files corresponding to each subject together with their writing history and the datetime of each post produced. Two ground truth files were

[3]https://erisk.irlab.org/2019/index.html

provided containing each user's responses to individual questions in the BDI questionnaire.

Instead, the dataset used to test the submitted systems is composed of 80 .xml files corresponding to each subject together with their writing history and the datetime of each text posted on Reddit.

3.2. Metrics

To evaluate the systems submitted in this task, the automatic filling of the questionnaire predicted by the system is compared with the questionnaire filled in by the actual subject, considering both the scores associated with each of the 21 questions and also considering the sum of these scores for each subject in the test set [30]. Specifically, in T3 each submitted system is evaluated for its performance in predicting the individual scores associated with each of the 21 symptoms and it is also assessed the ability to predict the levels of depression defined as the sum of the scores of all 21 questions for each subject. Depression levels are associated with the following four categories: *minimal depression* with the sum of scores ranging from 0 to 9; *mild depression* ranging from 10 to 18; *moderate depression* ranging from 19 to 29; and *severe depression* from 30 to 63.

The following metrics are considered in evaluating the performance of a system in T3:

- **Average Hit Rate** (AHR) - That is, Hit Rate (HR) averaged over all users. This metric computes the ratio of cases in which the scores predicted by the system are exactly the same as those present in the real questionnaire.
- **Average Closeness Rate** (ACR) - Closeness Rate averaged over all users. This metric computes the absolute difference between the score predicted by the system and the real score for each of the questions contained in the BDI.
- **Average Difference between Overall Depression Levels** (ADODL) - This metric calculates the absolute difference between the levels of depression - obtained with the sum of the scores associated with each of the 21 questions - predicted by the system against the real levels.
- **Depression Category Hit Rate** (DCHR) - It measures the fraction of cases in which the questionnaire filled by the system leads to a category of depression equivalent to the actual one of the subject.

4. Methodology

In this section, we describe the pre-processing steps carried out on the textual data provided by the organizers along with the data representations methods we used to discover pieces of text in which the user describes himself or herself as the subject of an action or experience - linguistically realized through the personal pronoun subject - and as an object realized through the personal pronoun object. In this context, we rely solely on the linguistic and grammatical information available in the data in order to represent self-referential textual segment.

In that, our goal is twofold: 1) to explore the information carried in those textual segments by exploiting grammatical categories and simple pre-processing techniques; and 2) once isolated these portions, to try to assess the severity of the depression representing the data both at the

word-level and at phrase-level. In order to accomplish the last point, i) we used LDA model [33] to get topic distributions over a vocabulary composed of nouns and adjectives and ii) we leveraged sentence embedding models for phrases.

We addressed T3 as a multi-class classification problem, by training and testing the models selected for each of the 21 questions in BDI, where each question can be assigned one of 4 labels with the exception of two questions with 6 possible labels.

4.1. Pre-processing

In order to prepare and extract the textual portions of our interest, we first concatenated together every post for each user in training dataset into larger documents and, then, we performed the following pre-processing steps:

- We transformed the Reddit posts in lower case
- We removed URLs, subreddits mentions and any posts *[removed]* using regular expressions
- We solved the English contractions for modal, auxiliary and negation
- We removed digits, single character tokens and punctuation except for the first person pronoun and the period mark respectively.

These steps were performed exactly in the order in which they are reported in order to apply then the sentence tokenizer, the PoS tagger and the dependency parser. The last three steps are performed using the English model [4] available for the Spacy library [5] in Python.

As first step in selecting linguistic elements for further processing, we used the PoS tagger to extract nouns and adjectives occurring in sentences in which "I" or "me" appears, for each transformed sentence in writing history. Moreover, we also added verbs and modifiers to the grammatical categories already extracted. Then, we used the dependency parser to segment and extract relations in sentences in which the self-referential user ("I" or "me") appears as in one of the following syntactic role: subject, direct object or indirect object.

Respectively, the features extracted with PoS tagger were used as input data for the first two models, UniorNLP_A and UniorNLP_B. While, the sentences extracted with syntactic relations were passed to UniorNLP_C and UniorNLP_D.

4.2. Models

Here we describe the data representation methods performed on the input linguistic features and the classification algorithms tested during T3.

We opted for Latent Dirichlet Allocation (LDA) model to represent the features deriving from the PoS tagger. For this model, different configurations were tested together with different supervised classification algorithms implemented in scikit-learn [6] using k-fold cross-validation with K value set at 4. Instead, each of the segmented sentences from the dependency parser were encoded using pre-trained models available in the Sentence-Transformers [7] library, then

[4] https://github.com/explosion/spacy-models/releases//tag/en_core_web_lg-3.0.0
[5] https://spacy.io/
[6] https://scikit-learn.org/stable/index.html
[7] https://www.sbert.net/index.html

these sentence representations were passed as inputs to several supervised classification algorithms, selecting the best performing model based on k-fold cross-validation with K value set at 4.

The best configurations of the models trained and tested are described below in detail:

- **UniorNLP_A** In this model, we represented nouns and adjectives extracted from sentences with an LDA model based on words unigrams. After training the LDA model on these data, a matrix of vectors is generated with the distribution of topics corresponding to each user. Then, we trained several classification algorithms on these vectors including logistic regression, support vector machine, ensamble classifiers and gaussian classifier. Considering the metrics reported above, we used k-fold cross-validation method for hyperparameter tuning of the classifiers and for the LDA model. The best performances were found using a gaussian process classifier [34] and 25 topics.
- **UniorNLP_B** In this model, we represented nouns, adjectives, verbs and modifiers extracted from sentences with an LDA model based on word unigrams and bigrams. We further removed the words unigrams and bigrams occurring less than three times. Another LDA model was trained on these data mapping to a vector of topics distribution each user. We trained several classification algorithms on these vectors including logistic regression, support vector machine, ensamble classifiers and gaussian classifiers. We used the k-fold cross-validation method for hyperparameter tuning of the classifiers and for the LDA model. In this configuration, the best performances were found using a logistic regression classifier with newton-cg solver and 28 topics.
- **UniorNLP_C** In this model, we applied a general purpose pre-trained model from Sentence-Transformers library[8] to embed each user sentences. Then, we applied a K-mean clustering to obtain group of similar sentences. We tested several classification algorithms on these vectors including logistic regression, support vector machine, ensemble classifiers and gaussian classifier. We used the k-fold cross-validation method for hyperparameter tuning of the classifiers and to choose the number of clusters. The best performance was reached by the extra-trees classifier with the number of clusters equal to 10.
- **UniorNLP_D** In this model, we repeated the previous configuration without using the clustering algorithm. We concatenated each of the sentence representations into a 300-dimensional vector for each user. These representations were passed to the extra-trees classifier.

5. Results

In this task the participants are asked to send a results file for each system tested in which each row corresponds to the predicted scores for each of the 21 questions of the BDI.
We report our offical results for T3 for all evaluation metrics along with the best results obtained in T3 for each metric. In Table 1 are shown the results achieved by our systems on the test set.

[8]We explored the following pre-trained models applying the average word embedding for some well-known word embedding methods: *Average_word_embeddings_glove.6B.300d*; *average_word_embeddings_levy_dependency* and *average_word_embeddings_komninos*. The best results were obtained using the last model in this list.

Table 1
Evaluation of Unior NLP's submissions in Task 3. The best results for each metric were added for comparison.

	AHR	ACR	ADODL	DCHR
Best Scores	35.36%	73.17%	83.59%	41.25%
UniorNLP_A	31.67%	63.95%	69.42%	08.75%
UniorNLP_B	31.61%	64.66%	74.74%	15.00%
UniorNLP_C	28.63%	63.31%	76.45%	20.00%
UniorNLP_D	28.10%	64.25%	71.27%	15.00%

6. Conclusions and Future Work

In this paper we presented the contributions of the Unior NLP team in the shared task 3 organized at eRisk 2021: Measuring the severity of the signs of depression.

We investigated the linguistic content in reference to textual portions in which the user places himself in a position of self-focus. In this context, we tried to exploit the linguistic knowledge associated with some grammatical and syntactic characteristics contained in the users' writing histories. We tried two well-known methods of representing linguistic data and several supervised classification algorithms. The results obtained are well below the best performances recorded in this edition of T3, especially for the DCHR metric.

Considering the potential of automated methods for recognizing symptoms and risky behaviors associated with depression in social media texts, we plan to further analyze and model the complex linguistic and social factors involved in the textual expression of mental disorders.

Acknowledgments

This research has been carried out in the context of an innovative industrial PhD project in computational stylometry supported by the POR Campania FSE 2014-2020 funds.

References

[1] J. A. Naslund, K. A. Aschbrenner, L. A. Marsch, S. Bartels, The future of mental health care: peer-to-peer support and social media, Epidemiology and psychiatric sciences 25 (2016) 113–122.

[2] B. S. Fraga, A. P. C. da Silva, F. Murai, Online social networks in health care: a study of mental disorders on reddit, in: 2018 IEEE/WIC/ACM International Conference on Web Intelligence (WI), IEEE, 2018, pp. 568–573.

[3] C. Burr, J. Morley, M. Taddeo, L. Floridi, Digital psychiatry: Risks and opportunities for public health and wellbeing, IEEE Transactions on Technology and Society 1 (2020) 21–33.

[4] M. De Choudhury, S. De, Mental health discourse on reddit: Self-disclosure, social support, and anonymity, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 8, 2014.

[5] A. Park, M. Conway, Harnessing reddit to understand the written-communication challenges experienced by individuals with mental health disorders: analysis of texts from mental health communities, Journal of medical Internet research 20 (2018) e121.
[6] G. Thornicroft, S. Chatterji, S. Evans-Lacko, M. Gruber, N. Sampson, S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, L. Andrade, G. Borges, et al., Undertreatment of people with major depressive disorder in 21 countries, The British Journal of Psychiatry 210 (2017) 119–124.
[7] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 28–39.
[8] R. A. Calvo, D. N. Milne, M. S. Hussain, H. Christensen, Natural language processing in mental health applications using non-clinical texts, Natural Language Engineering 23 (2017) 649–685.
[9] S. Chancellor, M. De Choudhury, Methods in predictive techniques for mental health status on social media: a critical review, NPJ digital medicine 3 (2020) 1–11.
[10] C. Karmen, R. C. Hsiung, T. Wetter, Screening internet forum participants for depression symptoms by assembling and enhancing multiple nlp methods, Computer methods and programs in biomedicine 120 (2015) 27–36.
[11] A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. Devylder, M. Walter, S. Berrouiguet, et al., Machine learning and natural language processing in mental health: Systematic review, Journal of Medical Internet Research 23 (2021) e15708.
[12] Y. Liang, X. Zheng, D. D. Zeng, A survey on big data-driven digital phenotyping of mental health, Information Fusion 52 (2019) 290–307.
[13] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, M. Kumar, Discovering shifts to suicidal ideation from mental health content in social media, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 2098–2110.
[14] A. B. Shatte, D. M. Hutchinson, S. J. Teague, Machine learning in mental health: a scoping review of methods and applications, Psychological medicine 49 (2019) 1426–1448.
[15] R. S. Campbell, J. W. Pennebaker, The secret life of pronouns: Flexibility in writing style and physical health, Psychological science 14 (2003) 60–65.
[16] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: Proceedings of the International AAAI Conference on Web and Social Media, volume 7, 2013.
[17] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, E. J. Langer, Forecasting the onset and course of mental illness with twitter data, Scientific reports 7 (2017) 1–11.
[18] M. Trotzek, S. Koitka, C. M. Friedrich, Word embeddings and linguistic metadata at the clef 2018 tasks for early detection of depression and anorexia., in: CLEF (Working Notes), 2018.
[19] Z. Jamil, D. Inkpen, P. Buddhitha, K. White, Monitoring tweets for depression to detect at-risk users, in: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality, 2017, pp. 32–40.
[20] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, J. Boyd-Graber, Beyond lda: exploring supervised topic modeling for depression-related language in twitter, in:

Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, 2015, pp. 99–107.
[21] T. Nalabandian, M. Ireland, Depressed individuals use negative self-focused language when recalling recent interactions with close romantic partners but not family or friends, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology, 2019, pp. 62–73.
[22] M. M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (2019) 44883–44893.
[23] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, M. Berk, Affective and content analysis of online depression communities, IEEE Transactions on Affective Computing 5 (2014) 217–226.
[24] K. Loveys, P. Crutchley, E. Wyatt, G. Coppersmith, Small but mighty: affective micropatterns for quantifying mental health from social media language, in: Proceedings of the fourth workshop on computational linguistics and clinical Psychology—From linguistic signal to clinical reality, 2017, pp. 85–95.
[25] F. Sadeque, D. Xu, S. Bethard, Measuring the latency of depression detection in social media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, 2018, pp. 495–503.
[26] A. Trifan, J. L. Oliveira, Bioinfo@ uavr at erisk 2019: delving into social media texts for the early detection of mental and food disorders., in: CLEF (Working Notes), 2019.
[27] P. Van Rijen, D. Teodoro, N. Naderi, L. Mottin, J. Knafou, M. Jeffryes, P. Ruch, A data-driven approach for measuring the severity of the signs of depression using reddit posts., in: CLEF (Working Notes), 2019.
[28] A. T. Beck, C. Ward, M. Mendelson, J. Mock, J. Erbaugh, Beck depression inventory (bdi), Arch Gen Psychiatry 4 (1961) 561–571.
[29] A. T. Beck, R. A. Steer, G. Brown, Beck depression inventory–ii, Psychological Assessment (1996).
[30] J. Parapar, P. Martin-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2021: Early risk prediction on the internet., in: Proceedings of the Twelfth International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, Cham, 2021, September, p. pp. tbp.
[31] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2019 early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2019, pp. 340–357.
[32] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk 2020: Early risk prediction on the internet, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 272–287.
[33] D. M. Blei, A. Y. Ng, M. I. Jordan, Latent dirichlet allocation, the Journal of machine Learning research 3 (2003) 993–1022.
[34] C. E. Rasmussen, Gaussian processes in machine learning, in: Summer school on machine learning, Springer, 2003, pp. 63–71.