

UNED-MED at eRisk 2022: depression detection with TF-IDF, linguistic features and Embeddings

Elena Campillo-Ageitos¹, Juan Martinez-Romo^{1,2} and Lourdes Araujo^{1,2}

¹NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Juan del Rosal 16, Madrid 28040, Spain

²IMIENS: Instituto Mixto de Investigación, Escuela Nacional de Sanidad, Monforte de Lemos 5, Madrid 28019, Spain

Abstract

Mental health problems, such as depression and anxiety, are conditions that can have very serious consequences if untreated, and cause the patient a lot of suffering. Research suggests that the way people write can reflect mental well-being and mental health risks, and social media provides a source of user-generated text to study. Early detection is crucial for mental health problems, and with this in mind the shared task eRisk was created. This paper describes the participation of the group UNED-MED on the 2022 T2 subtask. Participants were asked to create systems to detect early signs of depression on users from Reddit. Our team proposes two approaches: feature-driven classifiers with features based on text data, TF-IDF terms, first-person pronoun use, sentiment analysis and depression terminology; and a Deep Learning classifier with pretrained Embeddings. The official task results show modest results that show the difficulty of working with depression data.

Keywords

early risk detection, depression detection, natural language processing, data extraction, data relabeling, CEUR-WS

1. Introduction

Mental health problems such as anxiety and depression are conditions that affect millions of people every year. People with depression may not seek medical attention in time, causing them unnecessary suffering. Some patients forego medical attention because they are not aware that they need it, but some still avoid it because of the stigma associated with it. Whatever the cause, untreated mental illnesses can worsen with time and lead to serious consequences, such as substance abuse, or even death.

Language is a tool we use to communicate with one another. Besides transmitting the intended message with it, we also transmit information about ourselves: our upbringing, our mood, our emotional well-being, etc. Several studies have shown a correlation between differences in the way people talk and write, and having a mental health condition [1, 2]. This use of language can be studied with Natural Language Processing (NLP) to detect untreated mental health problems.


CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy

✉ ecampillo@lsi.uned.es (E. Campillo-Ageitos); juaner@lsi.uned.es (J. Martinez-Romo); lurdes@lsi.uned.es (L. Araujo)

🆔 0000-0003-0255-0834 (E. Campillo-Ageitos); 0000-0002-6905-7051 (J. Martinez-Romo); 0000-0002-7657-4794 (L. Araujo)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Social media sites such as Twitter, Tumblr or Reddit present a natural collection of user-generated texts. They use these sites to communicate with their friends, follow celebrities, and to express themselves. There is an insurmountable amount of information that can be applied to NLP techniques with various purposes. Recent research has been applying these techniques to automatically detect users who suffer from several mental health issues.

In the case of mental health, early treatment is especially important, since it increases the probabilities of a good prognosis. The longer a patient suffers without medical treatment, the more likely they are to suffer from associated risks. Early detection helps detect these cases before they become bigger problems. Most work done in the literature focus on detecting people who already suffer from conditions, but we believe focusing on early detection is important to allow for a faster diagnostic and faster intervention.

The eRisk shared task was created with this objective in mind. The shared task focuses on the early detection of mental health problems in social networks. Previous editions have focused on anorexia, self-harm, and pathological gambling. The 2022 eRisk shared task proposed three subtasks, and this paper presents our participation to the subtask T2: early detection of depression. An overview of the task and the overall results for all participating teams can be found on [3].

The following sections are organized as follows: Section 2 describes the data we used to train our models; Section 3 details the models we proposed for this task; Section 4 explains the experiment setup; Section 5 summarizes our results after the test stage; finally, Section 6 presents our conclusions and ideas for future work.

2. Dataset Description

Our system for the T2 task was trained and evaluated with two different datasets: 1) The official eRisk 2022 shared Task 2 dataset, and 2) The UNED-MED depression Reddit dataset. This section briefly describes each dataset.

2.1. eRisk 2022 shared Task 2 dataset

This is the official eRisk dataset given to the participants. It is a Reddit early risk detection dataset first presented by Losada et al. at [4]. The dataset is comprised of a collection of documents. Each document contains the post submission history of a user from Reddit. The users are labeled as one of two classes: Positive (at risk of depression) and negative (control group).

This year, the training data was formed with the training and testing data of 2017 and 2018 eRisk depression tasks. We decided to use the data from 2017 to train our model, and the data from 2018 for evaluation.

Each post in a user document is either a text submission or a comment to another users' submission. The posts are ordered chronologically, from earliest to latest. The number of posts for an user is indeterminate, as is the length of each post. Additional data comes with each post, such as the date of submission.

The classes are deeply unbalanced, as can be seen in table 1 represented by the Original eRisk column: there are 7 times more negative users than positive users. This can be a challenge

when training classifiers, especially when the most important class is the least represented one.

Table 2 shows some statistics about the dataset. The average length of messages for positive users is 206, while for negative users it is 170. The 25%, 50%, and 75% values show that, while below the median, most posts are short for both groups of users, positive users tend to write longer posts, even if the longest post was written by a negative user.

One brief exploration of the data shows some challenges from the textual data. This is no surprise, since they are not formal texts; they are sentences written by people on the Internet to communicate between each other. It is widely observed that Internet speech has deep particularities, such as loose grammar, incorrect spelling (sometimes in purpose, and with meaning), emoji use, etc. A whole paper could be written on social cues only observed on Internet speech. We also find metadata, such as hyperlinks, references to other users, etc.

2.2. UNED-MED 2022 depression Reddit dataset

One of the most challenging aspects of the official dataset is the deeply unbalanced aspect of the classes. Detecting positive users is arguably much more important than detecting negative users, but they are underrepresented in the training dataset. We have curbed this problem in previous editions of the shared task by applying data oversampling. In this occasion, we obtained additional data from Reddit, following the strategies described in [4].

We used the PRAW Python Reddit API Wrapper to extract new data ¹. We searched Reddit submissions with the following search queries:

- diagnosed AND depression
- I AND am AND diagnosed AND depression

We searched on r/all, and on the following subreddits related to mental health and depression:

- r/addiction
- r/adultdepression
- r/anxiety
- r/anxietyhelp
- r/depression
- r/depression_help
- r/mentalhealth
- r/mentalillness
- r/sad
- r/suicidewatch

Results were manually reviewed to make sure users were talking about themselves, and had been officially diagnosed with clinical depression. A list was compiled, and we obtained the post and comment history of each user. Users with less than ten submissions were discarded.

The resulting dataset is a collection of 235 users. Table 2 shows some statistics for the resulting dataset. We can see that the statistics resemble those of the positive users of the original eRisk dataset.

¹<https://praw.readthedocs.io/en/stable/>

Table 1

Breakdown of the Original eRisk 2022 dataset’s number of positive and negative users, plus the additional extracted data.

	Original eRisk	New eRisk	Combined
Positive users	214	235	449
Negative users	1493	0	1493
All users	1707	235	1942

Table 2

Analysis of the messages text length of the Original eRisk 2022 dataset and the UNED-MED 2022 dataset, represented as New eRisk

	Original eRisk	Original positive	Original negative	New eRisk
count	1076582	90222	986360	106588
mean	172.91	205.54	169.92	208.87
std	538.48	398.54	549.41	481.77
min	1	1	1	1
25%	40	40	40	33
50%	73	90	72	79
75%	160	213	156	199
max	38663	18175	38663	31638

3. Proposed Model

We present three early risk models, depending on the classifier algorithm we use: 1) Random Forest, 2) XGBoost, and 3) CNN. Models 1 and 2 are based on traditional machine learning techniques, while model 3 applies Deep Learning. Features for models 1 and 2 are a combination of TF-IDF and text-based. Features for model 3 were Embeddings.

Each model is conformed of three stages: 1) Data pre-processing, 2) features, and 3) classification. The models take one message by one user and predict whether this user is at risk of depression (1) or not (0). As is established by the eRisk task, a positive decision is final, but a negative decision may be rectified later.

3.1. Training data

Based on the data described in section 2, we created three different training sets:

- Original eRisk: This training set was formed by combining the eRisk 2022 shared Task 2 train and test datasets.
- Augmented eRisk: This training set was created by incorporating the UNED-MED 2022 depression Reddit dataset to the Original eRisk training set.
- Relabeled eRisk: This training set was created by applying a relabelling strategy based on sentiment analysis to the Original eRisk training set.

Table 3

Analysis of the messages text length of the Original eRisk 2022 dataset after applying relabeling.

	Relabel eRisk	Relabel positives	Relabel negatives
count	531394	11521	519873
mean	160.67	203.26	159.73
std	397.83	224.69	400.77
min	1	1	1
25%	39	61	39
50%	72	134	72
75%	159	262	157
max	38216	4033	38216

3.1.1. Relabeled eRisk

Labels in the Original eRisk dataset are applied at user level, not at post level. This means that every post from a positive user is labeled positive, and vice-versa. We propose the hypothesis that not all posts by users at risk contain relevant information that can be detected by an early risk system, and that training a system with these posts labeled as positive makes the system perform worse.

We could approach this hypothesis in different ways: for example, we could apply unsupervised learning, or treat the problem as a zero-shot classification problem. As a first approach to the problem, we chose to re-classify only posts labeled as positive by using sentiment analysis. Posts with a negative sentiment analysis above a certain threshold would keep their classification as positive, while others would be re-classified as negative.

This training set was created with this strategy applied to the Original eRisk training set. Posts from positive users were relabeled based on the sentiment analysis strategy. We applied a twitter-XLM-roBERTa-base model trained on tweets and finetuned for Sentiment Analysis ² [5].

Table 3 shows statistics of the dataset after applying the relabeling.

3.2. Preprocessing

Standard text preprocessing was applied to the text from each post. Posts were cleaned, tokenized, and stems were obtained. Stop words were kept as part of the text, since we believe they are important for this particular task.

We used the Python library *redditleaner* ³ to clean the textual data. We removed Markdown formatting, separated contractions, removed hyperlinks, HTML tags, numbers and multiple spaces. Finally, all text was made lowercase.

3.2.1. Windowfying

Some texts are long, while others are exceptionally short. To curve this difference and make sure a significant length of text is processed in each step without compromising speed, we applied a sliding window to the posts. After cleaning, we joined the text of a post from its previous w

²<https://huggingface.co/cardiffnlp/twitter-xlm-roberta-base-sentiment>

³<https://pypi.org/project/redditleaner/>

messages, where w is a configurable parameter. Features, explained next, were calculated on this message window instead of only on the text of the particular post.

3.3. Features

We used two different strategies for features, depending on whether the classifier algorithm was a traditional machine learning algorithm (models 1 and 2), or a Deep Learning one (model 3).

3.3.1. Traditional features

The features applied to traditional machine learning were an adapted version of the ones used for the eRisk 2021 T2 [6].

Text-based features We applied two features in this category: 1) text length and 2) number of words. We showed in section 2 that positive users are more likely to write longer texts, so we keep track of this information with these two features.

Similarly to our previous eRisk participations, we applied a collection features tailored to the depression dataset. Features were normalized by text length and discretized to a fixed number of bins.

First-person pronouns First-person pronouns: Several works [7] [8] have established that people with mental health problems such as depression tend to use more first-person pronouns when they speak. We create a feature that counts the number of times a first-person singular pronoun is used in a text.

Depression-related words In previous editions of the shared task, we applied a wordset of self-harm related terms as a feature. This year, we applied a collection of words related to clinical depression, and the moods and topics associated with it extracted from [9]. This feature counts the number of depression-related words that appear in a text.

We combined these features to TF-IDF-based features using Scipy Hparse.

3.3.2. TF-IDF features

Similarly to previous years, we trained a TF-IDF featurizer on the positive users of the data and used this featurizer to obtain TF-IDF features for each message window. The featurizer was trained exclusively on positive data (in the case of the relabeled dataset, it was trained only on those messages that remained positive) because we want to detect words used frequently by positive users.

3.3.3. Embeddings

Embeddings were used for the Deep Learning model. We applied Stanford's pre-trained GloVe [10] Wikipedia 2014 100d word Embeddings. Posts were windowfied and then padded to a sufficiently long length in order to include the longest messages before applying the Embeddings.

Table 4
UNED-MED eRisk 2022 T2 runs configurations.

run	dataset	model	training window size	test sliding window size
0	Original eRisk	XGBoost	30	30
1	Augmented eRisk	Random forest	10	10
2	Relabeled eRisk	XGBoost	100	100
3	Relabeled eRisk	XGBoost	100	10
4	Original eRisk	CNN	10	10

3.4. Classifier Algorithms

We worked with traditional machine learning models, and with one Deep Learning model. The classifiers predict whether a message window belongs to a user at risk of being "positive" or "negative". Like the task specifies, a positive decision is final, but a negative decision may be revised later.

We worked with two traditional machine learning models: Random Forest and XGBoost; and one Deep Learning model: a Convolutional Neural Network (CNN).

- Random Forest: We used the scikit-learn⁴ implementation of the Random Forest Classifier [11].
- XGBoost⁵: This model is a type of ensemble model. It learns by optimizing a distributed gradient on learning algorithms of the Gradient Boosting framework.
- CNN: We implemented a Convolutional Neural Network using Keras. The Neural Network was formed by a CNN layer of size 64, a GlobalMaxPooling layer, a Dense layer with relu activation, and an output Dense layer with sigmoid activation.

3.4.1. Training strategy

We applied descending training weights to positive posts. This was in order to make our system prioritize earlier messages, and detect positive users as fast as possible.

Messages created by negative users were all assigned the same training weight (1). Messages created by positive users were assigned descending weights, from oldest to most recent message, through a fixed rate (2 to 1). Our working notes from the eRisk 2021 task [6] present a thorough explanation of the algorithm used to calculate these training weights.

4. Experimental Setup

Table 4 shows the parameter configurations for the five different runs. Each run uses a different combination of training data, classifier model and training and test window size.

They all used weighted training.

⁴<https://scikit-learn.org>

⁵<https://xgboost.readthedocs.io/en/stable/>

Table 5

eRisk 2022 T2 decision-based evaluation. Our teams' results (UNED-MED) are compared to the best results for each metric. Our best results for each metric and the overall best results for the rest of the teams are bolded.

team name	run id	<i>P</i>	<i>R</i>	<i>F1</i>	<i>ERDE</i> ₅	<i>ERDE</i> ₅₀	<i>latency</i> _{tp}	<i>speed</i>	<i>latency-weighted F1</i>
UNED-MED	0	.119	.969	.212	.091	.056	18	.934	.198
UNED-MED	1	.139	.980	.244	.079	.046	13	.953	.233
UNED-MED	2	.122	.939	.215	.086	.057	15	.945	.204
UNED-MED	3	.131	.949	.231	.084	.051	15	.945	.218
UNED-MED	4	.084	.163	.111	.079	.078	251	.252	.028
NLPGroup-IISERB	0	.682	.745	.712	.055	.032	9	.969	.690
BLUE	0	.395	.898	.548	.047	.027	5	.984	.540
UNSL	2	.400	.755	.523	.045	.026	3	.992	.519

5. Results and Discussion

In this section we analyze the task results of our participation.

Table 5 shows our five runs metric results, plus the results for the best groups according to different metrics. Our results this year were modest, with a best latency-weighted F1 of 0.233 compared to NLPGroup-IISERB's 0.690.

When comparing our different runs, we observe that run 1 obtained the best results overall in all the available metrics, followed by run 3. In the following ranking metrics we will also observe our best results in these two runs. Run 1 used the Random forest model trained on the augmented dataset, while run 3 was an XGBoost model trained on the relabeled dataset. Other differences between these runs are the sizes of the feature windows: 10 for run 1, and 10 for run 3.

While modest, we believe these differences show that the strategies to improve the training dataset worked favorably. Our best results were obtained with a model trained on the augmented dataset, which used an increased amount of positive users during training. It would be interesting to see how the models would behave if we trained the XGBoost model on the augmented dataset instead, and the Random Forest model on the relabeled dataset. Unfortunately, due to the amount of combinations we wanted to test, we could not fit these combinations in the official task.

Smaller feature window sizes appear to yield better results, as can be also seen by the ranking of our results. Run 1 was trained on window sizes of size 10, and evaluated on sliding window sizes of the same size. Run 3 was trained on window sizes of size 100, but evaluated on sliding window sizes of size 10. It appears that using smaller sizes on test is better, but it may not be necessary during training.

Our worst results were obtained with run 4, the Deep Learning model. We can only wonder as to why this happened, since usually Deep Learning models perform better than traditional learning models in similar circumstances. Despite the sliding window size being 10, the same as runs 1 and 3, the latency value is also very high compared to our other runs (251 compared to less than 20 for all other runs). This makes us think that maybe something went wrong with

Table 6

Ranking-based evaluation. Our team's results (UNED-MED) are compared to the best team's results. The best results overall are bolded.

team	run	1 writing			100 writings		
		<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>	<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>
UNED-MED	1	.50	.44	.26	.70	.76	.50
UNED-MED	3	.80	.82	.29	.60	.44	.31
BLUE	0	.80	.88	.54	.60	.56	.59
BLUE	1	.80	.88	.54	.70	.64	.67
BLUE	2	.80	.75	.46	.40	.40	.30
TUA1	0	.80	.88	.44	.60	.72	.52
TUA1	2	.80	.88	.44	.60	.72	.52
Sunday-Rocker2	3	.80	.88	.41	.50	.50	.23
UNSL	1	.80	.88	.46	.60	.73	.64
Sunday-Rocker2	1	.70	.81	.39	.90	.93	.66
NLPGroup-IISERB	1	.30	.32	.13	.90	.81	.27
NLPGroup-IISERB	4	.00	.00	.04	.90	.93	.66
NLPGroup-IISERB	0	.00	.00	.02	.90	.92	.30
CYUT	3	.10	.07	.12	.70	.70	.57
CYUT	4	.10	.06	.12	.60	.68	.55

the implementation of this model.

Overall, we believe the depression task has been significantly more challenging than previous edition of the eRisk task. We observed this too while preparing our systems during the training phase, and we believe this might be due to the nature of the data. eRisk datasets are obtained by searching users on Reddit that have explicitly said that they were diagnosed with the mental health problem the task is about (depression, in this case). While other problems such as anorexia and self-harm still have a lot of stigma, openly talking about one's depression is seen more and more in this day and age. Therefore, it is more possible that most positive users in previous years, where anorexia or self-harm were detected, were accounts created exclusively to talk about that specific problem (Reddit users call these kind of accounts "throwaways"), while positive users in the depression dataset are normal users that use their Reddit account to talk about their hobbies, interests, etc.

Table 6 shows our ranked results compared to teams that obtained the best results in this category. In this case we can see some better results in some of the categories for runs 1 and 3. We can see that our results are better in the beginning, when only one message has been processed, and they decrease as time goes by. This might indicate that our system performs better when only a limited amount of messages for every user are observed, and that observing too many messages results in yielding too many false positive results. This is in concordance with results from table 5, where we can see that Recall is very high for four of the five runs, while Precision is very low.

Table 6

Ranking-based evaluation. Our team’s results (UNED-MED) are compared to the best team’s results. The best results overall are bolded. *Continuation*

team	run	1 writing			100 writings		
		<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>	<i>P@10</i>	<i>NDCG@10</i>	<i>NDCG@100</i>
UNED-MED	1	.60	.64	.47	.80	.74	.50
UNED-MED	3	.80	.73	.36	.40	.51	.30
BLUE	0	.80	.81	.66	.80	.80	.68
BLUE	1	.80	.84	.74	.80	.86	.72
BLUE	2	.30	.35	.20	.30	.38	.16
TUA1	0	.60	.67	.52	.70	.80	.57
TUA1	2	.60	.67	.52	.70	.80	.57
Sunday-Rocker2	3	.60	.69	.34	.00	.00	.00
UNSL	1	.60	.73	.66	.60	.71	.66
Sunday-Rocker2	1	.90	.88	.65	.00	.00	.00
NLPGroup-IISERB	1	.80	.84	.33	.00	.00	.00
NLPGroup-IISERB	4	.90	.92	.69	.00	.00	.00
NLPGroup-IISERB	0	.90	.92	.33	.00	.00	.00
CYUT	3	.70	.72	.59	.80	.74	.60
CYUT	4	.60	.69	.59	.80	.84	.61

6. Conclusions and Future Work

This paper presented the UNED-MED participation for the eRisk 2022 T2 task. We developed several classifier models based on TF-IDF, text and specially-tailored features, and a Deep Learning classifier model with Embeddings. We also implemented several strategies to reduce the imbalance of the training data: we obtained more data from Reddit, and we relabeled the original training data. The test results show that our systems obtain modest results, and that more effort is needed to achieve state-of-art results.

In future work, we would like to keep exploring strategies to relabel the data, and maybe experiment with zero-shot learning. This would allow the system to be portable from one kind of disease to another with minimal effort.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the DOT-HEALTH Project (MCI/AEI/FEDER, UE) under Grant PID2019-106942RB-C32, as well as project EXTRAE II (IMIENS 2019), the research network AEI RED2018-102312-T (IA-Biomed), and a predoctoral contract UNED - Santander.

References

- [1] M. De Choudhury, S. Counts, E. Horvitz, Social Media as a Measurement Tool of Depression in Populations, in: Proceedings of the 5th Annual ACM Web Science Conference, WebSci ’13, Association for Computing Machinery, New York, NY, USA, 2013, pp. 47–56.

- [2] J. W. Pennebaker, M. R. Mehl, K. G. Niederhoffer, Psychological Aspects of Natural Language Use: Our Words, Our Selves, *Annual Review of Psychology* 54 (2003) 547–577.
- [3] L. D. Parapar J., Martín Rodilla P., C. F, Overview of erisk 2022: Early risk prediction on the internet., in: *Experimental IR Meets Multilinguality, Multimodality, and Interaction. 13th International Conference of the CLEF Association, CLEF 2022*, Springer International Publishing, 2022.
- [4] D. E. Losada, F. Crestani, A Test Collection for Research on Depression and Language Use CLEF 2016, Évora (Portugal), *Experimental IR Meets Multilinguality, Multimodality, and Interaction (2016)* 28–29. URL: <https://tec.citius.usc.es/ir/pdf/evora.pdf>.
- [5] F. Barbieri, L. E. Anke, J. Camacho-Collados, Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond, 2021. URL: <https://arxiv.org/abs/2104.12250>. doi:10.48550/ARXIV.2104.12250.
- [6] E. Campillo-Ageitos, H. Fabregat, L. Araujo, J. Martínez-Romo, Nlp-uned at erisk 2021: self-harm early risk detection with tf-idf and linguistic features, in: *CLEF, 2021*.
- [7] J. W. Pennebaker, *The Secret Life of Pronouns What Our Words Say About Us*, Bloomsbury Press, 2011.
- [8] T. Edwards, N. S. Holtzman, A meta-analysis of correlations between depression and first person singular pronoun use, *Journal of Research in Personality* 68 (2017) 63–68. URL: <http://dx.doi.org/10.1016/j.jrp.2017.02.005>.
- [9] D. Mowery, H. Smith, T. Cheney, G. Stoddard, G. Coppersmith, C. Bryan, M. Conway, Understanding depressive symptoms and psychosocial stressors on twitter: A corpus-based study, *Journal of Medical Internet Research* 19 (2017) e48. URL: <https://doi.org/10.2196/jmir.6895>. doi:10.2196/jmir.6895.
- [10] J. Pennington, R. Socher, C. D. Manning, GloVe: Global Vectors for Word Representation, in: *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. URL: <http://www.aclweb.org/anthology/D14-1162>.
- [11] L. Breiman, *Machine Learning* 45 (2001) 5–32. URL: <https://doi.org/10.1023/a:1010933404324>. doi:10.1023/a:1010933404324.