Exploring Depression Symptoms through Similarity Methods in Social Media Posts

Notebook for the eRisk Lab at CLEF 2023

Naveen Recharla[1,*], Prasanthi Bolimera[1], Yash Gupta[1] and Anand Kumar Madasamy[1]

[1]National Institute of Technology Karnataka, Surathkal, India

Abstract

Regardless of age, gender, or color, depression affects people all over the world. People feel increasingly at ease sharing their opinions on social networking sites practically every day in the present era of communication and technology. Reddit is a social networking site consisting of subreddits, or single-topic communities, created, maintained, and frequented by anonymous users. Users have the ability to post, comment on, and reply to posts within subreddits. Data for this suggested model is gathered from user posts on Reddit. Our approach involves ranking sentences from a collection of Reddit posts according to their relevance to a depression symptom for the 21 symptoms of depression from the BDI-II Questionnaire.

Keywords

Depression, Social Media, BDI-II, eRisk, Word2Vec, Sentence Transformers

1. Introduction

Social media sites like Reddit, Twitter, Instagram, and Facebook are extremely important in our daily lives. During the pandemic, these platforms' popularity has grown substantially. According to studies, when the Covid-19 pandemic started, people were more prone to express their moods and emotions on Reddit. While negative emotions are more likely to reflect a person's real feelings, positive emotions are less frequently linked to greater life satisfaction. One of the biggest issues during the Covid-19 pandemic has been depression, which has become increasingly prevalent over the past year as stress, one-sidedness, and negative thoughts have all increased. It can be difficult for society to diagnose or treat a person's depression because people aren't always willing to talk to others about their worries.

Mental health conditions such as depression develop gradually over time and have early-stage signs that can be identified. Such disorders might be prevented or better managed. If they are identified early in the disease's progression, additional attention and therapy can be given. Therefore, making assumptions about someone's mental state based on how they act or look is a sophisticated psychological science that has not yet been mechanized. Concerningly, many persons experiencing depression symptoms do not seek professional assistance or psychiatric

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

*Corresponding author.

veninavi36@gmail.com (N. Recharla); prasanthib151@gmail.com (P. Bolimera); guptayash1104@gmail.com (Y. Gupta); m_anandkumar@nitk.edu.in (A. K. Madasamy)

guidance because of the negative social stigma attached to it. People are therefore turning to informational tools like social media to help them with their concerns. Since the advent of social media, those who struggle with depression have found solace in sharing their thoughts and feelings in online forums, tweets, and blogs.

The BDI is used to assess a depressive illness' severity. It can be used to detect depression and track the effectiveness of therapy. Despite not being a diagnostic test, the BDI can aid clinicians in making a diagnosis. A score on the BDI-II between 10 and 18 indicates mild depression, while a score of 30 or more indicates severe depression. The progression of the treatment can also be tracked using the BDI. A consistent drop in scores suggests that the person's symptoms are getting better.

According to psychological research, there is a significant correlation between a person's language use and depression. Reddit is one of the most widely used social media sites today,[1] where users can post anonymously and are very likely to share depressive views. According to researchers, analysis of Reddit posts can be used to detect sadness and other mental health issues. These online activities inspire them to create innovative systems for the early diagnosis of depression and future healthcare solutions. The project involves ranking sentences from a collection of Reddit posts of different users of a particular period of time. They are ranked based on 21 different symptoms of the BDI-II questionnaire and are going to be ranked according to their relevance to one of the depression symptoms of the 21 symptoms from the BDI Questionnaire.

2. Related Work

In [2], they created distributed Reddit users' text data representations using the Skip-gram model. Then for the purpose of encoding they made use of a Convolution Neural Network or Bidirectional Long short term memory model. This model predicts the outputs for the 21 standard BDI questions for each Reddit user text. Ultimately, the user's comprehensive survey is generated by choosing, for each BDI inquiry, the answer that appears most frequently.

In [3], the authors utilized less complex learning models such as SVMs and logistic regression. They approached the problem as a multi-label, multi-class issue, training a separate model for each BDI question.

The approach involves authors incorporating psycholinguistic and behavioral characteristics in their efforts to link the user's posts with their BDI responses in the paper [4].

Filtering posts based on their relevance is advantageous for training classifiers to address different questions as demonstrated in [5]. The process of feature extraction on several pre-trained models such as BERT, ELMo, and SpanEmo. Then these features are passed into a random forest classifier. Using the described approach [6] got an average hit rate of 32.86

The pre-processed data is sent into the BERT or RoBERTa model in [7]. They worked on a total of 21 models, one for each of the 21 questions on the BDI questionnaire, developing one model for each question. The majority of responses are then taken into account to produce the final output for a user, and using this as a base, weights are then applied in subsequent executions to provide the final output.

Three techniques were used in [8] to automatically complete the BDI questionnaire. The first

two approaches classified each user's choices for each BDI item using cosine similarity and well-known classifiers like SVM and Random Forests. They used a language model based on a phrase called SBERT to represent the subjects' posts. The third technique improved a RoBERTa model that was used to predict the respondents' responses for the collection.

The authors used BERT-based classifiers that were specifically trained for each task in [9]. They compared the training data to a number of pre-trained models, including BERT, DistillBERT, RoBERT, and XLM-RoBERT. They approached the topic as a multi-class labeling problem, treating each BDI question as a separate issue.

The eRisk shared activity, which was launched by CLEF in 2017, asks participants to create experimental models for foreseeing risks to their mental health using social media data sets provided by the conference organizers [10]. Our search strategy identified the studies related to eRisk tasks conducted in 2017 (8/54, 15%), 2018 (5/54, 9%), 2019 (1/54, 2%), and 2020 (1/54, 2%) because these tasks used data sets gathered from Reddit and had a focus on depression detection. As more observations on the distinctive contributions of the eRisk shared tasks to the general landscape of research employing Reddit data have been made, this group has been referred to as the CLEF eRisk studies throughout the Results and Discussion sections.

The process of feature extraction using Latent Dirichlet Allocation (LDA) and a pre-trained model from Sentence Transformers library [11]. Further, these feature representations are passed onto several supervised machine learning classification algorithms namely, logistic regression, support vector machine, ensemble classifier, and Gaussian classifier.

In [12], the task was approached by the authors using several different methods, including topic modeling algorithms such as LDA and Anchor Variant, neural models with three distinct architectures (Deep Averaging Networks, Contextualizers, and RNNs), and an approach based on writing styles. Some of the variations considered stylometric factors, such as Part-of-Speech, common n-grams, punctuation, word/sentence length, and use of uppercase letters or hyperlinks.

In 2020, a task measuring the severity of the signs of depression was provided [13]. The details of the overview of the task, the submissions provided by several teams, and the results of all participating teams can be found in [14]. The dataset utilized for the task was the user Reddit posts along with the golden truths for their respective BDI-II questionnaire. [9, 2, 12, 4, 3] are several works of different teams.

In 2021, a task that is a continuation of the previous year's task[13] namely, measuring the severity of the signs of depression was provided. The details of the overview of the task, the submissions provided by several teams, and the results of all participating teams can be found in [15]. The dataset utilized for the task was the user Reddit posts along with the golden truths for their respective BDI-II questionnaire. [16, 17, 8, 18, 7] are several works of different teams.

3. Dataset

Given dataset, of eRisk2023 [19], consists of Reddit posts from 3107 different users. Each file consists of different Reddit posts of a user for a period of time (TREC format) and each post is split into sentences along with tags for each sentence that are unique. To be able to work on the dataset, all the lines are combined into a single CSV file and after combining the data of all the users, the total number of sentences is 4267799.

Table 1
Statics on the eRisk2023 dataset

Number of users	3107
Number of sentences	4267799
Average number of sentences per user	1373

4. Proposed Methodology

4.1. Hybrid Word2Vec

The first methodology mostly involves data pre-processing, word2vec encoding, obtaining phrase embeddings using bow and tf-idf representations, and computing soft cosine similarity. As we can observe in figure 1, the dataset proceeded for pre-processing, which is then encoded

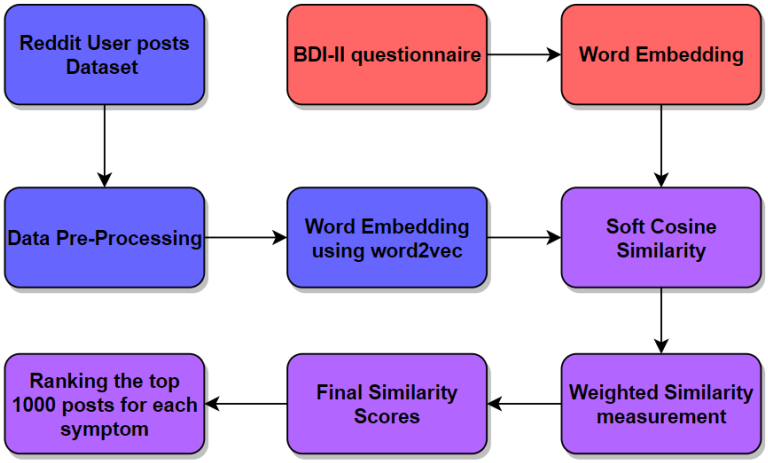


Figure 1: Workflow of the Hybrid Word2Vec methodology

using word2vec. This would be used as our data corpus. Then the BDI-II questionnaire, which is used as the query, is first embedded then the relevant soft cosine similarity scores are calculated. Then the weighted similarity would be computed further. After all these processes are completed, the final top 1000 posts are ranked.

4.1.1. Data Pre-processing

The dataset is initially pre-processed by changing the text's case to lowercase and eliminating punctuation. Next, unnecessary spaces, numerals, emojis, and contractions are removed, and the dataset is tokenized.

4.1.2. Encoding using word2vec

Word embeddings, which are high-dimensional vector representations of words that capture their semantic and grammatical similarities, are created using the machine learning algorithm

Word2Vec. The Word2Vec model is trained on the corpus with the aim of predicting the context in which a word appears. The corpus is built from the sentences in the dataset. Since word embeddings capture the semantic and syntactic links between the words in the corpus, they can be employed after the model has been trained. The similarity between sentences can then be calculated using these embeddings.

4.1.3. Soft Cosine Measure for retrieving relevant sentences

The Soft Cosine Measure (SCM) method allows us to properly compare two articles even if they don't share any words. It employs a word similarity metric that can be discovered using word2vec vector embeddings.

It has been demonstrated to surpass a variety of cutting-edge algorithms in the semantic text similarity challenge when used for replying to public questions. The technique also makes use of the papers' vectorized bag-of-words format. The corpus is used to generate a term similarity matrix, word2vec model, TF-IDF model, and dictionary. The fundamental principle of the method is that we compute standard cosine similarity while assuming that the document vectors are encoded on a non-orthogonal basis, where the angle between two basis vectors is determined by the angle between the respective word2vec embeddings.

The top 1000 sentences for each of the symptoms in the BDI-II questionnaire are chosen as the final result of the retrieval task after the relevance score (similarity) between each symptom and all the phrases in the dataset is determined. The minimum score that can be achieved in the BDI questionnaire is 0 and the maximum is 63, with each question having 0 as the minimum score and 3 as the maximum score.

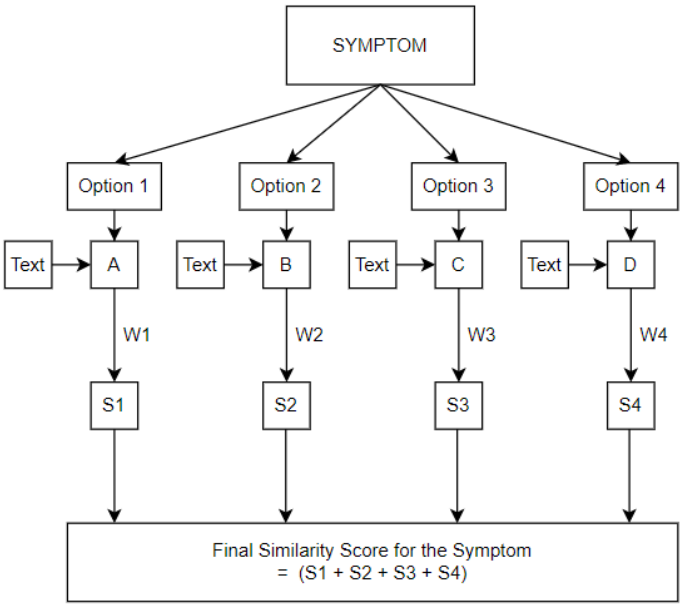
Depression levels can be divided into 4 categories:

- minimal (score of 0–13)
- mild (score of 14–19)
- moderate (score of 20–28)
- severe (score of 29–63)

4.1.4. Weighted Aggregation of the score for symptoms' options

Each option for a symptom is given a specific weightage when computing the similarity score between the phrases in the dataset and each symptom of the BDI. The first option has been given the least weight because it contributes the least to the final score, and the intensity level of depression grows with each alternative. The last option, which means there is a high likelihood that the symptom would appear in that specific user, has also been given the highest weighting, meaning that it will contribute the most to the final score. All the weights are in ascending order.

Soft Cosine Similarity is then calculated between each sentence of the dataset and each option of the symptom. The final score of a sentence is obtained by multiplying the similarity values of each option by its corresponding weight. A similar procedure is followed for all the sentences of the dataset.



*

Figure 2: Similarity Score Calculation for a Symptom

As we can observe, figure2, it illustrates how a sentence's similarity score to a certain symptom is calculated. Now that we are aware that there are four possible causes for each symptom, we have given these causes the numbers 1, 2, 3, and 4 correspondingly. For each option, we determine the sentence's similarity score to a certain ailment. As shown in figure 2, the sentences with options 1, 2, 3, and 4 have similarity scores of A, B, C, and D, respectively. The weights described for the various alternatives are W1, W2, W3, and W4. The weighted similarity scores of the statement with each of the available possibilities for the symptom are S1, S2, S3, and S4.

Weighted similarity score of an option with the sentence is the multiplication of the weights with the respective similarity score. Then the final similarity score of the sentence with the symptom would be the summation of all the weighted similarity scores of the sentence with the respective options of the symptom.

4.2. Sentence Transformers Method

As it can be observed in Figure 1, the dataset proceeds for pre-processing which is then encoded using Sentence Transformers. This would be used as the data corpus in this paper. Then the BDI- II questionnaire which is used as the query is first embedded then the relevant cosine similarity scores are calculated

4.2.1. Sentence Transformers

Sentence Transformers is an embedding framework for sentences, paragraphs, and images. This enables the generation of embeddings with semantic meaning, which is useful for applications

such as semantic search and multilingual zero-shot classification.

This approach uses the para-MiniLM-L3-v2 sentence transformer model. Comparable sentences from various languages can be transformed into comparable vector spaces using multilingual sentence transformers. The model of the technique can be applied to tasks such as clustering and semantic search since it maps phrases and paragraphs to a 384-dimensional dense vector space.

The embeddings for all of the dataset's sentences are constructed and saved in a separate file.

4.2.2. Cosine Similarity

Regardless of size, cosine similarity is a statistic that can be used to determine how similar two data objects are. To determine similarities between two texts, use Python's Cosine Similarity function. Each piece of data in a dataset is treated as a vector via cosine similarity. Its formula is given as follows:

$$Cos(a, b) = a.b/||a|| * ||b|| \tag{1}$$

In the equation 1, a.b is the product (dot) of the vectors namely 'a' and 'b', $||a||$ and $||b||$ is respectively the length of the two vectors 'a' and 'b' and $||a|| * ||b||$ is the cross product of the two vectors 'a' and 'b'.

Embeddings for each symptom are calculated using the model and cosine similarity is then used to calculate the similarities between each sentence of the dataset with each symptom.

The total scores for each sentence, by taking weights, are calculated in a similar way done for the previous method.

5. Results

In the task1 of erisk2023 [20], the evaluation was done metrics was done using the methods namely, majority-based and unanimity. In the tables, we can observe several values for standard performing metrics such as Average Precision (AP), R-Precision, Precision at 10, and NDCG at 1000. There are four runs presented by our team namely:

- SentenceTransformers_0.25: We used the sentence transformers method in this run. After calculating the similarity of a sentence for the four options of a symptom, the final similarity score is obtained by aggregating the similarity scores by providing the same weightage of 0.25 for the four options.
- SentenceTransformers_0.1: We used the sentence transformers method in this run. After calculating the similarity of a sentence for the four options of a symptom, the final similarity score is obtained by aggregating the similarity scores by providing the different weightages of 0.1,0.2,0.3, and 0.4 for the four options according to their intensities. The greater the intensity the greater the weightage.
- result2: We used the hybrid word2vec method in this run. After calculating the similarity of a sentence for the four options of a symptom, the final similarity score is obtained by aggregating the similarity scores by providing the same weightage of 0.25 for the four options.

- word2vec_0.1: We used the hybrid word2vec in this run. After calculating the similarity of a sentence for the four options of a symptom, the final similarity score is obtained by aggregating the similarity scores by providing the different weightages of 0.1,0.2,0.3, and 0.4 for the four options according to their intensities. The greater the intensity the greater the weightage.

Table 2
Ranking-based evaluation (majority voting)

Run	AP	R-PREC	P@10	NDCG@1000
SentenceTransformers_0.25	**0.319**	**0.375**	**0.861**	**0.596**
SentenceTransformers_0.1	0.308	0.359	0.861	0.584
result2	0.086	0.170	0.457	0.277
word2vec_0.1	0.092	0.176	0.5	0.285

In Table 2, which is a majority voting Ranking-based evaluation table, the SentenceTransformers_0.25 got run got the top score among all the teams that participated in the task. For all the evaluation metrics the SentenceTransformers_0.25 got the top score.

Table 3
Ranking-based evaluation (unanimity)

Run	AP	R-PREC	P@10	NDCG@1000
SentenceTransformers_0.25	0.268	**0.360**	**0.709**	**0.615**
SentenceTransformers_0.1	**0.293**	0.350	0.685	0.611
result2	0.079	0.155	0.357	0.290
word2vec_0.1	0.085	0.163	0.357	0.299

Table 3 depicts the unanimity ranking-based evaluation scores. For the Average Precision metric, the SentenceTransformers_0.1 run got top performance among all the teams. For the other three metrics, the SentenceTransformers_0.25 run provided top performance among all the teams in the erisk2023 task1 evaluation process.

6. Conclusion

In this paper, we have proposed a methodology to search for the symptoms of depression. It is demonstrated in the achieved results that our methodology of Sentence Transformers achieved top results, especially the run of Sentence Transformers with equal weightage for all four options. In the experimentation, we can observe that two different weightage distributions lead to different sentence relevance scores. The choice of the usage of the method and weightage depends on the interest in the metric of use. Finally, the methods of using similarity scores proved to be a better field of exploration for the search for depression symptoms. The usage of the relation between symptoms while in search of symptoms might prove to be an advantageous direction in future exploration.

References

[1] J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo, Advances in Information Retrieval: 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2–6, 2023, Proceedings, Part I, volume 13980, Springer Nature, 2023.

[2] A. Madani, F. Boumahdi, A. Boukenaoui, M. C. Kritli, H. Hentabli, Usdb at erisk 2020: Deep learning models to measure the severity of the signs of depression using reddit posts., in: CLEF (Working Notes), 2020.

[3] A.-S. Uban, P. Rosso, Deep learning architectures and strategies for early detection of self-harm and depression level prediction, in: CEUR workshop proceedings, volume 2696, Sun SITE Central Europe, 2020, pp. 1–12.

[4] L. Oliveira, Bioinfo@ uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases, in: Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece, 2020, pp. 22–25.

[5] D. Inkpen, R. Skaik, P. Buddhitha, D. Angelov, M. T. Fredenburgh, uottawa at erisk 2021: Automatic filling of the beck's depression inventory questionnaire using deep learning., in: CLEF (Working Notes), 2021, pp. 966–980.

[6] H. Alhuzali, T. Zhang, S. Ananiadou, Predicting sign of depression via using frozen pre-trained models and random forest classifier., in: CLEF (Working Notes), 2021, pp. 888–896.

[7] S.-H. Wu, Z.-J. Qiu, A roberta-based model on measuring the severity of the signs of depression., in: CLEF (Working Notes), 2021, pp. 1071–1080.

[8] C. Spartalis, G. Drosatos, A. Arampatzis, Transfer learning for automated responses to the bdi questionnaire., in: CLEF (Working Notes), 2021, pp. 1046–1058.

[9] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020 (2020).

[10] F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. H. Bürki, L. Cappellato, N. Ferro, Experimental ir meets multilinguality, multimodality, and interaction, in: Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019). Lecture Notes in Computer Science (LNCS), volume 11696, Springer, 2019.

[11] R. Manna, J. Monti, Unior nlp at erisk 2021: Assessing the severity of depression with part of speech and syntactic (2021).

[12] D. Maupomé, M. D. Armstrong, R. M. Belbahar, J. Alezot, R. Balassiano, M. Queudot, S. Mosser, M.-J. Meurs, Early mental health risk assessment through writing styles, topics and neural models., in: CLEF (Working Notes), 2020.

[13] A. Pérez, J. Parapar, Á. Barreiro, Automatic depression score estimation with word embedding models, Artificial Intelligence in Medicine 132 (2022) 102380.

[14] D. E. Losada, F. Crestani, J. Parapar, Overview of erisk at clef 2020: Early risk prediction on the internet (extended overview)., CLEF (Working Notes) (2020).

[15] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk at clef 2021: Early risk prediction on the internet (extended overview)., CLEF (Working Notes) (2021) 864–887.

[16] D. Maupomé, M. D. Armstrong, F. Rancourt, T. Soulas, M.-J. Meurs, Early detection of

signs of pathological gambling, self-harm and depression through topic extraction and neural networks., in: Clef (working notes), 2021, pp. 1031–1045.
[17] A. Basile, M. Chinea-Rios, A.-S. Uban, T. Müller, L. Rössler, S. Yenikent, M. A. Chulvi-Ferriols, P. Rosso, M. Franco-Salvador, Upv-symanto at erisk 2021: Mental health author profiling for early risk prediction on the internet, in: Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021, CEUR, 2021, pp. 908–927.
[18] A.-M. Bucur, A. Cosma, L. P. Dinu, Early risk detection of pathological gambling, self-harm and depression using bert, arXiv preprint arXiv:2106.16175 (2021).
[19] F. Crestani, D. E. Losada, J. Parapar, Early Detection of Mental Health Disorders by Social Media Monitoring: The First Five Years of the ERisk Project, volume 1018, Springer Nature, 2022.
[20] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2023: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 14th International Conference of the CLEF Association, CLEF 2023, Springer International Publishing, Thessaloniki, Greece, 2023.