MindwaveML at eRisk 2024: Identifying Depression Symptoms in Reddit Users

Notebook for the eRisk Lab at CLEF 2024

Raluca M. Hanciu[1,*]

[1]Master in Data Science, University of Bucharest, Bucharest, Romania

Abstract

In this work, we tackle the first task of the eRisk Lab for 2024: analyzing Reddit comments to identify symptoms of depression. Our objective is to use information collected from Reddit's various user communities to identify sentiments associated with symptoms of depression listed in the BDI-II questionnaire. As a place where people may openly discuss ideas and experiences, Reddit provides a unique perspective on how people express themselves, including their difficulties with mental health. Through utilizing this content created by users, we hope to bring to light how depression presents itself in digital environments.

1. Introduction

Millions of people worldwide suffer from depression, a widespread mental health illness that raises serious public health concerns [1]. The amount of cases of depression has increased in the fast-paced, globally connected society of today, where social demands and stressors are prevalent. Numerous causes, such as higher expectations at work, unstable economy state, loneliness, and the widespread effect of social media, are contributing to this trend [2, 3].

Early detection of depression is critical since it allows for prompt support and care for those who are affected. But conventional diagnosis techniques frequently depend on patients' self-reported symptoms during professional evaluations, which can be arbitrary and vulnerable to underreporting because of stigma or ignorance [4]. Because of this, a large number of depression patients remain undetected or untreated, resulting in severe outcomes and ongoing suffering for the individuals.

Through the use of artificial intelligence (AI) algorithms to examine social media information, researchers can find hidden patterns and possibly even diagnose depression. These realizations can guide the creation of support networks, preventative initiatives, and treatments that are suited to the requirements of those who are at risk [5]. Moreover, by providing scalable and affordable alternatives to conventional diagnostic techniques, AI-driven technologies have the potential to widen access to mental health care [5].

eRisk seeks to close the gap among research and practice by arranging shared tasks and provide a place for practitioners and researchers to work together to create efficient early risk detection. eRisk is dedicated to utilizing digital data in order to facilitate the creation of solutions that are both scalable and easily accessible, with the ultimate goal of improving the lives of those who are dealing with mental health concerns [6, 7].

Under the eRisk system, we focus on the first task for 2024, which aims to identify depressive symptoms from user-generated content. This involves ranking sentences from a sample of Reddit user postings based on their relevance to depressive symptoms. Participants are tasked with providing rankings for the 21 symptoms outlined in the Beck Depression Inventory (BDI) questionnaire [6].

The Beck Depression Inventory (BDI) is a widely used self-report questionnaire carefully constructed to assess the intensity of depression symptoms. Each one of the 21 items depicts a different depression symptom, and participants rate their level of experience with each symptom during a certain time

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

*Corresponding author.

rmhanciu@yahoo.com (R. M. Hanciu)

period [8]. In the given setting of this task, a sentence is regarded as relevant to a BDI symptom when it offers information about how the user is feeling with respect to that symptom. Therefore, a sentence can still be considered significant even if it suggests that the user does not have the symptom [6].

2. Related Work

The first task in eRisk 2024 is an extension of the one in 2023. In 2023, the goal has been similar to this year: score sentences from a selection of user writings based on their relevance to a depressive symptom [9]. The participants will have to rank the 21 indicators of depression from the BDI Questionnaire.

FormulaML's approach to eRisk 2023 included preprocessing and encoding the dataset with Sentence Transformers, namely the MiniLM-L3-v2 model. They used the BDI-II questionnaire as a query, computed cosine similarity scores, and applied weighted scoring to assess sentence relevance. This strategy allowed for a complete investigation of symptoms recognition and similarity assessment [10].

OBSER-MENH compared texts and symptoms using Sentence Transformers (ST), which convert them into fixed-sized vectors. They used pre-trained models such as BERT to build these vectors, which can represent phrases in a dense vector space. For example, the all-mpnet-base-v2 model converts phrases into 768-dimensional vectors, whereas the all-distilroberta-v1 model is trained on a large dataset using the distilroberta-base model. They also employed the all-MiniLM-L12-v2 model [11].

BLUE team used an approach inspired by current research to enhance data using LLMs, creating synthetic Reddit posts associated with every BDI-II symptom to increase dataset diversity. While they expected that adding more diverse data would improve results, their findings revealed that the model using original BDI-II responses outperformed the model using produced data. They emphasized the problem of the specificity of ChatGPT data and proposed improving prompts for semantically comparable but heterogeneous material. Despite this, they acknowledged the useful character of the created text and pointed out the potential of LLMs to generate mental health data for future research [12].

In 2023, Formula-ML received the top scores across all parameters, followed by OBSER-MENH and then the BLUE team. In total there were 10 participating teams.

3. Dataset

Regarding the dataset utilized in this study, it comprised two main folders: one containing the training data and the other the test data. Within the training data folder, a subfolder contained a total of 3107 TREC files, each file representing a distinct user.

Furthermore, the training folder encompassed two CSV files containing subsets of 'docnos' and corresponding labels categorized by query. One CSV file documented labels assigned by the majority of annotators for each 'docno,' while the other recorded consensus labels. In the former, a '1' label signified agreement among at least two out of three annotators regarding the text's relevance, whereas in the latter, a '1' indicated unanimous agreement among all three annotators or disagreement among them.

Notably, the aggregated texts from all TREC files in the training set totaled approximately 4.2 million, illustrating that the annotated data represented only a fraction of the entire corpus. This disparity is expected, given the substantial volume of texts that are unrelated to any of the queries.

It is also important to note the inclusion of the test dataset folder, which has the same structure as the training dataset. This folder contains 553 TREC files, each with numerous sentences similar to the training dataset. These files are critical for evaluating the performance of models trained on the training dataset, since they allow us to analyze generalization capabilities and model efficacy on previously encountered data. This thorough assessment procedure guarantees the model's predictive potential is strong and reliable, hence improving the general quality and validity of the study's results.

4. Method

4.1. Feature Engineering

4.1.1. Cosine Similarity

Capturing semantic nuances in natural language processing (NLP) presents a major challenge due to language's innate complexity and ambiguity [13]. To solve this difficulty, several techniques have been developed, each with its own set of strengths and limitations. One typical technique is to use pre-trained language models, that were developed and fine-tuned on large amounts of text data, to learn complicated syntax and semantic correlations [14]. Hugging Face offers such a pre-trained language model called paraphrase-MiniLM-L12-v2 [15]. This model distinguishes itself by focusing on paraphrase detection, which entails detecting pairs of phrases that contain the same idea but are expressed differently. By training on a variety of paraphrase data sets, the paraphrase-MiniLM-L12-v2 model has gained a thorough understanding of semantic equivalency, allowing it to build embeddings that successfully represent semantic similarities between text sections [15, 16].

The major goal of this newly designed feature is to compute cosine similarity scores for Beck Depression Inventory (BDI) symptoms and dataset sentences. Cosine similarity is a frequently employed measure in NLP that assesses the similarity of two vectors by calculating the cosine of their angle [17]. By computing cosine similarity scores, we hope to obtain the semantic similarity between BDI symptoms and dataset sentences, effectively reflecting the degree of resemblance between them.

The four alternative responses to each of the 21 BDI symptoms are encoded into dense embeddings using the paraphrase-MiniLM-L12-v2 model. These embeddings capture the semantic information contained in each response, converting textual data into high-dimensional vector representations [15]. The sentences are also encoded with paraphrase-MiniLM-L12-v2, and the cosine similarity of each text to its associated query responses is determined. In other words, for each paragraph, there would be four cosine similarities, one for each possible response to the query symptom. I used np.max to get the maximum out of these 4 similarities and used it as a feature, because a text would be deemed as relevant no matter how severe they experience the specific symptom.

4.1.2. First Person

Apart from the fact that a sentence is relevant even if the person affirms they don't have that symptom, a text is only relevant if it is talking about the user only, so:

- A text such as " I feel sad lately" would be labeled as 1 for the first BDI query which is 'sadness', but
- A text such as " My sister is very sad"/ "that is sad", would be label 0 as it is not written from a subjective point of view

As this is an important aspect of our texts, I have created an additional feature called "first_person" which can get values of 0 or 1. If in that specific text there are pronouns such as: ['i', 'me', 'my', 'mine', 'myself', 'im'], then it would be relevant for this new feature.

However, there are many sentences written in first person that actually are irrelevant for a specific query, such as this one: "im sure that symbolizes loneliness" labeled 0 for query 1, which denotes sadness.

To balance this impediment, I have added a condition to this "first_person" feature to only be labeled as 1 if there any of those pronouns are present in the text AND also the cosine similarity is greater than 0.4 for that query.

4.2. Algorithms

In the beginning phase of my research, I undertook a thorough review of numerous machine learning models to establish a performance baseline. I specifically explored with models known for their

computing efficiency, such as Logistic Regression, MLPClassifier, DecisionTreeClassifier, and ensemble models such as GradientBoostingClassifier, XGBClassifier, and LGBMClassifier. My motivation for choosing those models was for a couple of reasons: computational efficiency and the possibility to provide first glimpses into the dataset's features [18].

Given the fact that Logistic Regression and MLPClassifier were the top performers according to f1_score at this point in my research, I have decided to continue working with these two models in my future undertakings.

The dataset has been split into training and testing subsets using StratifiedKFold, which preserves class distributions. This splitting approach preserves the consistency of the evaluation process by reducing biases and increasing the reliability of the outcomes [19].

In each iteration, the model's hyperparameters are fine-tuned using a GridSearchCV pipeline. This pipeline systematically investigates a variety of configurations contained within the parameter grid [20], including for example for MLPClassifier parameters such as: hidden_layer_sizes, alpha, and learning_rate_init [21], and for LogisticRegression: C, penalty [22]. Its major goal is to determine the ideal parameters that result in the highest F1 results.

A key component of the pipeline is the RandomOverSampler, a mechanism used to solve class imbalance. This approach works by oversampling minority class instances within each fold, which corrects the dataset's disproportionate distribution of class labels [23]. By boosting the proportion of minority class samples, the RandomOverSampler helps ensure the model is trained on a more balanced dataset, lowering the risk of biased predictions and improving the classification model's overall performance.

Utilizing joblib, I've preserved the models that demonstrated the highest F1 scores across all five folds for each individual query. Given the performance variations between Logistic Regression and MLPClassifier across different query contexts, I've made the decision to retain both models for future use in making predictions on the test dataset. This approach ensures that we can leverage the strengths of each model depending on the specific characteristics and complexities of the queries encountered during testing, thus optimizing our predictive capabilities and ensuring robustness in our analysis.

I submitted three entries structured as follows:

1. The first submission served as a benchmark to gauge the performance of my models. For this submission, I exclusively relied on the cosine similarity scores generated by the paraphrase-MiniLM-L12-v2 model. These scores were arranged in descending order for each query, and only the top 1000 scores per query were considered.
2. In the second submission, I utilized the probability scores outputted by my algorithms for each query. Depending on the query, I selected either MLPClassifier or Logistic Regression, based on their respective performance during the training phase. Subsequently, I computed the mean of the cosine similarity score and the algorithm's probability score per query.
3. The third and final submission employed both MLP and Logistic Regression models per query, determined by their superior performance during training. However, a distinctive feature of this submission was the utilization of different thresholds to generate scores:
    - For queries with an F1 score ranging between 0.60 and 0.75, the new relevance score was computed as:
    $$0.3 \times \text{prediction\_probability} + 0.7 \times \text{cosine\_similarity\_score}$$
    - Queries with an F1 score between 0.75 and 0.85 utilized a relevance score calculated as:
    $$0.5 \times \text{prediction\_probability} + 0.5 \times \text{cosine\_similarity\_score}$$
    - For queries achieving an F1 score between 0.85 and 0.95, the relevance score was determined as:
    $$0.7 \times \text{prediction\_probability} + 0.3 \times \text{cosine\_similarity\_score}$$

Table 1
Ranking-based evaluation for Task 1 (majority voting)

Team	Run	AP	R-PREC	P@10	NDCG@1000
MindwaveML	MindwaveMLMiniLML12MLP_weighted	0.159	0.240	0.567	0.396
MindwaveML	MindwaveMLMiniLML12MLP_0.5	0.149	0.231	0.538	0.378
MindwaveML	MindwaveMLMiniLML12	0.133	0.212	0.490	0.33
Official Best results					
NUS-IDS	Config_5	**0.375**	**0.434**	0.924	**0.631**
APB-UC3M	APB-UC3M_sentsim-all-MiniLM-L6-v2	0.354	0.391	**0.986**	0.591

Table 2
Ranking-based evaluation for Task 1 (unanimity voting)

Team	Run	AP	R-PREC	P@10	NDCG@1000
MindwaveML	MindwaveMLMiniLML12MLP_weighted	0.158	0.238	0.471	0.427
MindwaveML	MindwaveMLMiniLML12MLP_0.5	0.147	0.227	0.457	0.408
MindwaveML	MindwaveMLMiniLML12	0.128	0.203	0.410	0.360
Official Best results					
NUS-IDS	Config_5	**0.392**	**0.436**	0.795	**0.692**
MeVer-REBECCA	TransformerEmbeddings_CosineSimilarity_gp	0.305	0.357	**0.833**	0.551

5. Results

In the presented tables we can see the outcomes of our team compared with the results of the team/teams that performed best, determined through voting by the annotators. Table 1 delineates the system performance rankings derived from a majority voting perspective, while Table 2 provides rankings based on unanimity among all three annotators.

Upon comparing the proposed methodologies, notable performance enhancements were observed with the approach integrating varied weights for cosine similarity and the probability score generated by the classification algorithm. Notably, the system named "MindwaveMLMiniLML12MLP_weighted" exhibited the highest performance, closely followed by "MindwaveMLMiniLML12MLP_0.5," which computed the mean between the probability score and cosine similarity. In contrast, "MindwaveMLMiniLML12" served as our comparative baseline. These findings affirm that while the improvements were not groundbreaking, the approach demonstrated efficacy by yielding higher scores in comparison to the baseline method.

6. Conclusions

In this research we have proposed an approach of ranking sentences based on their relevance to 21 depression symptoms derived from the Beck Depression Inventory (BDI) questionnaire.

Results from the evaluation process revealed insights into the performance of the developed computational models. Notably, models incorporating a weighted combination of cosine similarity scores and probability outputs from classification algorithms demonstrated superior performance compared to baseline approaches.

Feature engineering is one area that might benefit from some improvements. Further research into more complex methods for obtaining significant characteristics from textual data may result in more accurate depictions of the underlying semantics.

Experimenting with ensemble learning techniques may also yield further insights for improving system performance. Ranking results may be more reliable and robust if several models are combined and their combined intelligence is utilized [24]. Furthermore, examining domain strategies specifically designed for textual data related to depression may aid in bridging the gap between general pre-trained models and the particular demands of mental health domain tasks.

References

[1] W. H. Organization, Depressive disorder (depression), 2023. URL: https://www.who.int/news-room/fact-sheets/detail/depression.

[2] J. F. Greden, The burden of recurrent depression: causes, consequences, and future prospects, Journal of Clinical Psychiatry 62 (2001) 5–9.

[3] S. Cunningham, C. C. Hudson, K. Harkness, Social media and depression symptoms: a meta-analysis, Research on Child and Adolescent Psychopathology 49 (2021) 241–253. URL: https://doi.org/10.1007/s10802-020-00715-7. doi:10.1007/s10802-020-00715-7.

[4] A. Handy , R. Mangal , T. S. Stead , R. L. Jr Coffee , L. Ganti , Prevalence and impact of diagnosed and undiagnosed depression in the united states, Cureus 14(8) (2022). doi:10.7759/cureus.28013.

[5] F. Zafar, L. Fakhare Alam, R. R. Vivas, J. Wang, S. J. Whei, S. Mehmood, A. Sadeghzadegan, M. Lakkimsetti, Z. Nazir, The role of artificial intelligence in identifying depression and anxiety: A comprehensive literature review, Cureus 16(3) (2024). doi:10.7759/cureus.564723.

[6] J. Parapar, P. M. Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. 15th International Conference of the CLEF Association, CLEF 2024, Springer International, 2024.

[7] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2024: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum CLEF 2024, Grenoble, France, September 9th to 12th, 2024, CEUR Workshop Proceedings, 2024.

[8] J. Upton, Beck Depression Inventory (BDI), Springer New York, New York, NY, 2013, pp. 178–179. URL: https://doi.org/10.1007/978-1-4419-1005-9_441. doi:10.1007/978-1-4419-1005-9_441.

[9] 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.

[10] N. Recharla, P. Bolimera, Y. Gupta, A. K. Madasamy, Exploring depression symptoms through similarity methods in social media posts, in: CLEF (Working Notes), 2023.

[11] J. Martinez-Romo, L. Araujo, X. Larrayoz, M. Oronoz, A. Pérez, Obser-menh at erisk 2023: Deep learning-based approaches for symptom detection in depression and early identification of pathological gambling indicators, in: CLEF (Working Notes), 2023.

[12] A.-M. Bucur, Utilizing chatgpt generated data to retrieve depression symptoms from social media, in: CLEF (Working Notes), 2023.

[13] E. Gabrilovich, S. Markovitch, Wikipedia-based semantic interpretation for natural language processing, Journal of Artificial Intelligence Research 34 (2009) 443–498. URL: https://doi.org/10.1613/jair.2669. doi:10.1613/jair.2669.

[14] H. Wang, J. Li, H. Wu, E. Hovy, Y. Sun, Pre-trained language models and their applications, Engineering 25 (2023) 51–65. URL: https://www.sciencedirect.com/science/article/pii/S2095809922006324. doi:https://doi.org/10.1016/j.eng.2022.04.024.

[15] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[16] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[17] B. Li, L. Han, Distance weighted cosine similarity measure for text classification, in: H. Yin, K. Tang, Y. Gao, F. Klawonn, M. Lee, T. Weise, B. Li, X. Yao (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2013, Springer Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 611–618.

[18] L. Shkurti, F. Kabashi, V. Sofiu, A. Susuri, Performance comparison of machine learning algorithms for albanian news articles, IFAC-PapersOnLine 55 (2022) 292–295. URL: https://www.

sciencedirect.com/science/article/pii/S2405896322030774. doi:https://doi.org/10.1016/j.ifacol.2022.12.037, 21st IFAC Conference on Technology, Culture and International Stability TECIS 2022.

[19] X. Zeng, T. R. Martinez, Distribution-balanced stratified cross-validation for accuracy estimation, Journal of Experimental & Theoretical Artificial Intelligence 12 (2000) 1–12. doi:10.1080/095281300146272.

[20] P. Liashchynskyi, P. Liashchynskyi, Grid search, random search, genetic algorithm: A big comparison for nas (2019).

[21] T. Windeatt, Ensemble MLP Classifier Design, Springer Berlin Heidelberg, Berlin, Heidelberg, 2008, pp. 133–147. URL: https://doi.org/10.1007/978-3-540-79474-5_6. doi:10.1007/978-3-540-79474-5_6.

[22] E. Bisong, Logistic Regression, Apress, Berkeley, CA, 2019, pp. 243–250. URL: https://doi.org/10.1007/978-1-4842-4470-8_20. doi:10.1007/978-1-4842-4470-8_20.

[23] R. Ghorbani, R. Ghousi, Comparing different resampling methods in predicting students' performance using machine learning techniques, IEEE Access 8 (2020) 67899–67911. doi:10.1109/ACCESS.2020.2986809.

[24] A. Mohammed, R. Kora, A comprehensive review on ensemble deep learning: Opportunities and challenges, Journal of King Saud University - Computer and Information Sciences 35 (2023). doi:10.1016/j.jksuci.2023.01.014.