Romanian Reddit Depression and Well-being Corpus

Petru-Theodor Cristea, Mihai Militaru, Mihai Dilirici

June 2025

Abstract

The scarcity of large, high-quality datasets in Romanian presents a significant challenge for building reliable natural language processing models, especially in the area of mental health detection. To bridge this gap, we introduce a novel corpus of 4721 Romanian sentences related to mental health and emotional state recognition. The dataset was constructed by collecting posts and top-level comments from Romanian-language subreddits on Reddit, which were manually annotated by our team as either depressive or non-depressive (positive/neutral). To further diversify the data, additional samples were sourced from an external corpus in English language, translated into Romanian using machine translation systems, and manually verified for quality and consistency. This dataset also includes a baseline experiment using BERT-based sentence embeddings combined with a Support Vector Machine (SVM) classifier, demonstrating strong classification performance. Our resource is intended to support further research on Romanian emotion classification, mental health analysis, and sentiment detection.

1 Introduction

Romanian Reddit Depression and Well-being Corpus is a manually annotated dataset designed for the task of binary text classification in the Romanian language, distinguishing between depressive and non-depressive content. The dataset consists of short sentences extracted from Reddit posts and comments originating from Romanian-speaking subreddits. Its main purpose is to enable the development and evaluation of models for detecting signs of depression and well-being in informal, user-generated texts.

The dataset consists of simple, standalone sentences extracted from Reddit posts and comments written in Romanian. Each data sample represents a single sentence (not full posts or paragraphs) to ensure clarity and reduce context dependency during classification tasks.

The sentences in the dataset originate from two primary sources. The first source, referred to as SCRAPED, includes sentences directly extracted from Romanian-language subreddits such as `r/romania`, `r/CasualRO`, and `r/bucuresti`.

An example of a depressive sentence from this source is: "În 2024 am trecut prin cea mai grea perioadă a vieții mele — o depresie care aproape m-a băgat în spital." Conversely, a non-depressive example from the same source is: "Când i-am zis la doctorul de familie a pus mâna pe telefon și a sunat ea."

The second source, labeled as MACHINE_TRANSLATION (MT), consists of sentences that were originally authored in other languages and subsequently translated into Romanian using a machine translation system. A representative depressive example from this category is: "Nu am nimic altceva decât dispreț pentru mine însumi." In contrast, a non-depressive example from the MT subset is: "Care sunt unele limite pe care le-ai stabilit cu părinții și socrii tăi?".

2 Data Provenance

The primary source of the data used in this corpus is Reddit, specifically a selection of subreddits relevant to Romanian-speaking communities, including `r/romania`, `r/CasualRO`, `r/WomenRO`, `r/bucuresti`, `r/brasov`, `r/cluj`, `r/iasi`, `r/Men_RO`, `r/moldova`, and `r/Roumanie`. Data was systematically extracted using the official `PRAW` API. To ensure relevance and diversity, only the top ten comments per post, ranked by Reddit's internal scoring mechanism, were included in the final dataset.

The dataset was constructed by scraping posts and comments from a selection of Romanian-language and region-specific subreddits on Reddit. A comprehensive Boolean search query was designed to retrieve posts containing depression-related and emotional keywords in Romanian. The query consisted of the logical OR combination of terms including *depresie*, *anxietate*, *panica*, *plans*, *suferinta*, *singuratate*, *dezamagire*, *frica*, *"probleme in familie"*, *"ganduri negre"*, *PTSD*, *tristete*, *sinucidere*, *suicidal*, and *angoasa*, aiming to maximize the recall of relevant emotional expressions.

For each subreddit, up to 1500 posts matching the query were retrieved using the Reddit API. From these posts, both the main post body (`selftext`) and the top 15 highest-scoring comments were processed. Texts were segmented into sentences based on punctuation delimiters and newline characters.

To ensure data quality and relevance, sentences were filtered by the following criteria:

- Sentences shorter than three words were discarded.
- Sentences containing no keywords from a predefined lexicon of emotion-related terms were excluded.
- Duplicate sentences within the dataset were removed to reduce redundancy.
- Only posts and comments containing at least ten words were considered for sentence extraction.

Each extracted sentence was stored along with contextual metadata including the original post title, subreddit name, and sentence source (post or comment). The processed data from each subreddit was saved into separate CSV files for subsequent manual annotation and analysis. The table 1 will contain how many samples we found for each subreddit.

Table 1: Number of Filtered Sentences per Subreddit by Class

Subreddit	**Depressive Sentences**	**Non-depressive Sentences**
Roumanie	–	400
moldova	–	203
CasualRO	1489	1805
WomenRO	1182	1523
Men_RO	574	1328
bucuresti	330	783
brasov	19	–
cluj	328	–
iasi	66	–

After preprocessing and manual annotation, the resulting dataset comprises a total of 4721 unique sentences. Of these, 2221 sentences were labeled as expressing depressive content, while 2500 were labeled as non-depressive. The sentences were derived from both post bodies and comment sections to ensure a balanced representation of user expressions across different interaction types on the platform. The overall distribution of depressive and non-depressive samples in the dataset is illustrated in Figure 1.

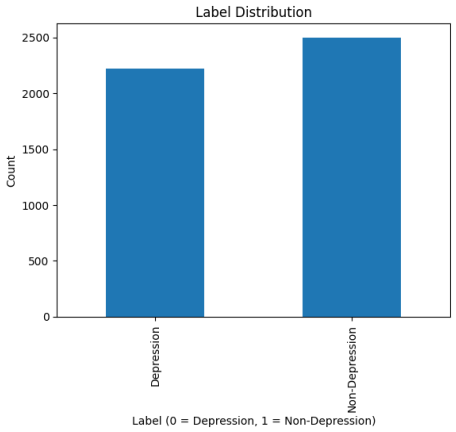

Figure 1: Data distribution of depressive vs. non-depressive sentences

Additionally, to compensate for the lack of sufficiently diverse data in Romanian, a subset of sentences originally authored in other languages was translated into Romanian using high-quality machine translation tools. These translated

samples were subsequently reviewed to maintain linguistic and contextual accuracy within the corpus. This dual-sourcing approach enhances the robustness of the dataset, making it suitable for training and evaluating machine learning models on depression detection tasks in low-resource languages such as Romanian. In Figure 2 we have all distribution including our scraped data and data available after Machine Translation process.

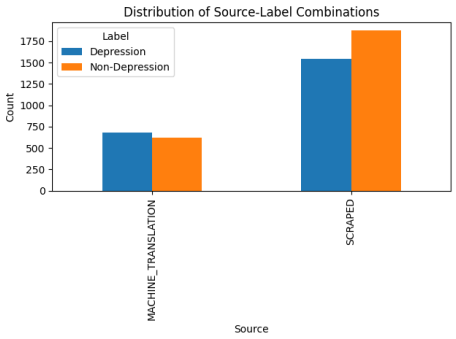

Figure 2: Detailed data distribution.

Part of the resources used in this study were inspired by publicly available datasets. Specifically, we referred to the dataset provided in this GitHub repository https://github.com/Inusette/Identifying-depression (Bucur et al., 2025), associated with the work of Inna Pirina and Çağrı Çöltekin (Pirina and Çöltekin, 2018). This repository was identified through the curated list of depression-related NLP datasets available in this here.

3 Language of the Data

The dataset is composed entirely of sentences in Romanian; however, the linguistic origin of these sentences is divided between two sources. Approximately 70% of the data consists of naturally written Romanian, extracted directly from posts and comments authored by native speakers on Romanian-language subreddits.

The remaining 30% of the dataset comprises sentences that were originally authored in English and subsequently translated into Romanian. The translation process was carried out using the OpenAI GPT-4.1 (OpenAI, 2024) language model via a prompt-based pipeline. The prompt used during this process explicitly instructed the model to translate the given English sentence into Romanian and to return only the translated text without additional comments or explanations. Manual verification of these translated samples was performed partially to assess the linguistic and contextual adequacy of the output.

4 Baseline Models

As a starting point for evaluation, two baseline models were developed using distinct feature representations. Prior to model training, the dataset underwent some preprocessing steps to ensure consistency and noise reduction. Each sentence was labeled with a binary tag: 0 for depressive content and 1 for non-depressive content, based on manual and semi-automatic annotation procedures. The dataset was subsequently split into training and testing subsets in an 80-20 ratio.

The first baseline employed a Word2Vec (Mikolov et al., 2013) based vectorization of the sentences, followed by classification using a Support Vector Machine (SVM). This model achieved an accuracy of approximately 69.7%. The classification report indicated a precision of 0.72 for the depressive class and 0.69 for the non-depressive class, with an overall macro-averaged F1-score of 0.69. The confusion matrix revealed that the model struggled particularly with correctly identifying depressive samples, resulting in a relatively high number of false negatives.

In contrast, the second baseline utilized sentence embeddings generated by a Romanian BERT model, also classified using an SVM (Cortes and Vapnik, 1995). This approach significantly outperformed the Word2Vec baseline, achieving an accuracy of approximately 84.9%. Precision, recall, and F1-scores exceeded 0.82 for both classes, indicating a more balanced and effective discrimination between depressive and non-depressive content. The confusion matrix for this model showed a substantial reduction in both false positives and false negatives compared to the Word2Vec-based baseline.

These baseline results demonstrate the advantage of using contextualized embeddings from Romanian BERT (Dumitrescu et al., 2020), over traditional static word vector representations when addressing the task of depression detection in Romanian texts. Table 2 is containing a detailed comparison of the performance metrics.

Table 2: Baseline Results Comparison

Model	**Acc.**	**Precision**		**Recall**		**F1-score**	
		0	**1**	**0**	**1**	**0**	**1**
Word2Vec + SVM	0.697	0.72	0.69	0.59	0.79	0.65	0.73
Romanian BERT + SVM	0.850	0.85	0.85	0.82	0.88	0.84	0.86

5 Limitations and Risks

Several methodological limitations should be considered when interpreting and generalizing the results of this study. Firstly, the manual and semi-automated annotation process inevitably introduces a degree of subjectivity, stemming from the individual perceptions and judgments of the human annotators involved.

Even when clear annotation guidelines are in place, the interpretation of linguistic content may vary between individuals, potentially introducing noise or inconsistency into the final dataset.

Another potential source of variation arises from the inclusion of automatically translated texts from English into Romanian. Although the translation was performed using a state-of-the-art model (GPT-4.1) and partially reviewed by human annotators, stylistic, lexical, or semantic discrepancies between native Romanian texts and machine-translated content may still exist. Such discrepancies could affect the statistical distribution of certain linguistic features and, consequently, the performance of machine learning models trained on this corpus.

Moreover, the study is subject to platform-specific bias. Data were collected exclusively from RedditInc., 2024, a platform with distinct demographic and discursive characteristics that may not reflect the language use or communicative behavior of users on other social media platforms (e.g., Facebook, Instagram, TikTok). This limitation could restrict the generalizability of the developed models to broader Romanian-language digital contexts.

6 Ethical and Licensing Considerations

The entire data collection and processing workflow was conducted in strict compliance with Reddit's public data policy. Only publicly available content was extracted, with no access to private or restricted information, thereby respecting user rights and platform guidelines.

The resulting dataset is intended exclusively for academic and research purposes, with no direct or indirect commercial usage. It will be made available to interested researchers under open science principles, but only under conditions ensuring data protection and intellectual property compliance.

In accordance with the General Data Protection Regulation (GDPR), the dataset contains no personally identifiable information (e.g., names, addresses, user accounts), and the textual data have been anonymized to remove any potentially sensitive references. As a result, the risk of re-identification of real users is considered negligible.

References

Bucur, A.-M., Moldovan, A., Parvatikar, K., Zampieri, M., Khudabukhsh, A., & Dinu, L. (2025). Datasets for depression modeling in social media: An overview. *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, 116–126. https://aclanthology.org/2025.clpsych-1.10/

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. https://doi.org/10.1007/BF00994018

Dumitrescu, S. D., Avram, A.-M., & Trandabat, D. (2020). The birth of romanian bert. *arXiv preprint arXiv:2009.02201*. https://arxiv.org/abs/2009.02201

Inc., R. (2024). Reddit data api terms and data use policy.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1301.3781

OpenAI. (2024). Chatgpt.

Pirina, I., & Çöltekin, Ç. (2018). Identifying depression on reddit: The effect of training data. *Proceedings of the 5th Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic (CLPsych)*, 9–15. https://aclanthology.org/W18-5903.pdf.