LHS712Team-1 at eRisk@ CLEF 2025: Searching for Depression Symptoms Using Various Natural Language Processing Algorithms

Notebook for the eRisk Lab at CLEF 2025 - Shared Task 1

Aisha Benloucif[1,†], Yashasvini Nannapuraju[1,†], Sripriya Bellam[1,†], Yuyan Hu[1,†], Zhe Zhao[2] and V.G.Vinod Vydiswaran[1,2,*]

[1] *School of Information, University of Michigan, Ann Arbor, MI 48109, USA*
[2] *Department of Learning Health Sciences, University of Michigan, Ann Arbor, MI 48109, USA*

Abstract

Depression is one of the most common mental health problems worldwide, affecting more than 280 million people. Early recognition and intervention are crucial to prevent severe consequences. This study participated in Task 1 of CLEF eRisk 2025, focusing on ranking the relevance of user-generated content to symptoms defined in the Beck Depression Inventory (BDI). Using various machine learning models and natural language processing techniques, including logistic regression (LR), Support Vector Machine(SVM), and BERT-based models, we aimed to advance early detection of depressive symptoms in online text, offering new tools for future mental health early prevention.

Keywords

Depression, Early Detection, Beck Depression Inventory (BDI), Natural Language Processing (NLP), Machine Learning, BERT,

1. Introduction

Depression is a mental disorder that has a staggering impact globally, affecting more than 280 million people, as noted by the World Health Organization [1]. Depression manifests itself as an array of symptoms, such as low mood, low interest in activities (anhedonia), sleep disturbances, alterations in appetite, etc. If not identified and addressed promptly, depression can severely affect one's ability to learn, work, and perform other daily activities, and in extreme cases lead to self-harm or suicidal actions [1]. Therefore, there is a critical need for accurate early recognition and intervention strategies to advance mental health research and public health.

The rise of social media and digital communication platforms has created an opportunity for mental health researchers to analyze user-generated content in real time. Unlike traditional clinical assessments, which are based on self-reports or structured interviews, online expressions such as tweets, Facebook posts, and forum discussions provide researchers with unique insight into individuals' emotional and psychological states [2, 3]. These data are particularly valuable because they capture everyday contexts in natural language, revealing subtle linguistic markers of distress, such as increased use of negative emotion words, first-person pronouns, or increased ruminations [3].

In recent years, a growing body of literature has explored the use of natural language processing (NLP) techniques to assess mental health from online content. For example, Coppersmith et al. used NLP methods to detect linguistic signals related to mental health from Twitter posts, showing how social media text can help identify mental health concerns [4]. Machine learning (ML) has greatly advanced the detection of mental health signals from text data. Traditional methods like Logistic Regression (LR)

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain

[*] Corresponding author.

[†] These authors contributed equally.

abenlouc@umich.edu (A. Benloucif); ynannapu@umich.edu (Y. Nannapuraju); sbellam@umich.edu (S. Bellam); yuyanhu@umich.edu (Y. Hu); zzhaozhe@umich.edu (Z. Zhao); vgvinodv@umich.edu (V.G.Vinod Vydiswaran)

used features such as n-grams and sentiment lexicons to classify text. Recently, pre-trained language models (PLMs) like Bidirectional Encoder Representations from Transformer (BERT) have enhanced NLP by capturing word context, allowing for better understanding of emotional and psychological nuances. BERT's ability to process large datasets and contextualize language has made sentiment analysis more accurate [5].

This paper summarizes our participation in Task 1 of CLEF eRisk 2025 [6, 7], a task aimed at ranking sentences from user writings on social media by their relevance to the 21 symptoms defined in the Beck Depression Inventory (BDI). The challenge of this task is that some sentences do not overtly express negative emotions. By applying various ML techniques, we examined their effectiveness in ranking sentences according to their relevance to depression symptoms. Our results propel further interdisciplinary inquiry at the intersection of NLP and mental health.

2. Task 1: Search for Symptoms of Depression

2.1. Task Description

This task focuses on detecting depression signals in user writings on social media. Participants are asked to build a system that could rank provided sentences for each of the 21 symptoms of depression from the Beck Depression Inventory II (BDI-II) questionnaire. A sentence is considered relevant if it contains the signal about the user's condition of a particular symptom.

The official TREC-formatted annotated corpus from the eRisk shared tasks 2023 and 2024 were provided [6, 7]. Each sentence in these data sets was annotated for relevance to one or more of the 21 symptoms listed in BDI-II. Two annotation schemes were provided: majority vote (agreement by at least two annotators) and full consensus (unanimous agreement). This corpus involved 9,000 social media users with a total of 17,553,441 sentences. The average number of words per sentence is 12.39.

2.2. Methods

In this study, a range of machine learning techniques were employed to analyze textual data and perform binary classification of sentence relevance, ranging from classical machine learning approaches to more advanced transformer-based deep learning models. The experimental design consisted of multiple run groups, each applying distinct vectorization and modeling strategies. Initial experiments used logistic regression and support vector machine (SVM) classifiers paired with CountVectorizer and TF-IDF vectorization, respectively. In subsequent experiments, more advanced feature representations, such as ClinicalBERT and Sentence-BERT (SBERT) embeddings, were incorporated to better capture semantic context. Models were evaluated using standard classification metrics, including precision, recall, F1-score, and computational efficiency. The final experiments explored a hybrid retrieval approach that combined BM25 lexical matching [8] with semantic re-ranking using SBERT, aiming to improve relevance ranking by leveraging both lexical overlap and contextual meaning.

During the training process, the datasets were merged and split into train and test sets in an 80/20 ratio. Relevance labels were binarized, assigning a value of 1 to relevant sentences and 0 to non-relevant ones. No external datasets were used during training or evaluation.

3. Results

3.1. Submitted Runs

We submitted five runs in total in the requested TREC format: "result" (LR with CountVectorizer), "LR file combined" (LR with TF-IDF representations), "SVM file combined" (SVM with TF-IDF Representations), "BERT CONSENSUS" and "BERT MAJORITY". Two types of evaluation schema were used to determine if sentences were correctly ranked to a symptom or not: majority-based (a sentence was deemed relevant if at least two of the three assessors marked it so) and unanimity-based (a sentence was deemed relevant only when all three assessors agreed) [6].

From both results table 1 and 2, the best-performing approaches involved fine-tuning the BERT [9] model. For each training instance, input was constructed in the format “[SYMPTOM] [SEP] sentence”, allowing the model to learn the relationship between a depression symptom and a candidate sentence. Separate models were fine-tuned for both majority-vote and consensus-labeled datasets to enable comparative evaluation. The training was performed over 5 epochs, using a batch size of 16, a learning rate of 2e-5, and the AdamW optimizer with weight decay. Model checkpoints were saved at the end of each epoch. The fine-tuned model learned to classify the relevance of sentences to specific BDI-II symptoms with high accuracy.

Given the scale of the eRisk 2025 test set (over 17 million unlabeled sentences), a symptom-aware keyword filtering pipeline was developed to pre-filter candidate sentences. This pipeline used symptom-specific keyword lists to reduce computational burden and improve inference speed. Filtered sentences were paired with their corresponding symptom prompts and passed to the BERT model for scoring. For each symptom, the top 1,000 sentences with the highest predicted relevance scores were selected and formatted according to TREC submission standards. The BERT model fine-tuned on unanimity-based labels achieved the highest scores on all five metrics in both ranking-based evaluation schema.

3.2. Unsubmitted Runs

3.2.1. SVM with ClinicalBERT Embeddings

We tried fitting SVM classifier with ClinicalBERT [10] embeddings, which is pretrained on clinical corpora. This approach achieved a slightly higher accuracy of 0.79 on the validation set, with strong performance for non-relevant sentences (F1 = 0.85) but moderate effectiveness for relevant ones (F1 = 0.63). The macro F1-score was 0.74, and the weighted F1-score was 0.79. Although ClinicalBERT provided richer semantic context, SVM with TF-IDF representations offered comparable results at a significantly lower computational cost, making it a viable alternative for large-scale or resource-constrained applications.

3.2.2. SVM/LR with SBERT

In this experiment, Sentence-BERT (SBERT) [11] was adopted, offering semantically meaningful sentence embeddings optimized for short text classification. Two classifiers – Support Vector Classifier (SVC) and Logistic Regression (LR) [12, 13] – were trained using embeddings generated by SBERT. The training data was derived from a merged 2023–2024 dataset, with sentence texts serving as input features and the binary “relevance” label as the output. To ensure consistency, the same 80/20 train-validation split from earlier experiments was used. Both SVC and LR models were trained on identical SBERT-based features to facilitate a fair comparison.

Despite SBERT’s semantic strength and computational efficiency, both classifiers achieved only moderate performance. Accuracy for both models were about 77% on the validation set, with strong precision and recall for the non-relevant class but lower metrics for the relevant class. These results suggest that while SBERT captures contextual meaning effectively, further optimization is required, particularly to address class imbalance and fine-tune model parameters.

3.2.3. BM25 with SBERT

A hybrid retrieval framework was implemented that combined traditional BM25 lexical retrieval with semantic re-ranking using a pre-trained Sentence-BERT model. The two-step approach was designed to leverage BM25’s strength in lexical term matching while addressing its limitations in semantic understanding through contextualized embeddings. While BM25 performs effectively in retrieving documents with overlapping vocabulary, it often fails to capture semantically related candidates phrased differently – an area where embedding-based models demonstrate superior performance.

Table 1
Ranking-Based Evaluation for Task 1 (Majority) — LHS712-Team-1

Run	AP	R-PREC	P@10	NDCG
results	0.000	0.000	0.000	0.000
BERT CONSENSUS	**0.075**	**0.153**	**0.305**	**0.264**
BERT MAJORITY	0.050	0.132	0.242	0.229
LR file combined	0.000	0.004	0.010	0.007
SVM file combined 4	0.000	0.004	0.010	0.007

Table 2
Ranking-Based Evaluation for Task 1 (Unanimity) — LHS712-Team-1

Run	AP	R-PREC	P@10	NDCG
results	0.000	0.000	0.000	0.000
BERT CONSENSUS	**0.078**	**0.169**	**0.344**	**0.287**
BERT MAJORITY	0.057	0.148	0.281	0.251
LR file combined	0.000	0.004	0.010	0.007
SVM file combined 4	0.000	0.004	0.010	0.007

The evaluation on the validation set demonstrated a weighted precision of 0.47, recall of 0.34, and an F1-score of 0.39. BM25 retrieval alone produced strong recall by retrieving highly relevant documents, whereas precision improved significantly after applying Sentence-BERT-based re-ranking. However, precision (0.09) and recall (0.18) for Class 1 (relevant) sentences remained notably lower than for Class 0 (non-relevant) sentences, indicating that while semantic matching improved, the consistent identification of truly relevant examples remains a challenge.

3.3. Discussion

The ranking evaluation demonstrates that fine-tuned BERT models consistently outperform traditional classifiers and unsupervised methods in retrieving clinically relevant sentences. Notably, the BERT models trained on unanimity-based labels achieve higher ranking scores across all metrics compared to those trained on majority-based labels. This suggests that when the labeling task is easier for human experts and/or they agree, the examples yield clearer relevance signals, enabling the model to rank pertinent sentences more effectively. Traditional machine learning classifiers such as logistic regression and SVM perform close to baseline levels in ranking metrics, highlighting their limited ability to capture nuanced clinical relevance for this task. The hybrid BM25 + SBERT approach improves lexical-semantic matching but still struggles with precision and recall, underscoring the importance of deep contextual understanding that transformer models like BERT provide. Overall, these results emphasize the value of transformer-based models trained on consensus labels for achieving improved retrieval performance in clinical sentence ranking tasks.

4. Conclusion

The results suggest that BERT outperformed traditional models such as logistic regression (LR) with term-frequency or TF-IDF features, primarily due to its ability to understand contexts within language. BERT's superiority is likely attributable to its transformer-based architecture, which enables nuanced comprehension beyond the keyword-level focus of term-frequency based features. Among the two traditional vectorization methods, TF-IDF yielded better results than term-frequency, emphasizing the importance of weighting term, though both approaches fell short in capturing semantic meaning.

Despite these promising findings, the study had several limitations. A key issue was the imbalanced labels: sentences labeled as "relevant" (Label 1) were underrepresented, negatively impacting the performance of LR in particular. Future research could address this limitation through oversampling techniques or contrastive learning methods to enhance minority-class representation. BERT, while effective, also introduced scalability concerns due to its high computational cost, which constrained the number of models and hyperparameter combinations that could be evaluated.

This supports BERT's potential for high-impact use cases where contextual understanding is essential. As NLP methods for classification tasks continue to evolve, integrating advanced NLP models like BERT or domain-specific variations, such as Opinion-BERT [14], which has been applied in mental health analysis to detect nuanced sentiment and psychological states, offers a path toward deeper and more reliable insights [15].

Declaration on Generative AI

During manuscript preparation, the authors used ChatGPT-4o for grammar and spelling check, then reviewed and edited the content as needed. They take full responsibility for the publication's content.

References

[1] World Health Organization, Depressive disorder (depression) (2023). URL: https://www.who.int/news-room/fact-sheets/detail/depression, retrieved March 31, 2023.

[2] 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, 2021, pp. 128–137. URL: https://doi.org/10.1609/icwsm.v7i1.14432. doi:10.1609/icwsm.v7i1.14432.

[3] J. C. Eichstaedt, R. J. Smith, R. M. Merchant, L. H. Ungar, P. Crutchley, D. Preoţiuc-Pietro, D. A. Asch, H. A. Schwartz, Facebook language predicts depression in medical records, Proceedings of the National Academy of Sciences of the United States of America 115 (2018) 11203–11208. doi:10.1073/pnas.1802331115, epub 2018 Oct 15.

[4] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in Twitter, in: P. Resnik, R. Resnik, M. Mitchell (Eds.), Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Association for Computational Linguistics, Baltimore, Maryland, USA, 2014, pp. 51–60. URL: https://aclanthology.org/W14-3207/. doi:10.3115/v1/W14-3207.

[5] E. C. Garrido-Merchan, R. Gozalo-Brizuela, S. Gonzalez-Carvajal, Comparing bert against traditional machine learning models in text classification, Journal of Computational and Cognitive Engineering 2 (2023) 352–356. URL: https://ojs.bonviewpress.com/index.php/JCCE/article/view/838. doi:10.47852/bonviewJCCE3202838.

[6] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet (extended overview), in: Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), Madrid, Spain, 9-12 September, 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[7] J. Parapar, A. Perez, X. Wang, F. Crestani, Overview of erisk 2025: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, Part II, Lecture Notes in Computer Science, Springer, 2025.

[8] G. Amati, BM25, Springer US, Boston, MA, 2009, pp. 257–260. URL: https://doi.org/10.1007/978-0-387-39940-9_921. doi:10.1007/978-0-387-39940-9_921.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423/. doi:10.18653/v1/N19-1423.

[10] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly available clinical BERT embeddings, in: A. Rumshisky, K. Roberts, S. Bethard, T. Naumann (Eds.), Proceedings of the 2nd Clinical Natural Language Processing Workshop, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 72–78. URL: https://aclanthology.org/W19-1909/. doi:10.18653/v1/W19-1909.

[11] 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: https://arxiv.org/abs/1908.10084.

[12] D. R. Cox, The regression analysis of binary sequences, Journal of the Royal Statistical Society: Series B (Methodological) 20 (1958) 215–232.

[13] C. Cortes, V. Vapnik, Support-vector networks, Machine learning 20 (1995) 273–297.

[14] M. M. Hossain, M. S. Hossain, M. F. Mridha, M. Safran, S. Alfarhood, Multi task opinion enhanced hybrid BERT model for mental health analysis, Scientific Reports 15 (2025) 3332. URL: https://www.nature.com/articles/s41598-025-86124-6. doi:10.1038/s41598-025-86124-6, publisher: Nature Publishing Group.

[15] A. Gaurav, B. B. Gupta, K. T. Chui, BERT Based Model for Robust Mental Health Analysis in Clinical Informatics, in: 2024 21st International Joint Conference on Computer Science and Software Engineering (JCSSE), Phuket, Thailand, 2024, pp. 153–160. doi:10.1109/JCSSE61278.2024.10613729.