Combining Recommender Systems and Language Models in Early Detection of Signs of Anorexia

Notebook for the eRisk 2024 Lab at CLEF 2024

Oskar Riewe-Perła[1,*], Agata Filipowska[1]

[1]Poznan University of Economics and Business, Al. Niepodleglosci 10, 61-875 Poznan, Poland

Abstract

The challenge of accurately classifying events in a stream, despite years of research, still demands attention, especially when it involves understanding complex events. An example of such complexity is identifying tweets written by individuals that may suggest the presence of a disease. This issue was addressed at the eRisk 2024 Lab during CLEF 2024, where the task focused on analyzing evidence to detect early signs of anorexia. The objective of the proposed solution is to determine, as early as possible from a sequence of events (messages), whether an individual is suffering from anorexia. To achieve this goal, we introduce a novel architecture that merges language models with recommender systems, facilitating the fast classification of new messages. Our model demonstrates good performance, with an average F1 score up to 0.68 and a recall rate of 0.97 on a real-life imbalanced dataset.

Keywords

Recommender systems, Language models, BERT

1. Introduction

Social interactions and personal conversations have largely shifted to the digital world due to the limitless availability of the Internet. While the mode of communication changes, the emotions involved remain the same, making social media platforms an important source of information for mental health studies. Text analysis offers a plethora of use cases for detecting mental health issues such as depression, anxiety or eating disorders. As these conditions become more prevalent and have a serious impact on individuals' health, there is a growing need for new detection methods.

Constructing text classification models to analyze a stream of messages presents a significant challenge, particularly when the objective is to identify specific phenomena, such as depression or anorexia, early on, with both high precision and recall, thereby achieving a high F1 score. This task becomes even more complex when the messages are composed by humans using unique online communication language (lingo). Such content can contain crucial information about the user, potentially revealing health issues or situations where intervention is necessary.

In the eRisk 2024 Lab [1, 2], we tackle the challenge of developing a method for early risk detection of anorexia (Task II). Our aim is to sequentially process textual evidence from social media and identify early indicators of anorexia as promptly as possible. To monitor user interactions across blogs, social networks, and other online platforms, texts are processed in the sequence they were posted.

Our research leverages the latest advancements in Natural Language Processing (NLP) along with a recommendation system architecture. By analyzing the behavior of models that suggest content based on user similarities, we assess whether the recommended content originates from individuals experiencing this condition. This approach enables us to issue early warnings for potential risks of exhibiting eating disorder symptoms.

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

[*]Corresponding author.

oskar.riewe-perla@ue.poznan.pl (O. Riewe-Perła); agata.filipowska@ue.poznan.pl (A. Filipowska)
0009-0009-9217-5572 (O. Riewe-Perła); 0000-0002-8425-1872 (A. Filipowska)

2. Proposed Method

2.1. Related work

Throughout the evolution of natural language processing (NLP), studies have highlighted its effectiveness in such areas as healthcare or psychology, i.e. in spotting symptoms and aiding in early detection, which can shorten the time to diagnosis [3, 4, 5]. The advent of extensive datasets from social media has been a game-changer for a field that previously suffered from a lack of measurable data on mental disorders [6]. There's been a noticeable change in the methodologies employed, paralleling NLP advancements. Initially, statistical models were the state of the art for analyzing natural language, often using Bag of Words (BoW) or Term Frequency-Inverse Document Frequency (TF-IDF) techniques [7, 8]. The advent of vectorization methods like Word2Vec or GloVe significantly enhanced NLP task outcomes by better capturing contextual nuances [9, 10, 11, 12]. More recently, the introduction of transformer-based language models, has revolutionized the field, establishing new benchmarks. Fine tuned models like BERT showcase state-of-the-art results on text classification tasks for mental health studies, outperforming statistical models like SVM, but also other deep learning architectures like LSTM or CNN [13].

The method used to distinguish between various message types is known as document classification. This traditional task involves, for example, assigning a tag or class to a document that appears within a stream of messages or is part of a collection. From a different angle, a recommendation system undertakes a similar task by attempting to predict whether a document aligns with user preferences, effectively performing a type of classification on a set of messages. Traditionally, the recommendation systems have been designed in two variants: content-based and collaborative recommendation systems. The former relies on aligning the characteristics of user profiles with the attributes of items, aiming to recommend content that is similar, and becomes more effective as it gathers more information about the user [14]. However, it encounters difficulties with the user cold start problem. The collaborative method, in contrast, seeks to identify similarities among users to introduce a variety of content, yet it faces challenges with new items that lack initial ratings, known as the item cold start problem. To leverage the strengths and overcome the limitations of both approaches, hybrid methodologies have been developed and are proving to be an effective strategy for integrating both item and user features to generate recommendations [15, 16, 17]. These systems have been successfully implemented across various domains, leading to the idea of applying them to the classification of user-generated messages.

The application of Deep Learning techniques has become increasingly evident in the field of recommender systems, particularly those that process textual or visual inputs and where contextual details are significant [18, 19]. Additionally, models originally designed for language processing, such as BERT, have been adapted to improve recommendation generation, surpassing the performance of existing top models [20]. Despite the widespread use of NLP techniques to enhance the effectiveness of recommendation systems, there appears to be a lack of research on employing these systems directly for the classification of mental health disorders.

2.2. Architecture

We propose a model designed for the early identification of potential mental health issues, utilizing recommender systems enhanced by natural language processing (NLP) techniques. This model is built upon three foundational elements: document embeddings, user embeddings, and a recommendation engine. The initial two components employ language models to transform textual data into numerical representations. Our approach incorporates the Sentence Transformer architecture, creating document embeddings that are well-suited for similarity assessments. This process facilitates the generation of user profiles based on the transformed data. By converting these textual and user representations into a format analogous to the classic recommendation scenario involving items and users, we can interpret a user's posting history as a matrix of interactions. In the final stage, we deploy a hybrid recommendation system that leverages both document and user embeddings to tailor recommendations to individual user profiles. The system evaluates recommended posts in relation to their authors, where a higher proportion of contributions from users experiencing mental health challenges is indicative of a positive detection.

2.3. Document embeddings

A key component of the model is the representation of text, which influences both the final document and the representation of the user. Among various NLP techniques suitable for this task, Sentence Transformers have been shown to produce high-quality embeddings that are ideal for similarity comparisons [21]. This approach not only eliminates the need for prior preprocessing but also accommodates raw text and supports multi-lingual datasets. It generates a fixed-sized numerical vector as output, which is applicable irrespective of the text's length. We decided to use *all-mpnet-base-v2* model, which is listed on top of the ranking[1] published by authors of the SBERT framework.

The size of the document embeddings is determined by the chosen model; in our case, they are 768-dimensional vectors. The high dimensionality complicates comparison using conventional methods. Although there are methods to handle sparse data, we recommend employing dimensionality reduction techniques to decrease the time required for generating recommendations. This strategy is advocated for e.g., Topic Modeling [22] and has been shown to effectively preserve the contextual information of texts in a more generalized form. For this purpose, we have utilized the UMAP model [23], which can be trained in a supervised manner. Providing binary labels for the training embeddings enhances the model's ability to distinguish between text representations from two distinct user groups.

2.4. User embeddings

To address the challenge of comparing authors and overcoming the issue of user cold start, we create user embeddings. This process falls under the scope of Author Profiling in the field of NLP. While our methodology primarily generates these embeddings from a history of documents authored by the users, it is possible to enrich them with supplementary data such as age, geographical location, profession, etc. For every user, we compute the average of their document embeddings, resulting in a fixed-size vector that distinctly characterizes authors based on their written works.

2.5. Recommender System

Document and user embeddings, along with an interaction matrix, constitute the input for our recommender system. To leverage the features of texts and their authors, and to accommodate profiles of new users, we opted for a hybrid approach - LightFM [17], which integrates content-based methods with collaborative filtering.

2.6. Classification

For every new user, we apply our model to generate recommendations of posts written by authors from the training dataset. From the top 20 documents, we measure the share of those coming from authors with anorexia. The positive classification is given for the users with scores higher than the predefined threshold of 50%. Additionally, to prevent too early classification based on limited information, we have set our threshold to 100% for the first five iterations. Meaning that the model delays triggering the alert, unless all recommended posts come from authors struggling with anorexia. According to the eRisk 2024 task constraints, positive classification also stops further processing of the user's stream of writings.

3. Experimentation and results

Our research involved analyzing a dataset aimed at the Early Detection of Mental Health Disorders [24]. This dataset comprises 823,850 social media posts along with their titles from 1,287 participants. Each

[1] https://www.sbert.net/docs/pretrained_models.html

Table 1
Model results for test and evaluation datasets.

	F1	Precission	Recall	ERDE$_5$	ERDE$_{50}$
Test dataset	0.68	0.56	0.85	0.067	0.026
Evaluation dataset	0.62	0.45	0.97	0.07	0.02

entry is timestamped, which facilitates sequential analysis, a point elaborated upon in discussions about the online learning component of our study. A notable aspect of this dataset is the uneven distribution between users dealing with anorexia (134) and the rest of the participants (1153). In our model, we leveraged language models for text representation, which allowed us to bypass traditional preprocessing steps. We combined the title and body of each post into a single text string for analysis.

In the experiment we perform, our model receives continuous updates with new posts from each user, mimicking the natural pattern of social media activity. The model is not retrained; instead, it updates the user profile for more precise recommendations. Once an alert for a user is activated, the model ceases to analyze subsequent posts from that individual.

Our findings indicate that posts from users with an eating disorder are more likely "to be recommended" to others facing similar challenges, aiding in early detection. We evaluated the effectiveness of our model by the proportion of recommendations that come from users with eating disorders, using a numerical score to determine when a positive classification is warranted.

The model was evaluated by us on a test dataset, as well as by eRisk Lab authors using a separate evaluation dataset, that was not accessible to us before the evaluation process. The model exhibits similar results on both datasets, as presented in Table 1, demonstrating its strong generalization capability. It has achieved very low ERDE loss (ranging between 0.02 and 0.07), indicating its ability of triggering an alert with minimal delay and proving that small number of social media posts is enough to perform accurate classification. The model's high recall score (0.85-0.97) indicates its proficiency in correctly identifying most positive cases. However, the lower precision (0.45-0.56) suggests that the model mistakenly identifies some negative samples as positive, an issue that could potentially be mitigated by increasing the classification threshold.

4. Discussion and future work

We have implemented a mental health disorder detection model that achieves robust F1 score of 0.68, high recall rate of 0.97, while maintaining a low ERDE loss ranging between 0.02 and 0.07. This model leverages advancements in the Natural Language Processing and combines them with recommender system architecture to detect similarities between users with and without anorexia.

This research contributes to the broader domain of tasks focused on early detection and intervention for mental health disorders through digital platforms. We suggest expanding already existing use cases for recommender systems with NLP methods for proactive health management, addressing the urgent need for early identification of eating disorders. The overall good results prove the feasibility of employing recommender systems in the task of mental health issues detection. Its simplicity and flexibility shows a potential in applying it in real-life application and enhancing the social impact of recommender systems. Its modular design ensures ongoing development in parallel with the advancements of NLP and emergence of more powerful language models. The promising results indicate the importance of continuous improvement and adaptation of these models to keep pace with the evolving nature of online communication and mental health challenges.

Future work should focus on improving the model's precision score, to eliminate false negative predictions. We propose investigating the optimal delay before triggering an alert, as too short posts might initially mislead the model. A similar approach to detect an adequate amount of data enough for prediction was previously suggested [25], by calculating the combined size of the text taken into analysis, instead of only relying on number of posts. Moreover, our suggested technique for creating user embeddings rely solely on the averaged document embeddings. More time-aware methods could be considered to reflect the transition of mental state between the posts, especially as post creation timestamps are available in the dataset.

References

[1] J. Parapar, P. Martín-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, Grenoble, France, 2024.

[2] 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, CLEF 2024. CEUR Workshop Proceedings, Grenoble, France, 2024.

[3] N. Viani, L. Yin, J. Kam, A. Alawi, A. Bittar, R. Dutta, R. Patel, R. Stewart, S. Velupillai, Time expressions in mental health records for symptom onset extraction (2018). doi:10.18653/v1/w18-5621.

[4] T. Zhang, A. Schoene, S. Ji, S. Ananiadou, Natural language processing applied to mental illness detection: a narrative review, NPJ Digital Medicine 5 (2022). doi:10.1038/s41746-022-00589-7.

[5] A. Wongkoblap, M. A. Vadillo, V. Curcin, Researching mental health disorders in the era of social media: systematic review, Journal of medical Internet research 19 (2017) e228.

[6] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in twitter, in: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, 2014, pp. 51–60.

[7] S. Paul, S. K. Jandhyala, T. Basu, Early detection of signs of anorexia and depression over social media using effective machine learning frameworks, in: Conference and Labs of the Evaluation Forum, 2018. URL: https://api.semanticscholar.org/CorpusID:51942457.

[8] M. Stankevich, V. Isakov, D. Devyatkin, I. V. Smirnov, Feature engineering for depression detection in social media., in: ICPRAM, 2018, pp. 426–431.

[9] F. Almeida, G. Xexéo, Word embeddings: a survey (2019). doi:10.48550/arxiv.1901.09069.

[10] F. Zhang, M. Jiang, W. Li, Z. Weng, Question amp; answering system based on retrieval, Applied and Computational Engineering 4 (2023) 301–307. doi:10.54254/2755-2721/4/20230477.

[11] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532–1543.

[12] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781 (2013).

[13] X. Wang, S. Chen, T. Li, W. Li, Y. Zhou, J. Zheng, Q. Chen, J. Yan, B. Tang, et al., Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis, JMIR medical informatics 8 (2020) e17958.

[14] P. Lops, M. De Gemmis, G. Semeraro, Content-based recommender systems: State of the art and trends, Recommender systems handbook (2011) 73–105.

[15] S. Geuens, Factorization machines for hybrid recommendation systems based on behavioral, product, and customer data, in: Proceedings of the 9th ACM conference on recommender systems, 2015, pp. 379–382.

[16] M. A. Ghazanfar, A. Prugel-Bennett, A scalable, accurate hybrid recommender system, in: 2010 Third International Conference on Knowledge Discovery and Data Mining, IEEE, 2010, pp. 94–98.

[17] M. Kula, Metadata embeddings for user and item cold-start recommendations, in: T. Bogers, M. Koolen (Eds.), Proceedings of the 2nd Workshop on New Trends on Content-Based Recommender Systems co-located with 9th ACM Conference on Recommender Systems (RecSys 2015), Vienna, Austria, September 16-20, 2015., volume 1448 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2015, pp. 14–21. URL: http://ceur-ws.org/Vol-1448/paper4.pdf.

[18] S. Zhang, L. Yao, A. Sun, Y. Tay, Deep learning based recommender system, Acm Computing Surveys 52 (2019) 1–38. doi:10.1145/3285029.

[19] S. Jeong, Y. Kim, Deep learning-based context-aware recommender system considering contextual features, Applied Sciences 12 (2021) 45. doi:10.3390/app12010045.

[20] F. Sun, J. Liu, J. Wu, C. Pei, X. Lin, W. Ou, P. Jiang, Bert4rec: Sequential recommendation with bidirectional encoder representations from transformer, in: Proceedings of the 28th ACM international conference on information and knowledge management, 2019, pp. 1441–1450.

[21] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[22] M. Grootendorst, Bertopic: Neural topic modeling with a class-based tf-idf procedure, arXiv preprint arXiv:2203.05794 (2022).

[23] L. McInnes, J. Healy, J. Melville, Umap: Uniform manifold approximation and projection for dimension reduction, arXiv preprint arXiv:1802.03426 (2018).

[24] F. Crestani, D. E. Losada, J. Parapar, Early risk prediction of mental health disorders, Early Detection of Mental Health Disorders by Social Media Monitoring (2022) 1–6. doi:10.1007/978-3-031-04431-1_1.

[25] D. Ramírez-Cifuentes, M. Mayans, A. Freire, Early risk detection of anorexia on social media, in: Internet Science: 5th International Conference, INSCI 2018, St. Petersburg, Russia, October 24–26, 2018, Proceedings 5, Springer, 2018, pp. 3–14.