BFH-AMI at eRisk@CLEF 2023*

Ghofrane Merhbene[2], Alexandre R. Puttick[2] and Mascha Kurpicz-Briki[2]

[2]Berner Fachhochschule BFH, Applied Machine Intelligence, Höheweg 80, 2502 Biel, Switzerland

Abstract

Mental health problems are a rising problem of today's society. Methods of machine learning and natural language processing provide interesting new possibilities for psychology and psychiatry. In particular, eating disorders (ED) are widespread and can be life-threatening if untreated. This paper describes the approach to Task 3 of the eRisk 2023 challenge of the BFH-AMI team. The task concerned the prediction of patients' answers to the Eating Disorder Examination Questionnaire (EDE-Q) based on their social media writing history. In our approach, we used a logistic regression model that was fed with a combination of user and question embeddings from the GPT-2 Large model.

Keywords

Early Detection System, Natural Language Processing, Machine Learning, Eating Disorder, Mental Health

1. Introduction

Eating disorders (EDs) represent a severe and potentially life-threatening mental health condition, especially if left untreated. They encompass a range of complex conditions characterized by disturbances in eating behaviors, distorted body image, and major psychological distress. The impact of these disorders is widespread, affecting millions of individuals worldwide. For instance, research conducted in 2015 revealed that anorexia, which is a common type of eating disorder, had already affected more than 2.9 million people [1]. Such statistics highlight the magnitude of the issue and emphasize the urgent need for effective intervention and treatment strategies.

Early detection and severity assessment of signs associated with EDs is paramount for effective intervention. Traditionally, the assessment of the severity of EDs has heavily relied on clinical evaluations which are known to often be time-consuming and labor-intensive. This resource-intensive assessment can be facilitated with computational approaches that can provide efficient pre-assessments on the severity of EDs.

The CLEF eRisk[1] Challenge is an academic research competition that encourages participants to develop text-based innovative solutions toward understanding health-related data. In its third task of the 2023 edition, the focus was on using Natural Language Processing (NLP) techniques to assess the severity levels of ED symptoms. To solve the task, the participants of the challenge were asked to design systems for predicting responses to an eating disorder questionnaire for

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

ghofrane.merhbene@bfh.ch (G. Merhbene); alexandre.puttick@bfh.ch (A. R. Puttick); mascha.kurpicz@bfh.ch (M. Kurpicz-Briki)
0000-0002-2142-0309 (A. R. Puttick); 0000-0001-5539-6370 (M. Kurpicz-Briki)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://erisk.irlab.org/

different patients, based on a history of their postings from social media. The Eating Disorder Examination Questionnaire (EDE-Q)[2] [2] was used to collect comprehensive and reliable data regarding participants' eating behaviors, body image concerns, and psychological distress associated with their eating disorder. The questionnaire covers multiple domains, including dietary restraint, eating concerns, shape concerns, and weight concerns.
In this paper, we document the approach of our research team BFH-AMI in this third task. Our aim was to leverage state-of-the-art NLP techniques to develop an efficient methodology for automatically detecting the severity of the signs of EDs. In the long term, technologies such as the one developed in this challenge can be further enhanced and validated as clinical tools. Such tools can support clinical professionals in their tasks and provide them with additional new insights based on data.

2. Related work

Traditional approaches conducted by clinical professionals, such as psychologists and therapists, to assessing people's emotions and traits through survey questionnaires and interviews have limitations in terms of cost, time, and scalability. However, recent advancements in NLP techniques offer promising new options to address these challenges and support the clinical professionals. Recent research has delved into various approaches to automate the identification of eating disorders. For example, López-Úbeda et al. [3] explored a range of strategies, including different machine learning techniques. They conducted experiments using five supervised learning models on a Spanish Anorexia dataset and achieved an F1-score of over 0.9 using Support Vector Machines (SVM) and Multilayer Perceptron (MLP). Other studies explored alternative techniques, such as Convolutional Neural Networks (CNN) and Short Term Memory (LSTM), e.g., [4].
Moreover, the exploration of recent NLP technologies, like BERT [5] embeddings, has demonstrated promise in predicting questionnaire responses by leveraging text from social media and survey questions, as demonstrated by Vu et al. [6] using a novel technique developed to address this task. By analyzing participants' social media texts and the text of the survey questions they are asked, the researchers used BERT to represent both the participants and the survey questions as embedding vectors. This enabled the prediction of responses for both new participants and new questions not seen during training. This method offers the possibility to study new participants or new questions without the constraints of costly data collection.
The proposed approach not only facilitates novel practical applications but also contributes to the advancement of psychological theory. Furthermore, the success of the model suggests a promising NLP-powered alternative to the resource-intensive use of traditional assessment methods.

3. Task and Data

During the training phase of the challenge, the eRisk team provided the entire history of writings and corresponding answers to the Eating Disorder Examination Questionnaire (EDE-Q) for

[2] https://www.corc.uk.net/media/1273/ede-q_quesionnaire.pdf

Table 1
Training data statistics

Nb. of Subjects	28
Min. Nb. of posts per Subject	12
Max. Nb. of posts per Subject	1143
Avg. Nb. of characters per Post	184.33

Table 2
Test data statistics

Nb. of Subjects	46
Min. Nb. of posts per Subject	5
Max. Nb. of posts per Subject	1161
Avg. Nb. of characters per Post	223.25

a specific set of training users. This allowed the participants to train their systems using the provided data.

The EDE-Q questionnaire consists of 28 items out of which only questions 1-12 and 19-28 were considered for the purpose of this competition. The questionnaire is designed to assess the range and severity of characteristics associated with a diagnosis of an eating disorder and it includes four sub-scales: Restraint, Eating Concern, Shape Concern, and Weight Concern, as well as a global score.

The training set consisted of 28 subjects. Each subject had a history of postings from the social media platform Reddit[3] as well as their answers to the EDE-Q questionnaire, the latter of which serving as ground truth labels for the task. This combined data allows for a comprehensive examination and analysis of the subjects' online interactions and self-reported information.

During the test stage, the writing history of a new set of users was provided. However, the test set did not include the answers to the questionnaire. Using the trained models, participants of the task had to generate predictions for the EDE-Q questionnaire. The testing set consisted of the writing history of 46 subjects on Reddit and was structured in a similar manner to the training set.

The statistics for the training and test data are presented in Table 1 and Table 2, respectively.

4. Evaluation Metrics

The evaluation of system performance in this third task is based on several measures of effectiveness. These measures have been defined by the organizers as follows:

- **Mean Zero-One Error (MZOE)**: To measure the average error rate.
- **Mean Absolute Error (MAE)**: To measure the deviation of the model's predictions from the actual values.

[3] https://www.reddit.com/

- **Macroaveraged Mean Absolute Error (MAEmacro)**: Is similar to MAE. It is the mean absolute difference for each class independently and then averages them across all classes. Here a class is defined as the set of all questions Q_i whose true answer is equal to $i \in \{0, 1, \ldots, 6\}$

MZOE, MAE, and MAEmacro each calculate a single score for every user, and the reported score is the average of all these values.

The measures presented below are derived from aggregated scores obtained from the questionnaires:

Restraint Subscale (RS):

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{\text{RS}}(u_i) - f_{\text{RS}}(u_i))^2}{|U|}}$$

where U is the user set, R_{RS} represents the real subscale ED score for user u_i, and F_{RS} represents the estimated subscale ED score for user u_i. The reported RMSE is the average over all RMSE values (mean RMSE over all users).

Eating Concern Subscale (ECS):

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{\text{ECS}}(u_i) - f_{\text{ECS}}(u_i))^2}{|U|}}$$

where R_{ECS} represents the real eating concern ED score for user u_i, and R_{ECS} represents the estimated eating concern ED score for user u_i. The reported RMSE is the average over all RMSE values (mean RMSE over all users.

Shape Concern Subscale (SCS):

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{\text{SCS}}(u_i) - f_{\text{SCS}}(u_i))^2}{|U|}}$$

where R_{SCS} represents the real shape concern ED score for user u_i, and F_{SCS}) represents the estimated shape concern ED score for user u_i. The reported RMSE is the average over all RMSE values (mean RMSE over all users).

Weight Concern Subscale (WCS):

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{\text{WCS}}(u_i) - f_{\text{WCS}}(u_i))^2}{|U|}}$$

where R_{WCS} represents the real weight concern ED score for user u_i, and F_{WCS} represents the estimated weight concern ED score for user u_i. The reported RMSE is the average over all RMSE values (mean RMSE over all users).

Global ED (GED):

$$RMSE(f, U) = \sqrt{\frac{\sum_{u_i \in U}(R_{\text{GED}}(u_i) - f_{\text{GED}}(u_i))^2}{|U|}}$$

A global score can be calculated by adding the scores of the four subscales scores and then dividing the resulting total by 4.
Additional details and information regarding the specific evaluation metrics employed during the evaluation phase can be accessed in the Overview of eRisk competition [7].

5. Methodology

Our main task is to assess to what extent the characteristics linked to the diagnosis of EDs in the questionnaire are reflected in a history of user writings. To accomplish this, we generated various embeddings for both user posts and questions. These embeddings were then combined and used as input for a logistic regression model. The process of generating the embeddings is described in detail below.
It is important to note that a single run was submitted to the challenge. Our submission involved a logistic regression model with an L2-regularization parameter = 1/50, which was chosen after fine-tuning using a hyperparameter search with values [1, 1/10, 1/20, 1/30, 1/40, 1/50]. All other models use L2-regularization parameters = 1. In the final model, the embeddings were obtained using GPT-2, combined with a method for extracting the most relevant user sentences via cosine similarity. This was the result of several incremental improvements detailed below.

5.1. Embeddings

To generate both the question embeddings as well as the user embeddings, we tried the method detailed below using two different models. This method was built upon techniques used for a similar task in [6]:

- **BERT Large (uncased)** [5] is a pre-trained language model with 336 million parameters. The ”uncased” aspect means that the model treats capitalization as irrelevant and converts all text to lowercase during training. This allows for better generalization across different cases of the same word.
- **GPT-2 Large** [8] is a pre-trained language model with 774 million parameters, which gives it an impressive ability to generate coherent and contextually relevant text.

Question Embeddings
To compute the embeddings for the 22 questions from the EDE-Q questionnaire, we used one of the two pre-trained models described above. To begin, we extracted hidden vectors from the last four layers of the model corresponding to each word in the input text, which contain valuable information regarding the semantic representation of the questions. We averaged these four embedding vectors over all of the words in a given question, resulting in four vectors representing that question. These four vectors were concatenated to obtain the final question embeddings. This technique, yielded in embedding vectors of dimension 4096 in the case of BERT Large and 5120 in the case of GPT-2 Large.

User Writings Embeddings
Method 1, Chunk Embeddings: Because some user posts were too long to feed into the model, we concatenated all user posts together (in chronological order), and broke the results text

into chunks of text corresponding of length *n* tokens, where *n* is the maximum input sequence length of the model (512 for BERT Large and 1024 for GPT-2 Large). The embeddings for each text chunk were computed in a manner identical to the one used to obtain question embeddings described above. Afterwards, these chunk embeddings were averaged to obtain user embeddings.

Method 2, Sentence Extraction: Since users' writing histories contained many posts that were not relevant for the task, we attempted to derive a method for extracting those sentences that were most relevant for the prediction of the corresponding user's degree of ED symptoms. To do this, we computed sentence embeddings for every sentence written by a given user in the same manner as for the question and text-block embeddings described above. These were compared to the 'topic' vector obtained by averaging all of the question embeddings to obtain a single vector. We extracted the 20 user-written sentences that were closest to this topic vector with respect to cosine similarity and averaged the corresponding sentence embeddings to obtain user embeddings. Note that this technique was only applied using the GPT-2 large model, and not the BERT model.

5.2. Baseline

To assess the performance of our model, we relied on a simple baseline approach for comparison. In this baseline method, for each question, the prediction is made by taking the average of all the users' answers from the training data. This served as a basic benchmark against which we could measure the effectiveness of our model.

6. Results

Table 3 summarizes our results when evaluating our different approaches using 10-fold cross-validation on the training data. Each sample consisted of the concatenation of the user embedding and question embedding pair, labeled by the user's response to the corresponding question. For each 10-fold split, a 7-class (responses from 0 to 6) logistic regression classifier was trained on nine folds and tested on the remaining fold. The evaluation metrics were computed by averaging over the ten folds.

The GPT-2 model with sentence extraction outperformed all other models, which is why it was chosen to be submitted to the competition. The performance results on the test data across all the metrics described in Section 4 are presented in Table 4.

Table 3
Performance over training data using 10-fold cross validation

Model	MZOE	MAE_{macro}	GED
Baseline average	0.96	2.10	1.96
GPT-2 with sentence embeddings (L2 = 1/50)	**0.73**	**1.30**	**1.37**
GPT-2 with chunk embeddings (L2 = 1)	0.78	1.50	1.61
BERT with chunk embeddings (L2 = 1)	0.78	1.70	2.20

Table 4
Performance over test data obtained by the Logistic regression and GPT-2 large sentence based embeddings

	MAE	MZOE	MAE_{macro}	GED	RS	ECS	SCS	WCS
Baseline all 0s	2.419	0.674	2.803	3.207	2.138	3.221	3.028	2.682
Baseline all 6s	3.581	0.834	3.995	3.839	4.814	3.650	3.950	3.318
Baseline average	2.091	0.859	1.957	2.391	1.592	2.398	2.162	2.002
BFH-AMI	2.407	0.719	2.729	3.169	2.597	2.854	2.923	2.144

Although our model performed considerably better than the baseline model during development, it only outperformed the baseline according to the MZOE metric on the test data. Having only submitted one run, it is difficult to discern the cause for this, but random chance associated to a very small training set (n = 28 users) maybe have played a role. We also observed significant differences in formatting between training and test data, which may have negatively affected performance. Given formatting differences and the inability to troubleshoot on test data, we cannot rule out simple implementation errors. Qualitative analysis of the sentences extracted using our cosine similarity based criteria suggest that the method indeed extracts sentences speaking about mental and physical health, food, weight etc., although plenty of less relevant sentences were also extracted, and we did not cross-check our method to gauge if the most important sentences were indeed extracted.

7. Conclusion

This paper documents the participation of our team BFH-AMI in the task 3 of the eRisk@CLEF 2023 edition. We investigated the severity of the signs of eating disorders, by developing a model that automatically generates responses to questions from the EDE-Q questionnaire, based on the user's writings on social media provided in anonymous form by the organizers. In our proposed approach, we used a logistic regression model that was fed with a combination of user and question embeddings extracted from the GPT-2 Large model.

The performance metrics demonstrate that there is substantial room for improvement across various evaluation criteria. Future work could investigate the following directions:

- *More powerful language models:* Larger language models could be employed in an attempt to capture more semantic information in user and question embeddings. However, experiments carried out using the 1.3 billion parameter version of GPT-Neo [9] did not yield significantly better results, although we did not have time to make a thorough comparison.
- *Improved sentence extraction:* Our sentence extraction method was based on comparison to a topic vector obtained by averaging all questions. In general, the more text that is averaged into an embedding vector of fixed length, the more the distinguishing features can become smoothed out. Therefore, it might be preferable to average only questions corresponding to a specific dimension of the EDE-Q, or even extract sentences separately for each question. We also observed that many of the extracted sentences were also

questions. This is not so relevant for the task, but is a feature that was likely encoded in the question embeddings and carried over to the most similar user sentences. In the future, this could be avoided by rephrasing each question into an analogous first person statement such as "I have been deliberately trying to limit the amount of food I eat to influence my shape or weight." Instead of taking the top 20 sentences, one might instead take only the sentences above a certain similarity threshold. This would allow flexibility in the amount of sentences extracted for each user, given relatively few when a user does not write about relevant topics and many when the user does.

- *Improved embedding methods:* Our methods relied on averaging over many words and sentences, following the methods in [6]. For transformer models like the ones used here, in cases where the input text is not too long, the embeddings obtained from only the final word in the sequence should contain information about the context preceeding that word. As mentioned above, averaging could have an undesired smoothing effect on the embedding vectors, and it may be preferable to either use only such "last word" embeddings or devise other strategies (such as sentence extraction), for decreasing the amount of text aggregated into each embedding vector. Furthermore, while we always constructed embedding vectors using the final four layers of the models, this number four could also be considered a hyperparameter and adjusted for ideal performance.
- *More deep learning:* In our methods, we only used deep learning models for feature extraction to then feed into a classical machine learning classifier (logistic regression). Fine-tuning weights within the large language models could improve performance, although, with so few data samples, there is a high danger of over-fitting.

Acknowledgments

The authors gratefully acknowledge the support of the Inventus Bern Foundation in our research in the field of augmented intelligence for the detection of eating disorders.

References

[1] T. Vos, et al., Global, regional, and national incidence, prevalence, and years lived with disability for 310 diseases and injuries, 1990–2015: a systematic analysis for the global burden of disease study 2015, The Lancet 388 (2016) 1545–1602. doi:10.1016/S0140-6736(16)31678-6.

[2] C. G. Fairburn, Z. Cooper, M. O'Connor, Eating Disorder Examination, 17.0D.

[3] P. López Úbeda, F. M. Plaza del Arco, M. C. Díaz Galiano, L. A. Urena Lopez, M. Martin, Detecting anorexia in Spanish tweets, in: Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019), INCOMA Ltd., Varna, Bulgaria, 2019, pp. 655–663. URL: https://aclanthology.org/R19-1077. doi:10.26615/978-954-452-056-4_077.

[4] N. Liu, Z. Zhou, K. Xin, F. Ren, TUA1 at erisk 2018, in: L. Cappellato, N. Ferro, J. Nie, L. Soulier (Eds.), Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018, volume 2125 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2018. URL: http://ceur-ws.org/Vol-2125/paper_121.pdf.

[5] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[6] H. Vu, S. Abdurahman, S. Bhatia, L. Ungar, Predicting responses to psychological questionnaires from participants' social media posts and question text embeddings, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, pp. 1512–1524. URL: https://aclanthology.org/2020.findings-emnlp.137. doi:10.18653/v1/2020.findings-emnlp.137.

[7] 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, Springer International Publishing, Thessaloniki, Greece, 2023.

[8] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[9] S. Black, L. Gao, P. Wang, C. Leahy, S. Biderman, GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow, 2021. URL: https://doi.org/10.5281/zenodo.5297715. doi:10.5281/zenodo.5297715, If you use this software, please cite it using these metadata.