CEUR-WS.org/Vol-3740/paper-75.pdf

ELiRF-VRAIN at eRisk 2024: Using LongFormers for Early Detection of Signs of Anorexia

Andreu Casamayor[1], Vicent Ahuir[1], Antonio Molina[1,*] and Lluís-Felip Hurtado[1]

[1] *Valencian Research Institute for Artificial Intelligence (VRAIN), Universitat Politècnica de València, Camino de Vera s/n, 46022 Valencia. Spain*

Abstract

This paper describes the approaches taken by the ELiRF-VRAIN team at the Task 2 of eRisk at CLEF 2024 focused on the early detection of signs of anorexia on English-language social media. Our work involved three distinct approaches: one using a Support Vector Machine (SVM) and the other two based on pre-trained Transformer models. Among the Transformer models, one approach employed BERT-like models, while the other used LongFormer models. To fine-tune our models, we implemented a data augmentation process on the dataset provided by the organization. In the validation phase, the models trained on the augmented dataset improved the F1 score results. In particular, F1 increased from 0.89 to 0.94 for the LongFormer model. During the testing phase the SVM model and LongFormer with data augmentation obtained the best results. LongFormer improved BERT-like model performance due to its ability to handle large contexts. Seeing the results achieved in the validation phase, we can say that the overall performance was not as good as expected. A detailed analysis of the results would be necessary to find out the reasons.

Keywords

Longformers, Transformers, Support Vector Machine, Anorexia

1. Introduction

Anorexia nervosa is the formal term for anorexia, and it's a complex, really multi-structural eating disorder. This is a disorder characterized by a fear of gaining weight and by the maintenance of a distorted body image through severe food restriction and excessive weight loss. It is hazardous for both males and females, but is most common among young women. Women account for 90-95% of those affected; the age range is usually between 12 and 25 years, and it is most common between 12 and 17 years of age. [1]

The impacts of anorexia extend to all aspects of one's health and functioning, extending far beyond malnutrition to nearly every organ system in the body, even when comorbid with other mental health issues like depression and anxiety. Little is done, anorexia is often difficult to detect and treat due to its insidious onset and the societal stigma surrounding mental health and eating disorders.

For this reason, the analysis of social interactions to detect risks of anorexia has recently become one of the most important ways of detection. This type of problem, anorexia detection, is complicated due to some reasons, such as the amount and quality of the data. CLEF eRisk created different tasks, to provide quality data and promote the creation of models for this early detection.

In 2024's edition, eRisk proposed three shared tasks [2, 3]: (1) Search for symptoms of depression, (2) Early Detection of Signs of Anorexia, and (3) Measuring the severity of the signs of Eating Disorders.

We focused our participation on the second shared task, where we used three different approaches to tackle the problem posed by the task:

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

[*] Corresponding author.

[†] These authors contributed equally.

ancase3@upv.es (A. Casamayor); vahuir@dsic.upv.es (V. Ahuir); amolina@dsic.upv.es (A. Molina); lhurtado@dsic.upv.es (L. Hurtado)

https://vrain.upv.es/elirf/ (A. Casamayor); https://vrain.upv.es/elirf/ (V. Ahuir); https://vrain.upv.es/elirf/ (A. Molina); https://vrain.upv.es/elirf/ (L. Hurtado)

0009-0003-6000-3828 (A. Casamayor); 0000-0001-5636-651X (V. Ahuir); 0000-0001-6537-8803 (A. Molina); 0000-0002-1877-0455 (L. Hurtado)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

1. The initial approach employs a traditional machine learning algorithm, Support Vector Machines (SVM). SVMs have shown meaningful performance in classifying lengthy texts, similar to this case. We use this approach to evaluate the effectiveness of classical models.
2. The second approach utilizes Transformers [4] by leveraging a pre-trained RoBERTa model [5] as a foundation, followed by a fine-tuning process to adapt it to the downstream task. We performed fine-tuning using two distinct datasets: one provided by the organization and the other created through data augmentation.
3. The final approach is similar to the second one but aims to capture more context by using a pre-trained LongFormer model [6]. This model accommodates larger input sizes, allowing it to grasp more contextual information. We fine-tuned the LongFormer model using the same dataset as in the previous approach.

We submitted four runs for Task 2, one for approaches 1 and 2, and two for approach 3. Before selecting the best model for each approach, we put them through a validation phase, where we tested different configurations and datasets used.

We have done this kind of experimentation before and had excellent results, proving how reliable and effective our approach is. In related topic works, we used similar methods and achieved substantial outcomes [7].

2. Description of Dataset and Task

Task 2 involves the early detection of anorexia risk by sequentially analyzing pieces of evidence to identify early signs of the disorder as promptly as possible. This task primarily focuses on evaluating natural language processing solutions, particularly those that analyze texts from social media. Texts must be processed in the chronological order in which they were created. This simulates better what the system would do: monitor real-time user interactions on blogs, social networks, or other online platforms.

The dataset in Task 2 consisted of a writing (post or comments) collection from a set of Social Media users formed from the datasets from previous editions of the task in 2018 and 2019. This collection has the same format as the one delivered in [8], where there are two different classes: users who suffer from anorexia and a control group (non-anorexia). Every user has a chronological collection of messages or writings.

Table 1 shows the distribution among the different labels in the dataset

Table 1
Distribution of samples across the 2018 and 2019 partitions of the Task 2 dataset.

	2018	**2019**	**Total**
None	411	742	1153
Anorexia	61	73	134
Total	472	815	1287

As mentioned, the primary goal of this competition is to predict signs of anorexia as promptly as possible. To simulate realistic conditions, the organizers set up a server that sequentially delivers data packets, each containing a message from a user. The system must predict the user's signs of anorexia, if any, by considering both the current message and all previous messages before receiving the next data packet.

3. Systems and Architecture and Techniques

In this type of task, a relevant factor to consider is the amount of context required for accurate detection. Since each user can have numerous messages, the size of the input to the system becomes a crucial consideration. One of our team's objectives was to examine the impact of context in these tasks.

Specifically, we aimed to evaluate the performance of different systems based on their ability to handle varying amounts of context. We selected three different systems to achieve this goal: the first based on Support Vector Machines (SVM), the second based on a RoBERTa model, and the third based on LongFormer model. Each system evaluated has a different size for context:

- Support Vector Machines (SVM) do not have a fixed limit on input size; they construct a vector with a length corresponding to the vocabulary size. This flexibility allows SVMs to handle a large and variable amount of data, as they can create feature vectors based on the entirety of the input text's vocabulary, accommodating diverse and extensive datasets.
- The selected RoBERTa model has a limit of 512 tokens in the input.
- The selected LongFormer model has a limit of 4096 tokens in the input.

Additionally, we developed two distinct datasets to train and evaluate the performance of the transformer-based systems.

Dataset 1. We created only one sample per user by aggregating all their messages, both for positive and negative labeled users. This approach ensures that the dataset effectively captures the overall context and messaging patterns of every user, facilitating a more accurate evaluation of the models' performance in distinguishing between positive and negative cases.

Dataset 2. If we had some a priori evidence of in which message a user begins to present symptoms of mental illness risk, we could label the samples from previous messages as negative, and the samples containing that message and subsequent ones as positive. In this way, we could increase the number of positive samples to achieve a more precise model. This data augmentation process is explained in the next section.

To conduct our experimentation, we split the original dataset into two partitions: training (80% of users) and development (20% of users). We ensured that both partitions maintained the same proportions of positive and negative samples to preserve the dataset's balance and integrity. Table 2 shows the distribution of samples in Dataset 1.

Table 2
Distribution of samples in Dataset 1 for training and development partitions

	Train	Development
None	920	233
Anorexia	109	25
Total	1029	258

3.1. Data Augmentation

The data augmentation process aims to generate additional samples for each positive user. As mentioned earlier, we need evidence of when a user begins to exhibit signs of anorexia in their messages. To identify this, we relied on predictions from the SVM-based classifier. We assume that all messages preceding the SVM decision point do not express signs of anorexia. To implement this, we followed these steps:

1. For positive users, we calculated how many messages the SVM needs to classify the user as positive. Each user has a different trigger value.
2. For false negatives, we used the mean of the true positive trigger values as the trigger value.
3. For each positive user in the original data set, let n be the number of messages that the SVM model needs to determine this user's mental disorder risk, MAX be the maximum number of messages the model supports as input, and m_i the ith message from the user.
    a) we created $n - 1$ negative samples as follows:

$$(m_1), (m_1 m_2), (m_1 m_2 m_3), ..., (m_1 ... m_{n-1})$$

b) and $MAX - n + 1$ positive samples:

$$(m_1...m_n), (m_1...m_n m_{n+1}), ..., (m_1...m_n...m_{MAX})$$

4. Note that the value of MAX depends on which model was used and the number of tokens in the messages. That is, we discard messages from an accumulated history of more than 512 tokens for RoBERTa and 4096 for LongFormer. So, if $n > MAX$ only negative samples are generated.
5. For negative users, we created new samples accumulating the history as before, stopping when the MAX was reached.

The result of this technique is a new dataset with a higher number of positive samples for the training. In the development partition, we held a sample per user, as in Dataset 1.

Table 3
Distribution of samples in Dataset 2 for training and development partitions

	Train	Development
None	18255	233
Anorexia	2272	25
Total	20527	258

3.2. Classical Machine Learning Classifier Approach

To evaluate the significance of the context, we aimed to use a classical machine learning classifier that is capable of handling all the available context. One of the major issues with Transformer-based models is that their ability to handle large texts is limited by the input size. This greatly affects performance because the input cannot contain the length of the sample, whereby crucial information may be lost. We would use such a classical machine learning model as SVM to create a vector as long as the size of the vocabulary to show the model's performance when it has no such restriction.

First, we experimented to compare different types of classical machine learning classifiers. We utilized the Scikit-learn library [9] for this purpose, employing its default classifiers to identify the best-performing model. The results, presented in Table 4, indicate that the Linear SVM emerged as the top performer among the classifiers tested.

Table 4
The results from different classifiers in the development partition. The scores are the Macro-precision, recall and f1-score.

	precision	recall	f1-score
Linear SVM	**0.83**	**0.80**	**0.81**
Gradient Boosting	0.72	0.75	0.74
K-Neighboors	0.45	0.50	0.47
AdaBoost	0.74	0.74	0.74

Once the classifier was chosen, we wanted to test different approaches:

- **Preprocess of Data**:
  1. First approach: Transform the text into tokens using TweetTokenizer and then eliminate stop words.
  2. Second Approach: Same as the first approach with the addition of methods to clean the text, eliminate non-alphanumerical characters and others, and lemmatize tokens.
- **Sentimental Analysis**: We used the model *"lxyuan/distilbert-base-multilingual-cased-sentiments-student"* [10] to perform sentiment analysis on every user message. This process yielded three

results: positive messages, negative messages, and neutral messages. These results were normalized and subsequently added as a new feature to the TF-IDF representation. This enhancement allowed us to incorporate sentiment-based insights into our analysis, potentially improving the performance and accuracy of our classification models.

- **TF-IDF**: We used the class `TfidfVectorizer` from *Scikit-learn* to vectorize the data. We experimented with different configurations for the *analyzer* and *ngram_range* parameters, while using the default values for other features. This approach allowed us to identify the optimal configuration for the task.

To find the best models for every approach, we did an exhaustive grid search over some specific parameters, such as regularization parameter C, different tols, and different loss.

We obtained 8 different approaches. Table 5 shows the different configurations used in the experimentation, the column TF-IDF refers to the type of analyzers (word or char) used and the number of n-grams. The last column refers to the best model found in the search grid.

Table 5
Summary of the different configurations of the SVM classifiers.

	Preprocess data approach	Sentiment analysis	TF-IDF	Best Model
SVM-1	1	No	"char_wb" , 4-5 n-gram	'C': 100, 'loss': 'hinge', 'tol': 0.01
SVM-2	2	No	"char_wb" , 4-5 n-gram	'C': 100, 'loss': 'hinge', 'tol': 0.01
SVM-3	1	Yes	"char_wb" , 4-5 n-gram	'C': 10, 'loss': 'hinge', 'tol': 0.1
SVM-4	2	Yes	"char_wb" , 4-5 n-gram	'C': 10, 'loss': 'hinge', 'tol': 0.1
SVM-5	1	No	"word" , 1-2 n-gram	'C': 1, 'loss': 'squared_hinge', 'tol': 0.01
SVM-6	2	No	"word" , 1-2 n-gram	'C': 1, 'loss': 'squared_hinge', 'tol': 0.01
SVM-7	1	Yes	"word" , 1-2 n-gram	'C': 10, 'loss': 'hinge', 'tol': 0.1
SVM-8	2	Yes	"word" , 1-2 n-gram	'C': 10, 'loss': 'hinge', 'tol': 0.1

The result shows in Table 6 the best configuration is the **SVM-1**, using the first preprocess for the data, without sentimental analysis, "char_wb" as the analyzer and (4-5) as ngram_range. This model was used for *Run0* in Task 2.

Table 6
Results of the different configurations of the SVM classifiers on development partition. In bold, the best result for each metric.

	Precision	Recall	F1-score
SVM-1	**0.92**	**0.89**	**0.91**
SVM-2	0.86	0.84	0.85
SVM-3	0.91	0.85	0.88
SVM-4	0.84	0.83	0.83
SVM-5	0.91	0.83	0.87
SVM-6	0.86	0.81	0.83
SVM-7	0.89	0.82	0.83
SVM-8	0.84	0.80	0.82

We tested adding sentimental analysis as a feature because it has been shown to be effective in improving performance in similar tasks using SVM. In particular, we achieved significant improvements in MentalRiskES 2024 [7], a shared task for the early detection of depression symptoms.

3.3. BERT-like Model Approach

It is well known that state-of-the-art models in NLP are based on Transformers. Models like BERT and RoBERTa typically offer excellent versatility for classification tasks. However, these models are often limited to handling a maximum of 512 tokens, which can be problematic for tasks requiring the processing of long contexts, such as the one at hand. To address this issue, we used one of these models as a baseline to compare against other models with a better capacity for managing large contexts. This comparison allows us to evaluate the performance trade-offs and benefits of different approaches in handling extended textual data.

We conducted research to find a base model trained in domains related to eating disorders; however, we did not find any pre-trained model specialized in eating disorders. While we were doing the research, we found the following: between 50% to 75% of those who struggle with an eating disorder will also experience symptoms of depression or anxiety [11]. Therefore, we used a pre-trained model related to mental disorders instead.

Research by Alireza Pourkeyvan [12] indicates that the state-of-the-art model in mental disorder detection is MentalRoBERTa [13]. MentalRoBERTa is a variant of the RoBERTa model that is specialized for mental health applications. It is pre-trained on a specialized corpus that includes texts from mental health forums, clinical notes, and general language corpus. This pre-training enables MentalRoBERTa to better understand and process language related to mental health, enhancing its applicability and effectiveness in this domain.

The model selected was *AIMH/mental-roberta-large* [14], a RoBERTa variant trained specifically on mental health-related posts from Reddit. This model is available on the HuggingFace [15] public hub (https://huggingface.co/AIMH/mental-roberta-large) and provides specialized capabilities for understanding mental health discourse.

We obtained two models by fine-tuning the base pre-trained model with two datasets: one using Dataset 1 (RoBERTa-1) and the other using Dataset 2 (RoBERTa-2), with the second incorporating data augmentation. Table 7 shows the configuration used in the fine-tuning process.

Table 7
Parameters for the fine-tuning process.

parameter	value
optimizer	AdamW
learning rate	7e-5
lr scheduler type	linear
weight decay	0.01
number of epochs	10
training batch size	16

Table 8 displays the results of each model on the development partition. The results indicate that **RoBERTa-2** obtained the best performance, a fine-tuned model with data augmentation. Consequently, we used this model for *Run1* in Task 2 of our participation.

Table 8
RoBERTa's result for Task 2 on development partition.

	Data Augmentation	Precision	Recall	F1-score
RoBERTa-1	No	0.88	0.85	0.86
RoBERTa-2	Yes	**0.92**	**0.90**	**0.91**

3.4. LongFormer Approach

As previously mentioned, one of the major drawbacks of BERT-like or RoBERTa-like models based on Transformers is their limited capacity to handle large contexts. However, there is a variant of Transformers called LongFormer, which can process longer texts effectively [6]

LongFormer, which stands for "Long-Document Transformer," is designed to process long contexts more efficiently than traditional Transformer models such as BERT or RoBERTa. The LongFormer architecture exhibits the following characteristics:

- **New attention mechanism**: An efficient attention mechanism that uses a sliding window, where each token only attends to a fixed number of neighborhood tokens, reducing the complexity.
- **Global attention selection**: The architecture can select which tokens are globally attended and which are just attended locally.

The pre-trained model chosen was *AIMH/mental-longformer-base-4096* [16] a pre-trained LongFormer for the mental health domain. This model can be found in https://huggingface.co/AIMH/mental-longformer-base-4096.

As in with the RoBERTa model, we fine-tuned the LongFormer with the two datasets: Dataset 1 without data augmentation (LongFormer-1), and Dataset 2 with data augmentation (LongFormer-2). We used the same fine-tuning parameters as in RoBERTa's experimentation; the configuration is in Table 7.

Table 9 shows the results of the experimentation, where *LongFormer-2* (fine-tuned with data augmentation) achieves better performance than *LongFormer-1* (fine-tuned without data augmentation). We used the two models in our participation, as *Run2* and *Run3*

Table 9
LongFormer's results for Task 2 on development partition.

	Data Augmentation	Precision	Recall	F1-score
LongFormer-1	No	0.91	0.89	0.89
LongFormer-2	Yes	**0.96**	**0.92**	**0.94**

4. Runs

Table 10 summarizes the selected model for each run, also the development performance is shown.

Table 10
Summary of the approaches chosen for each run. Also, the performance achieved by each system in the development partition.

	Task	Model	Precision	Recall	F1-score
Run0	1	SVM-1	0.92	0.89	0.91
Run1	1	RoBERTa-2	0.92	0.90	0.91
Run2	1	LongFormer-1	0.91	0.89	0.89
Run3	2	LongFormer-2	**0.96**	**0.92**	**0.94**

The rationale for selecting these models was to evaluate the significance of context in predicting anorexia. Each model varies in its capacity to handle input length, allowing for the processing of different context sizes. By comparing models with varying context-handling capabilities, we aim to determine how the extent of context affects the accuracy and effectiveness of mental illness prediction.

The results demonstrate that the SVM model, despite being less powerful in general, achieved performance comparable to MentalRoBERTa. This can be attributed to the SVM's ability to handle large texts, leveraging the full context provided by the input data. On the other hand, LongFormer models outperformed both BERT-like models and the SVM in this task. The performance of LongFormer can be credited to its capability to process larger contexts while maintaining the powerful features of Transformer-based models. This combination allows LongFormer to capture more comprehensive contextual information, leading to more accurate predictions in mental illness detection tasks.

4.1. Run Configuration

Besides, to select the model for each run, the classification systems contained additional parameters that needed to be set:

- For every round in the competition, we used as the input classifier a new sample created combining the new message of the user with the previous ones.

- Each system has an initial context, in other words, we made our systems wait until the initial context was sufficiently large. This context was different in each system:
  - **SVM**: An initial context of 50 tokens after the pre-process.
  - **RoBERTa and LongFormer**: An initial context of 100 tokens.
- The RoBERTa and LongFormer system has a limit of tokens, when the system was full we just returned the last prediction made.

5. Results

Table 11 shows the results achieved by our teams in Task 2. The structure of the Table 11 is the following: rows refer to each run and a special row refers to the highest values of the competition. The systems in the competition were ranked using the Macro-F1 score (last column). A total of 46 different systems (runs) participated in this task.

Table 11
Results for the 4 runs on Task 2. *Highest* refers to the highest values achieved in the competition. The values inside the parenthesis indicate our position in the ranking.

	Model	Precision	Recall	F1-score
Run0	SVM	**0.43** (15)	0.99	**0.60** (8)
Run1	RoBERTa	0.41	**1.00** (1)	0.58
Run2	LongFormer-1	0.32	0.99	0.49
Run3	LongFormer-2	**0.43** (15)	0.99	**0.60** (8)
Highest	-	0.73	1.00	0.790

Table 11 shows how the best systems are Run 0 and Run 3 if we take F1-score as the evaluation metric. Run 0 refers to **SVM-1**, a Support Vector Machine without sentimental analysis and a basic preprocess for the data. Run 3 refers to the **LongFormer-2**: pre-trained LongFormer fine-tuned with the data augmentation. These two runs achieved the eighth position in the global table at the competition.

However, our first thought was that LongFormer would perform better because of its power and capacity to handle large text, SVM has proven to achieve equal results thanks to its ability to deal with long texts. This indicates that classical approaches like SVMs continue to be useful in detecting mental illnesses because of their ability to handle large contexts. Therefore, SVMs still well-fitted in situations with low computational resources.

On the other hand, the results show how data augmentation has improved the performance of our models if we compare Run2 and Run3. Data augmentation helped our model learn more about positive samples and fit into the problem.

6. Conclusion

In this paper, we have presented the participation of the ELiRF-VRAIN team in Task 2 of eRisk at CLEF 2024: early detection of signs of anorexia. In addition to testing classic classification models and state-of-the-art Transformer models, we used LongFormers models to expand the context when making the decision. In addition, a proposal for data augmentation was presented with successful results during the training process.

For future work, two lines of improvement are identified. On the one hand, try to improve early detection so that the system does not need as much context to make the right decision; on the other hand, use Explainable Artificial Intelligence (XAI) techniques to understand the system's behavior better.

Acknowledgments

This work is partially supported by MCIN/AEI/10.13039/501100011033, by the "European Union" and “NextGenerationEU/MRR”, and by “ERDF A way of making Europe” under grants PDC2021-120846-C44 and PID2021-126061OB-C41. Partially supported by the Vicerrectorado de Investigación de la Universitat Politècnica de València PAID-01-23. It is also partially supported by the Spanish Ministerio de Universidades under the grant FPU21/05288 for university teacher training.

References

[1] FEACAB, Anorexia, 2015. URL: https://feacab.org/anorexia/, accessed: 2024-05-28.

[2] 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.

[3] 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, CEUR Workshop Proceedings, Grenoble, France, 2024.

[4] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017). URL: https://arxiv.org/abs/1706.03762, accessed: 2024-05-15.

[5] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, RoBERTa: A robustly optimized BERT pretraining approach, arXiv preprint arXiv:1907.11692 (2019). URL: https://arxiv.org/abs/1907.11692.

[6] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020). URL: https://arxiv.org/abs/2004.05150.

[7] A. M. Mármol-Romero, A. Moreno-Muñoz, F. M. P. del Arco, M. D. Molina-González, M.-T. Martín-Valdivia, L. A. Ureña-López, A. Montejo-Ráez, Overview of mentalriskes at iberlef 2024: Early detection of mental disorders risk in spanish, Procesamiento del Lenguaje Natural 73 (2024).

[8] D. E. Losada, F. Crestani, A test collection for research on depression and language use, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the 7th International Conference of the CLEF Association (CLEF 2016), 2016, pp. 28–39. URL: https://doi.org/10.1007/978-3-319-44564-9_3. doi:10.1007/978-3-319-44564-9_3.

[9] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, É. Duchesnay, Scikit-learn: Machine learning in python, Journal of Machine Learning Research 12 (2011) 2825–2830. URL: https://jmlr.org/papers/v12/pedregosa11a.html.

[10] L. X. Yuan, distilbert-base-multilingual-cased-sentiments-student (revision 2e33845), 2023. URL: https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student. doi:10.57967/hf/1422.

[11] N. I. of Mental Health, Eating disorders, n.d. URL: https://www.nimh.nih.gov/health/statistics/eating-disorders, accessed: 2024-05-30.

[12] A. Pourkeyvan, R. Safa, A. Sorourkhah, Harnessing the power of hugging face transformers for predicting mental health disorders in social networks, IEEE Access 12 (2024) 28025–28035. URL: http://dx.doi.org/10.1109/ACCESS.2024.3366653. doi:10.1109/access.2024.3366653.

[13] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, MentalBERT: Publicly available pretrained language models for mental healthcare, in: N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, S. Piperidis (Eds.), Proceedings of the Thirteenth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2022, pp. 7184–7190. URL: https://aclanthology.org/2022.lrec-1.778.

[14] AIMH, Mentalroberta: A robustly optimized bert pretraining approach for mental health, 2024. URL: https://huggingface.co/AIMH/mental-roberta-large, accessed: 2024-05-15.

[15] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (2020). URL: https://www.aclweb.org/anthology/2020.emnlp-demos.6.

[16] AIMH, Mentallongformer: A long-document transformer model for mental health, 2024. URL: https://huggingface.co/AIMH/mental-longformer-base-4096, accessed: 2024-05-15.