Measuring the Severity of the Signs of Eating Disorders Using Machine Learning Techniques

Notebook for the eRisk Lab at CLEF 2024

Sachin Prasanna[1,†], Abhayjit Singh Gulati[1,†], Subhojit Karmakar[1,†], M Yoga Hiranmayi[1,†] and Anand Kumar Madasamy[1,†]

[1]*National Institute of Technology Karnataka, Surathkal, Mangaluru, 575025, India*

Abstract

The paper presents the results submitted by Team SCaLAR-NITK for task 3 of eRisk Lab at CLEF 2024 [1]. The dataset provided by the task organizers consisted of 74 subjects for training and 18 for testing. We begin by describing the data cleaning and preprocessing steps. Subsequently, we outline various approaches used to address the problem, such as Word2Vec, TF-IDF, Backtranslation and Dimensionality Reduction, among others. Finally, we summarize the results obtained from each approach. Our solutions demonstrated strong performance, achieving the best results in 7 out of the 8 evaluated metrics.

Keywords

Machine Learning, Word2Vec, TF-IDF, Backtranslation

1. Introduction

Eating disorders, such as anorexia nervosa, bulimia nervosa, and binge eating disorder, are serious mental health conditions characterized by abnormal eating habits and distorted body image [2]. These disorders often stem from a combination of genetic, psychological, and social factors. Individuals may restrict food intake, engage in binge eating followed by purging behaviors, or compulsively overeat. Eating disorders can have devastating effects on physical health, leading to malnutrition, electrolyte imbalances, and organ damage. Psychological impacts include depression, anxiety, and low self-esteem. Treatment typically involves a combination of therapy, nutritional counseling, and medical supervision to address both physical and mental aspects of the disorder.

In this paper, we discuss methodologies for detecting signs of eating disorders from users' Reddit posts. The first approach involves creating separate models for each of the 22 questions, with each model learning the distribution specific to its corresponding question. The second approach reshapes the dataset to include the text of the question in a separate column, allowing the use of a single model for both training and predicting, rather than maintaining 22 individual models. The third approach employs Principal Component Analysis (PCA) for dimensionality reduction, ensuring that the importance of the text related to the question is weighted at a ratio of 3:1. These methodologies aim to effectively identify patterns and indicators of eating disorders in user's reddit posts.

2. Dataset

The dataset was given in the form of XML files which had to be cleaned for further processing. The XML files consisted of the user names, the posts they had posted and their timestamps. The answers given by the subjects (true labels) were given as a separate text file.

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

[†] These authors contributed equally.

sachinprasanna.211it058@nitk.edu.in (S. Prasanna); abhayjitsinghgulati.211ee102@nitk.edu.in (A. S. Gulati); subhojit.211it071@nitk.edu.in (S. Karmakar); myogahiranmayi.211it038@nitk.edu.in (M. Y. Hiranmayi); m_anandkumar@nikt.edu.in (A. K. Madasamy)

0009-0001-0890-064X (S. Prasanna); 0009-0004-8987-7083 (A. S. Gulati); 0009-0000-1356-7266 (S. Karmakar); 0009-0006-7413-4900 (M. Y. Hiranmayi); 0000-0003-0310-4510 (A. K. Madasamy)

2.1. Data Cleaning

Several XML had loading issues and were incorrectly formatted. So as a first step, these issues were resolved and the XML file for each user was converted to CSV files for further cleaning and usage. It was also found that column names had a mismatch and was fixed.

Further cleaning involved the removal of emojis from the posts of users using the *emoji* library. It was also found that several posts were enclosed with starting with **b"** and ending with **"**. These were removed accordingly. Unicode representations were replaced with their actual representations. For example, *\xe2\x80\x99* means **'** and instances such as these were found and replaced.

2.2. Data Preprocessing

After cleaning, all posts from a single user were concatenated into one single chunk and preprocessed using standard preprocessing methods. This text was lowercased and URLs were removed. All kinds of punctuation were removed. Stop words were removed so that the machine learning models are not heavily influenced by the effects of these words. Finally, the words were lemmatized using the *WordNetLemmatizer* function from the *nltk.stem* library.

3. Methodology

Three different approaches are proposed in our solution to the problem. They are described and explained below.

3.1. Different Models for Each Question

This approach consisted of fitting a model to each question. Since there were 22 questions answered by each subject, we made 22 different models to learn the distribution of each question's answers. Thereafter, the **Linear Support Vector Machine** algorithm was used for classification [3].

The **Linear Support Vector Machine** model implemented here utilizes stochastic gradient descent with a hinge loss function for maximum-margin classification. It applies L2 regularization to prevent overfitting and employs hyperparameters for regularization strength, reproducibility, the maximum number of iterations and the tolerance of stopping criteria. It is particularly suited for large datasets and online learning scenarios.

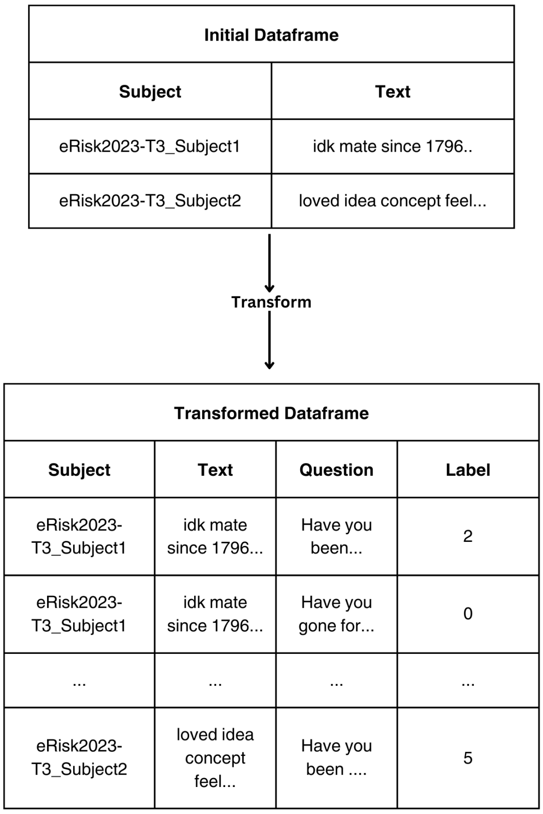
For word embeddings for input to the models, a pipeline was constructed. The first component was the *CountVectorizer*, which converts the collection of text documents into a matrix of token counts. Then, the standard *TF-IDF* approach is taken and the corresponding matrix is constructed. These embeddings are taken for each document and fed into the machine learning algorithms and neural network for each question.

This approach demonstrated good results, and got the best score in 2 of the measuring metrics. The reason as to why we went with the simpler approach is because complex models would overfit the data, hence resulting in poorer scores, as observed in the results of the same task last year. Clearly, the approach paid off with good results.

Although a good approach with decent results, the importance of the questions was not given as inputs to the models. It was just a case of learning 22 arbitrary distributions and then predicting the same. This emerged as a drawback as the questions were not given any importance in the predictions. Also, since the training set was small, there was a chance that the training was not sufficient for predicting the validation labels correctly.

3.2. Extending Dataset and using Questions along with Word2Vec and Backtranslation

To overcome the drawback of the previous approach, the dataset was reshaped to include the text of the question in a separate column. This also includes the size of the dataset and makes use of a single model for training and predicting, rather than 22 models as in the previous approach. A pictorial representation of the transformation is shown in Figure 1.



Initial Dataframe	
Subject	**Text**
eRisk2023-T3_Subject1	idk mate since 1796..
eRisk2023-T3_Subject2	loved idea concept feel...

Transformed Dataframe			
Subject	**Text**	**Question**	**Label**
eRisk2023-T3_Subject1	idk mate since 1796...	Have you been...	2
eRisk2023-T3_Subject1	idk mate since 1796...	Have you gone for...	0
...	...	...	...
eRisk2023-T3_Subject2	loved idea concept feel...	Have you been	5

Figure 1: Transformation of the Dataframe to include questions

After the transformation, word embeddings for both text and question was constructed using Word2Vec. Pretrained Word2Vec models, offer precomputed word embeddings trained on vast amounts of text data, such as Google News articles as used in our work. The loaded model, pre-trained on Google News corpus (3 billion running words) word vector model (3 million 300-dimension English word vectors), captures intricate semantic nuances and gives numerical meaning to text [4]. Since each word is represented by a 300 size vector, the average vector of all the words was taken as the final representation of each subject. If a word was not present in the Word2Vec corpus, then its vector was taken as 0.

Separate word embeddings are generated for the text and question. Each of these are 300 dimensional vectors, and were concatenated to form the final embedding vector which was used for training which turned out to be a 600 dimensional vector. By this way, we ensured that both the text and question are given equal weightage during prediction of labels. An illustration of our methodology is shown in Figure 2.

The imbalance in training data was taken care of by Backtranslation. Backtranslation is the process of translating text from one language to another and then translating it back to the original language [5]. We translated our text from English to French and vice versa. This technique is commonly used in natural language processing for data augmentation and improving the robustness of machine learning models. An illustration of Backtranslation is shown in Figure 3. The split of labels is represented pictorially in Figure 4.

Our implementation handles cases where the input text may exceed character limits specified by GoogleTranslate API (4999 characters) by splitting it into smaller chunks. It utilizes the Google Translator

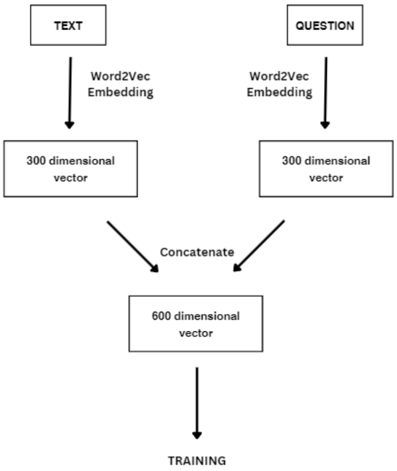
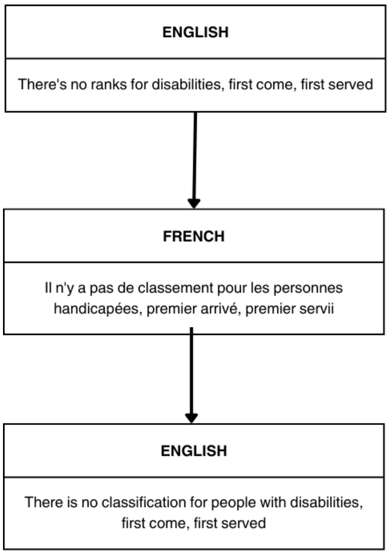


Figure 2: Pictorial representation of Word2Vec with Backtranslation

library to perform translation tasks efficiently.

Backtranslation was performed for only the text column and only for those instances with labels 2,3,4 and 5. These were then added to the dataset as new instances, which increased the weightage of the minority classes. This can be pictorially seen in Figure 5.



There's no ranks for disabilities, first come, first served

Il n'y a pas de classement pour les personnes handicapées, premier arrivé, premier servii

There is no classification for people with disabilities, first come, first served

Figure 3: Backtranslation using the French Language

Using these new embeddings - **Linear Support Vector Machine** and **Gradient Boosting** algorithms were used to train the data.

The **Gradient Boosting** model utilized in this implementation is a powerful ensemble learning method widely used for classification tasks. It works by sequentially adding weak learners, typically decision trees, to correct the errors made by preceding models. Each subsequent learner focuses on the residual errors of the previous model, gradually improving the overall predictive performance [6]. The key hyperparameters include the number of estimators and the learning rate, which control the model's complexity and the rate at which each additional learner contributes to the ensemble.

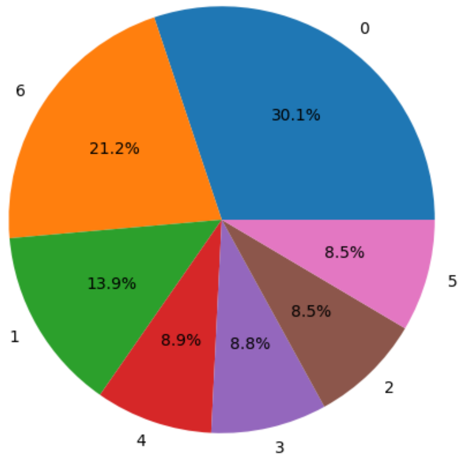


Figure 4: Label distributions of the Training Set

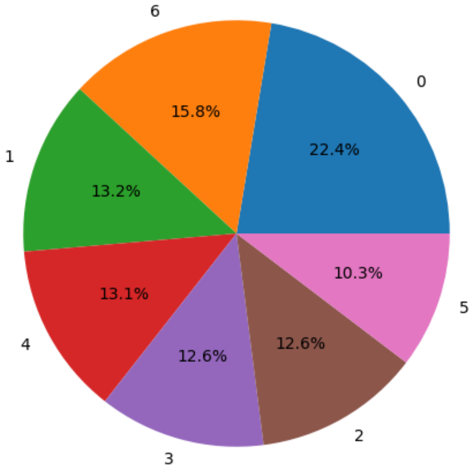


Figure 5: Label distributions of the Training Set after Backtranslation

3.3. Using Word2Vec with Backtranslation and Dimensionality Reduction

An innovative method to reduce the weightage of the question embeddings is proposed using dimensionality reduction. Principle Component Analysis (PCA) was used for the purpose of dimensionality reduction.

PCA (Principal Component Analysis) is a dimensionality reduction technique that transforms high-dimensional data into a lower-dimensional representation while preserving most of its variance. It achieves this by identifying the principal components, which are orthogonal directions in the original feature space that capture the maximum variance in the data [7]. It was used to reduce to the dimensions of the question embeddings from size 300 to size 100.

This ensured that the importance of text to question is in the ratio of 3:1. The approach is illustrated in Figure 6.

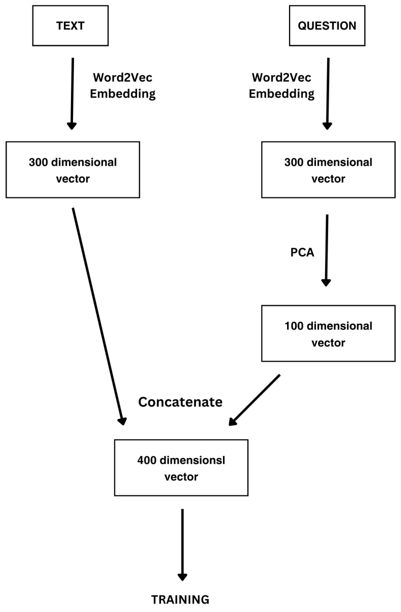


Figure 6: Pictorial representation of Word2Vec with Backtranslation and Dimensionality Reduction

4. Results

The results of our endeavours are shown in Table 1. For this task, 5 teams took part and 14 different solutions or runs were submitted, along with 3 baseline solutions described by the authors. Our solutions performed exceedingly well and we had the best results in 7 out of the 8 given metrics.

Citing previous year results, complex deep learning based solutions like transformers did not give great results - most likely because of the overfitting problem. Hence we decided to go with a simpler machine learning based approach to tackle the problem. It can be inferred that when the size of the dataset is small, simpler solutions offer better results than complex Deep Learning based solutions.

Table 1
Results of SCaLAR-NITK team for Task 3 in performance results.

Run	MAE	MZOE	MAE_{macro}	GED	RS	ECS	SCS	WCS
0	1.912	0.591	**1.643**	2.495	2.713	1.568	**1.536**	2.098
1	1.980	0.664	1.972	2.570	2.562	**1.553**	1.960	2.066
2	1.879	**0.568**	1.942	2.158	2.477	2.222	2.245	2.364
3	1.932	0.586	1.868	**2.117**	2.430	2.046	2.242	2.407
4	**1.874**	0.672	1.820	2.292	**2.140**	1.557	1.880	2.061

5. Conclusion

This paper outlines Team SCaLAR-NITK's involvement in Task 3 of the eRisk@CLEF 2024 edition, where we investigated diverse techniques for assessing eating disorders across multiple users based on their Reddit contributions. Our investigation focused on two primary approaches to tackle the problem. The first approach employed 22 distinct models, each tailored to a specific question, while the second approach utilized a single model capable of capturing the essence of both the questions and the posts. Innovative approaches such as Backtranslation was used to balance the label distribution and Principle Component Analysis (PCA) was used to better assign the weightages of text and question. Our contributions aim to address complex societal challenges in mental health detection and intervention.

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] T. Anwar, M. Fuller-Tyszkiewicz, H. K. Jarman, M. Abuhassan, A. Shatte, W. Team, S. Sukunesan, Edbase: Generating a lexicon base for eating disorders via social media, IEEE Journal of Biomedical and Health Informatics (2022).

[3] S. Ghosh, A. Dasgupta, A. Swetapadma, A study on support vector machine based linear and non-linear pattern classification, in: 2019 International Conference on Intelligent Sustainable Systems (ICISS), 2019, pp. 24–28.

[4] D. Wallace, T. Kecahdi, Outlier detection in health record free-text using deep learning, in: 2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2019.

[5] Q. Xu, Y. Hong, J. Chen, J. Yao, G. Zhou, Data augmentation via back-translation for aspect term extraction, in: 2023 International Joint Conference on Neural Networks (IJCNN), 2023.

[6] D. Agrawal, S. Minocha, A. K. Goel, Gradient boosting based classification of ion channels, in: 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021.

[7] S. Sehgal, H. Singh, M. Agarwal, V. Bhasker, Shantanu, Data analysis using principal component analysis, in: 2014 International Conference on Medical Imaging, m-Health and Emerging Communication Systems (MedCom), 2014.