Sunday Rockers at eRisk 2022: Early Detection of Depression

Raluca-Andreea Gînga[1], Andrei-Alexandru Manea[1] and Bogdan-Mihai Dobre[1]

[1]*Faculty of Mathematics and Computer Science, University of Bucharest*

Abstract

In this paper, we describe the participation of Sunday Rockers team at eRisk 2022 on "Early Detection of Depression" task. The main goal of this task is to sequentially process pieces of evidence and detect as quickly as possible early traces of depression. Our approaches for this task varied from using TF-IDF and linguistic statistical features extracted from Reddit writings to using MixUP technique, a new approach for sentence classification to augment the data through interpolation. We obtained our best results with a Voting Classifier with hard voting applied on Text features and with an SVM applied on both textual features and numerical features as well.

Keywords

CLEF eRisk, Sunday Rocker2, BERT, Depression, MixUP, Voting Classifier, Support Vector Machines, XGBoost

1. Introduction

The last three decades have brought a huge development of information and communication technology. The advent and proliferation of the Internet have made it possible to create complex global networks of communication and collaboration. These new technologies have transformed the way we learn, communicate and work - they have fundamentally transformed the way we live. This evolution has brought with it various opportunities and benefits from an economic and social point of view, but also the emergence of new challenges.

One of these challenges is that even though the Internet was not designed with children in mind, many of the users of this environment are children and young people with specific needs and vulnerabilities that need to be recognized. Given the opportunities that the Internet offers in accessing knowledge, the benefits that it can have in developing the skills needed for the 21st century, but also the wide range of risks and dangers to which children may be exposed, it is necessary that decisions taken for the younger generation are based on real needs and current data.

Generally speaking, social networks gather individuals who have common goals. They were developed rapidly in the last 20 years, becoming the main way to create and maintain virtual relationships. Social networks may be classified by user experience: sharing photos (Pinterest), movies (YouTube, Netflix), thoughts (Facebook, Twitter, Reddit), and others for

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 5-8, 2022, Bologna, Italy

gingaraluca@gmail.com (R. Gînga); andrei.manea.dev@gmail.com (A. Manea); dobrebogdan98@gmail.com (B. Dobre)

CEUR Workshop Proceedings (CEUR-WS.org)

career management (LinkedIn).

Although we have access to a lot of information that can be interesting and useful, social networks might also be used to detect certain mental illnesses. A fairly common mental health condition is depression.

A study by the World Health Organization Evans-Lacko et al. [1] found that depression affects about 121 million people worldwide. It is estimated that annually, 3-15% of the general population has a depressive episode, of which 0.4 - 5% are severe depressive episodes. In Europe, 58 out of 1,000 adults have a major depressive disorder (33.4 million people). All these worrying numbers have made depression one of the most studied mental disorders at the moment.

Early detection of depression can help mitigate these threats, but most studies on early detection are based on patient diagnoses based on questionnaires or reported experiences Halfin [2] that are financially expensive and time-consuming, and a major part of some countries with a public health system do not have the necessary support for these diagnoses. Fortunately, social media can come to our aid with a solution to this problem, as many studies have been able to easily extract the data and content of social media posts to analyze and predict the mental well-being of a user (Choudhury et al. [3], Moreno et al. [4], Sadeque et al. [5], Schwartz et al. [6]).

In this paper, we present the 5 models for "Early Detection of Depression" task that we sent to the eRisk 2022 edition [7]. The main goal of this task is to sequentially process pieces of evidence and detect as quickly as possible early traces of depression. Our approaches for this task varied from using TF-IDF and linguistic statistical features extracted from Reddit writings to using MixUP technique, a new approach for sentence classification to augment the data through interpolation.

The rest of the paper is organized as follows. In section 2, the related work is introduced. In section 3 the eRisk 2022 task is described, along with the origin of the datasets and the problem to be solved. Then, section 4 describes the experiments and methods approached, along with the results obtained in the 5 runs. The paper ends with some conclusions about the best approach and other methods that could be tried in the future for this task.

2. Related Work

There has been significant research in the area of early detection of depression in the scientific community. It was demonstrated that early detection and treatment of depression can alleviate some of the emotional symptoms of depression and lower the costs of care Halfin [2].

The idea that certain patterns in text might be an indication of depression was enforced by many studies. Stirman and Pennebaker Stirman and Pennebaker [8] discovered a clear predisposition on the side of depressed poets to use self centered words such as "I, me, my". Zinken et al. [9] showed that depressed individual who were unable to successfully participate in guided self help tend to use less complex syntax and fewer causation words than those who were able to successfully complete the program. That brings us to the conclusion that linguistic markers may not only be useful for detection of depression but also for assessing the severity of the disease.

Shen et all Shen et al. [10] proposed an approach to detect depression in the case of Twitter users and, more recently, Jooshi and Kanoongo Joshi and Kanoongo [11] researched methods for depression detection on Twitter, Facebook and Live Journal.

There was some research that focused on detecting certain mental illnesses in the case of Reddit users. The study of Shen and Rudzics Shen and Rudzicz [12] dealt with detecting anxiety at Reddit users and obtained an accuracy of 0.98 and a precision of 0.99 using Neural Networks with lexicon based features and N-grams.

There has been some research on depression detection and related domains on Reddit. De Choudhury et al. [13] found some linguistic markers that are associated with users who engage in suicidal ideation such as heightened self-attentional focus and poor linguistic coherence.

Tadesse et al. [14] achieved 0.91 accuracy and 0.93 F1 score for the dataset of Reddit users posts from CLEF eRISK 2018 (Conference and Labs of Evaluation Forum for Early Risk Prediction) using a Multilayer Perceptron and LIWC, LDA and bigrams as features. The challenge consisted of sequentially processing the posts of reddit users and flagging potentially depressed users as soon as possible.

In this paper we will approach a similar task: Early Detection of Depression from CLEF eRisk 2022 workshop, containing eRisk 2017 and 2018 datasets as training datasets.

3. Task Description

The Early Detection of Depression task is part of the CLEF eRisk 2022 workshop[1]. This task consists of performing the early detection of the risk of depression by analyzing a series of comments and content published by certain users on the Reddit platform[2]. In this regard, a system receives user-generated content as input and must return a prediction about the likelihood that the user will be prone to depression.

This challenge consists of sequentially processing evidence and detecting any signs of depression as soon as possible. This task refers to the evaluation of Text Mining solutions and thus focuses on texts written on the Reddit network. Texts must be processed in the order in which they were created.

The dataset for this task contains the datasets from previous editions of eRisk from 2017 and 2018. The data collection for this task is described in detail in Losada and Crestani [15]. The dataset consists in a collection of writings (either in the form of posts or comments) from a set of users extracted from Reddit platform. There are two categories of users: depressive (at risk) and non-depressive (control group) and, for each such user, his collection of writings. The task was organized in two different stages:

- Training phase, in which participants had access to training data (those of 2017 and 2018) with the entire writing history and associated labels (positive or negative)
- Test phase, in which participants had to iteratively extract the writings of a number of 2001 users and send the predictions to the eRisk server

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

[2] https://reddit.com.

The evaluation was not only based on the correctness of the system result (i.e. whether or not the user was diagnosed with depression), but also on the delay with which this decision was issued and other metrics described in Parapar et al. [16].

The following table 1 contains a general overview of the training set, where "Avg. time span" is the average number of days between the first and the last submission and a submission stands for any post or comment. It can be seen the high imbalance between the positive subjects and the negative subjects, so the first problem of this task was the classes imbalance.

Table 1
General observation of the training dataset

Year	Label	Num. subjects	Num. submissions	Avg. num. of. submissions/user	Avg. time span
2017	negative	752	481837	641	625
2017	positive	135	49557	367	587
2018	negative	741	504523	681	703
2018	positive	79	40665	515	787

4. Experiments

We requested the maximum of 5 submissions, presented in the following subsections.

Three of the 5 submitted models (Voting Text, XGB Metadata and SVM Combined) were based on 3 variations of the dataset:

1. Dataset based only on text (i.e. on writings written by different users) - called "Voting Text", because we used Voting Ensemble only on the textual features
2. Dataset based on numeric features (detailed in the next subsection) - called "XGB Metadata", since we used an XGBoost only on the numerical features
3. Dataset based on both text and numeric features (ie a combination of the two) - called "SVM Combined", since we used an SVM on this dataset variation (combination of both textual features and numerical features extracted from text as well)

The other two models were BERTs: one based on Mental-Bert, and the other one based on an augmentation technique for text classification, MixUp.

In order to see the generalization of our models and have some preliminary results regarding the efficiency of the algorithms developed, we proceeded into a train-test split of the dataset (80% reserved for training, 20% for validation). Both training and validation datasets were saved for further experimenting and ensuring all of the experiments have the same source of data.

4.1. Text Preprocessing

The datasets formed were based on chunks of 100 writings in order to have a smaller dataset and capture even the writings of some users that are positive (depressed), having writings that don't necessary belong to "depression" category.

The preprocessing steps consisted in:

- removing the links from the writings
- fixing the contractions (i.e. "I'll" turns into "I will")
- removing the "[removed]" comments/posts from Reddit
- replacing emoticons and emojis with plain text using FlashText[3] and Emot[4] modules
- adding to the existing user writing the emotional affect using NRCLex library[5]
- removing the numbers from the posts
- ignoring the mentions (using "@" sign) and keeping the words through hashtag
- removing the words with less than 3 characters
- removing the additional whitespaces

4.2. Features

We implemented and combined two types of features: 1) TF-IDF-based, and 2) textual-based features (called metadata).

4.2.1. TF-IDF features

A TF-IDF vectorizer was trained on the texts. This vectorizer was then used to obtain TF-IDF features with a maximum of 5000 features. These features were passed then on to the classifier. We experimented with different n-grams: unigrams, bigrams and trigrams, but we obtained better performance using only unigrams. We have also tried different variations of TF-IDF analyzers: word and character-based, obtaining better scores on word analyzer. Thus, we continued with unigrams & word analyzers as TF-IDF arguments.

4.2.2. Textual-based features (Metadata)

For this kind of features, we used different types of text-based features. All of them were normalized by the text length.

- Language-style features (here we included the total number of words, nouns, adjectives, verbs, adverbs, negations in users' writings, number of question marks, exclamation marks, capitalized words and other punctuation marks.
  Another interesting features belonging to this category are Formality scores, Trager coefficients, readiness to action & aggressiveness coefficients, activity index.

[3] https://flashtext.readthedocs.io/
[4] https://github.com/NeelShah18/emot
[5] https://github.com/metalcorebear/NRCLex

Formality score Heylighen et al. [17] measures the degree of formality of a text and the quality of a text in terms of the abundance of a vocabulary. F-measure score is given by the following formula:

$$f\_measure = \frac{(nounFreq + adjFreq + prepFreq + artFreq)}{2} - \frac{(pronFreq + verbFreq + advFreq + intjFreq) + 100}{2} \tag{1}$$

Trager coefficient, as pronominalisation, measures the stability/instability, dynamism of the text. It is given by the number of verbs divided by the number of nouns.
Readiness to Action indicates the certainty of some sort of action and the degree of socialization (or expressing ideas to an audience).
Aggressiveness coefficient is given by the excessive use of verbs or verbs forms in comparison with the rest of words.
Activity index is given by the number of verbs / (number of verb + adjective + adverbs).

- User behaviour. Here we included *time posting level* as a user-related feature, since users with sleep disorders tend to post late at night. That is why because lack of sleep and insomnia occur in 60%-80% of depressed patients Luca et al. [18].

- Self-preoccupation. There are some studies suggesting that people who are using more first-person pronouns on average are more depressed than people who use the third person (Nerbonne [19], Edwards and Holtzman [20]). In this regard, we decided to include another feature that is counting the occurrence of first-number pronouns. Another feature we included was the count of over-generalization, since depressed users tend to over-generalize and include intense quantifiers, like *all, everybody, nobody, everyone, etc.*. For example, instead of criticizing, judging a specific category of people, a person may write "All men/women are bad".

- Reminiscence. Mowery et al. [21] showed that depressive users tend to make reference to past more frequently than control users. We defined a feature to capture the reference to past called *temporal_past*, where we counted the number of past tense verbs related to time that were used by the users in their writings.

- Drugs, feelings (symptoms) and relevant depression lexicon. There is also evidence linking depression to non-suicidal self-harm Greaves [22], so tracking this information would prove beneficial for our task. In this regard, we decided to build some dictionaries related to depression-domain: antidepressants, psychoactive drugs, unpleasant feelings and NSSI-words Greaves [22].
  - The unpleasant feelings, antidepressant drugs, psychoactive drugs were obtained from Choudhury et al. [23] and Wikipedia.
  - The trigrams and 5-grams related to suicide and negative feelings were taken from Colombo et al. [24].

Feature	Type	Description
no_of_words	numeric	number of words
no_of_nouns	numeric	number of nouns
no_of_adjectives	numeric	number of adjectives
no_of_verbs	numeric	total number of verbs
no_of_adverbs	numeric	number of adverbs
no_of_negations	numeric	number of negations
punctuation_count	numeric	punctuation count (".:;-")
questions_count	numeric	number of question marks
exclamations_count	numeric	number of exclamation marks
capitalized_count	numeric	number of words written in uppercase
formality_score	numeric	formality score as stated in Heylighen et al. [17]
trager_coefficient	numeric	Trager coefficient as stated in Vasyliuk [25]
readiness_to_action_coefficient	numeric	directness coefficient as stated in Vasyliuk [25]
aggressiveness_coefficient	numeric	aggressiveness coefficient as stated in Vasyliuk [25]
activity_index	numeric	activity index as stated in Havigerová et al. [26]
time_level	numeric	number of writings posted between 12 AM and 7 AM (in deep night)
first_person_count	numeric	number of 1^{st} person pronouns
overgeneralization_count	numeric	number of words than tend to overgeneralize (everybody, all, nothing, nobody etc.)
temporal_past	numeric	number of words related to past (yesterday, before, last etc.)
antidepress_count	numeric	number of words related to antidepressants
three_grams_suicide_count	numeric	number of trigrams suicide related terms
five_grams_suicide_count	numeric	number of 5-grams suicide related terms
psychoactive_count	numeric	number of words psychoactive-related drugs
unpleasant_feeling_count	numeric	number of words that reveals an unpleasant feeling
nssi_count	numeric	number of NSSI words

Table 2
Characteristics and Features extracted from the texts written by users

- NSSI-words. We used a word dictionary of terms related to non-suicidal self injury (NSSI words) from Greaves [22]. This dictionary is divided in several categories: 1) Methods of NSSI; 2) NSSI terms; 3) Instruments used; 4) Reasons for NSSI. We decided to build a feature where we tracked the number of NSSI-related words.

Table 2 is showing the list of features described above.

The correlation between the features is shown in the figure 1. As we can see, formality metric is in an inverse correlation with the label, so we can assume that the users who are prone to depression write in a more informal way. Readiness to action seems to have a positive and moderate correlation with label, expressing the fact that the users tend to socialize and write long texts on Reddit platform. These features were useful to the prediction in combination with the textual features as well.

4.3. Threshold BERT

We used a BERT model pretrained Ji et al. [27] on mental health datasets, including eRisk 2018 and achieved a similar performance with the first BERT. It was modified with a sequence of 2 layers of fully connected neurons, having GELU Hendrycks and Gimpel [28] as activation function and a dropout of 20%.

We decided to train only the last 5 layers of the BERT pretrained encoder layers, the BERT pooler and the final classifier. The optimizer was AdamW Loshchilov and Hutter [29], with the default parameters, including learning rate of 10^{-3}. It also included a scheduler for reducing it

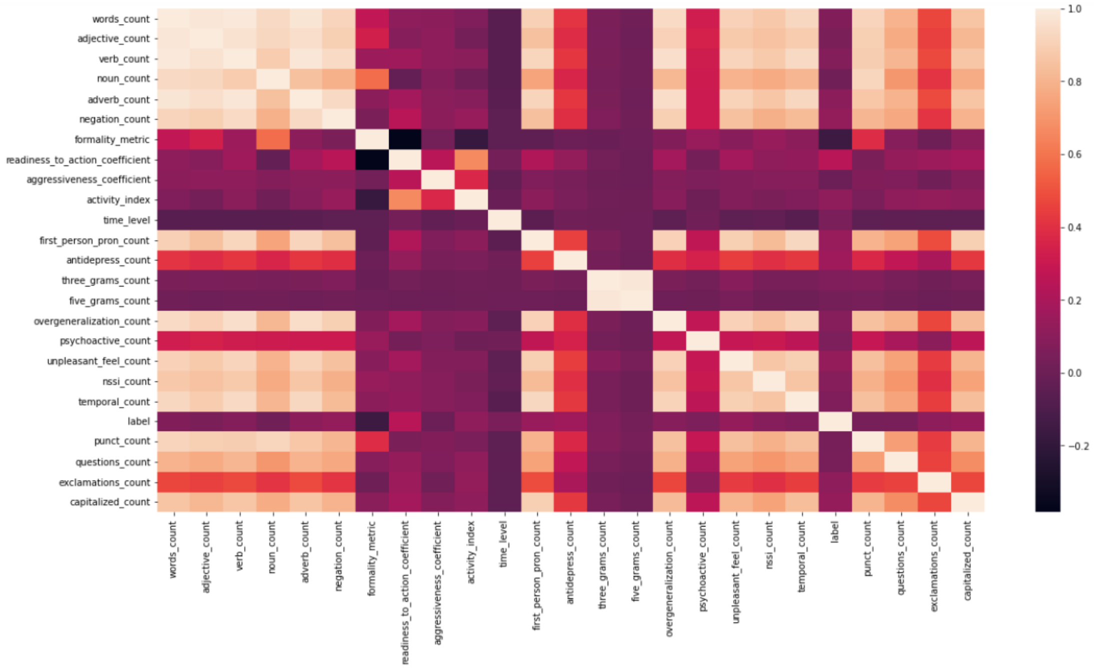

Figure 1: Correlation between the features and label

by dividing by 10, if the validation loss is locked on plateau was not improving in 3 epochs. The whole training procedure was written in Pytorch framework.

The first attempt was to create a weighted binary cross entropy loss, in a Pytorch class. The weights were manually computed as $[w_-, w_+]$ = $[\frac{N_-}{N}, \frac{N_+}{N}]$, where N_- and N_+ represents the number of negative and positive samples from the training set, and $N = N_- + N_+$. The model became biased after 2-3 epochs of training, by predicting only the majority class (negative).

So, the second attempt was to train the model with a regression loss, the mean squared error. Also, the learning rate from the optimizer was changed to 2×10^{-5} and $\varepsilon = 10^{-8}$.

The output of this model, can be considered a class probability. In order to provide a classification result, we may choose a threshold θ for a proper class separation. The decision function will be:

$$d_\theta(x) = \begin{cases} 0 & x \leq \theta \\ 1 & x > \theta \end{cases}$$

where we consider x, the output probability of the model. So, having a fixed value of θ, we can collect all the probabilities, x_i for each chunk from the validation set, and obtain the result $\hat{y}_i = d_\theta(x_i)$. Then, we can compute the F1 score for both class 0 and 1. F1 for a class c represents the harmonic mean of the precision and recall computed solely on the class c. For $\theta = 0.5$, the validation F1 score for class 1 was 19.23%.

So we decided to plot the histogram of the probabilities (x_i), displayed in figure 2. In order to find the best separation, we took 100 equidistant values of $\theta \in [0, 1]$, and chose the one which

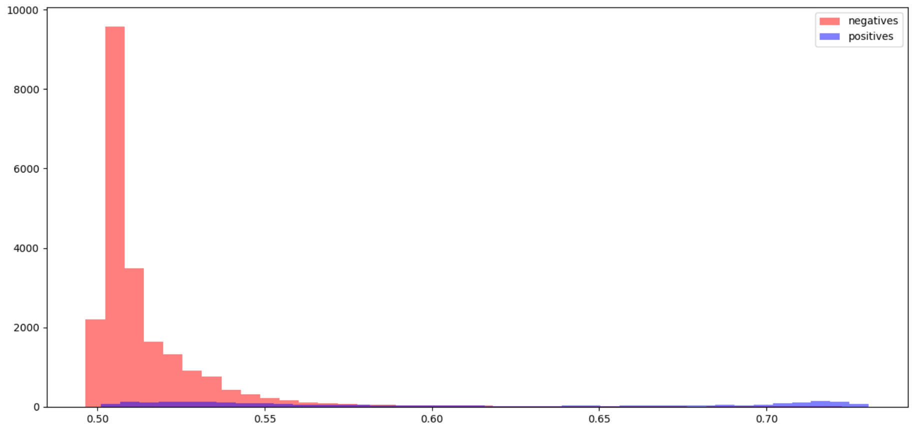

Figure 2: Histogram of the probabilities (x_i) obtained from the best BERT model on the validation set. The most of the negative probabilities have a distribution similar to exponential density function, and most of them lies in the interval of $[0.5, 0.55]$, whereas the outputs from the negative class are are double-normal distributed in the region of $[0.5, 0.55]$ and $[0.7, 0.75]$

maximizes the F1 score on positive class. The evolution can be seen in figure 3 and the best parameter found $\theta = 0.556$, obtaining f1_score_0 = 95%, f1_score_1 = 60%. By choosing this value, the model will also ignore the negative cases from the interval [0.5, 0.55], resulting in more confusions, but a better class separation according to the model understanding of the text.

Another attempts has been made by freezing all BERT layers, but they didn't outcome the first separation result.

4.4. Voting Text

In the second run, we decided to use a Voting Classifier on the textual features. Voting Classifier is that type of Machine Learning estimator that is training various base models or estimators and is predicting on the basis of aggregating the findings of each base estimator.

For this Voting Classifier, we decided to use as estimators: Logistic Regression, Light Gradient Boosting Machines and Support Vector Machines with linear kernel, since we've come to the conclusion that these algorithms provided very good results on the datasets on their own (without using a Voting system).

To overcome the imbalance present between the classes (depressed - control groups), we added a class weight to each algorithm in order to balance the datasets and provide better results.

In our case, the Voting Classifier consisted in aggregating the predicted class on basis of hard voting (voting is calculated on the predicted output class).

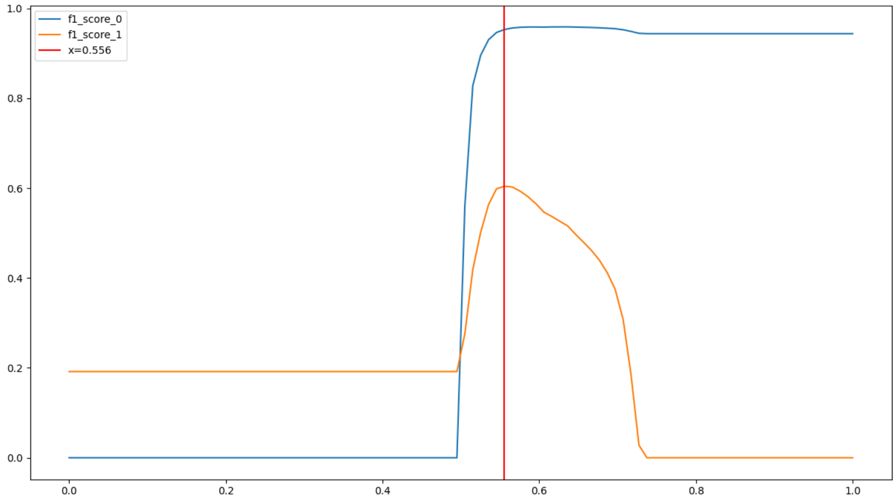

Figure 3: F1 score evolution on both 0 and 1 class evolution, according to the threshold θ. The significant interval is around [0.5, 0.75] where both scores have are ascending and then descending. The red vertical line representing the best option, which is also the local maxima point for both of the functions.

4.5. XGB Metadata

In the third run, we used a XGBoost model on the metadata features presented above. It was at first trained on the metadata features extracted from the whole training data. Then, for each new set of user posts received, the complete text for each user from their posts so far was created and the metadata extracted from these texts was used for prediction. Multiple models were evaluated on the metadata extracted from the train data and XGBoost had the best performance, hence it seemed to be the most promising for this approach. The model used the default parameters from the XGBoost Python library, with the exception of scale_pos_weight which had the value 10 in order to control the balance for positive and negative weights. This parameter is important when working with unbalanced datasets like the one in this task.

4.6. SVM Combined

In the fourth run, we used a Linear Support Vector Machines Classifier on the textual features + metadata features (numerical features with Standard Scaler transformer applied on them).

We have also tried other variations of Support Vector Machines algorithm, with different kernels (polynomial, RBF, sigmoid), but linear kernel worked best in our case. This algorithm provided as well the best results on the combined features (texts + numerical).

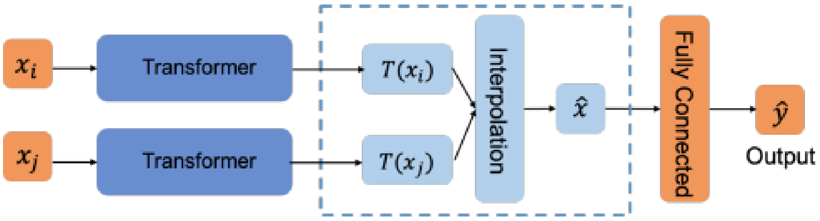

Figure 4: The overall framework of *mixup*-transformer, x_i, x_j are two separate sentences, fed to the same Transformer T. $T(x_i)$ is the representation of the input x_i generated by T, $\hat{x}$ and $\hat{y}$ are the interpolated representation and label, respectively.

4.7. MixUp BERT

The fifth model used was BERT using MixUp technique.

The MixUp approach, proposed by Zhang et al. [30] in 2017, is a more innovative one and less explored, especially in image classification field, that brought the best results for image classification tasks.

Usually, we input a data point X into a neural network with parameters, obtaining a predicted class/label along with the true label. We also have a loss function that compares what is the output with the true label and then trying to make that loss smaller. Briefly, we want to adjust our parameters such as the next time, the output of X, to be a little closer to the reality. We are calling this technique *Empirical Risk Minimization (ERM),* because our data point X comes from some data distribution $\mathcal{D}$ like the space of all natural images, but what we actually have is a dataset of a finite amount of data that we can sample X and y from, so instead of minimizing our true risk, we minimize the empirical risk. What is the problem of ERM? The problem is that we can get overly confident about our data points and nothing else and thus, it will hurt the generalization.

The paper of Zhang et al. [30] proposes to train these classifiers on all the data points that are between some two random samples. In mixup technique, it is starting by taking two samples from the dataset and mixing them together, resulting in a data point that's in between the other two samples. This mixing is done in a linear fashion and what is really important is that it is applied both on the feature vectors and the target labels. Coefficient λ is sampled randomly, however, it's always going to be in $[0, 1]$ interval.

Unlike image which consists of pixels, sentence is composed of a sequence of words. Therefore, a sentence representation is often built to aggregate information from a sequence of words. The first step of text classification is to use the word embedding to convert each word into a vector representation. In our approach, instead of using the traditional encoding methods, we use transformer-based pre-trained language models to learn the representations for text data. The BERT model we've used was fine-tuned with the *mixup* data augmentation method. Formally, mixup-transformer constructs virtual hidden representations dynamically during the training process as seen in figure 4.

In this approach, we used Bert Tokenizer with *bert-base-uncased* to tokenize the original texts, writings and *BERT-base-uncased* as BERT model. We trained the model for 5 epochs, with a learning rate of $2 * 10^{-5}$, using as training dataset the mix-up augmentation of the original training dataset with $\lambda = 0.5$. At the end, we added as callback Early Stopping with a patience of 2.

4.8. Results

The validation results we obtained on our validation data for each model are presented in table 3.

Table 3
Results obtained on our evaluation dataset

Model	Validation F1 score
Threshold BERT	0.604
Voting Text	0.644
XGB Metadata	0.390
SVM Combined	**0.647**
MixUp BERT	0.583

The evaluation results we received from the organizers for the 5 runs are in table 4. The runs are indexed from 0 to 4.

Table 4
Results from decision-based evaluation

Team (Run)	P	R	F1	ERDE$_5$	ERDE$_{50}$	latency$_{TP}$	speed	latency$_{F1}$
Sunday-Rocker2 (0)	0.091	**1.000**	0.167	0.080	0.053	4.0	0.988	0.165
Sunday-Rocker2 (1)	0.355	0.786	0.489	0.068	0.041	27.0	0.899	0.439
Sunday-Rocker2 (2)	0.092	0.388	0.149	0.088	0.083	117.5	0.575	0.085
Sunday-Rocker2 (3)	0.283	0.816	0.420	0.071	0.045	37.5	0.859	0.361
Sunday-Rocker2 (4)	0.108	**1.000**	0.195	0.082	0.047	6.0	0.981	0.191
NLPGroup-IISERB (0)	0.682	0.745	**0.712**	0.055	0.032	9.0	0.969	**0.690**

The order in which we used the models for the runs is the following: Threshold BERT, Voting text, XGB Metadata, SVM Combined and MixUpBERT. In terms of F1 score, we can see that the second model, *Voting text*, achieves the best result of 0.489 while the 2 BERT models have similar performance. The best F1 score achieved by any team was 0.712. In both 2017 and 2018, for the same task, the best F1 score was 0.64.

Comparing to the other teams on the decision-based metrics, presented in table 4, we had high recall and low precision on the deep models, whereas the others were more balanced. The latency_weighted_f1 of the second model was the 5^{th} best in the ranking.

On the other evaluation metrics, presented in table 5, the 4^{th} model reached state-of-the-art performance on 1 writing for P@10 and NDCG@10. The last column tuple is full of zeros,

Table 5
Results from ranking-based evaluation

		1 writings			100 writings			500 writings			1000 writings		
Team	Run	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100	P@10	NDCG@10	NDCG@100
Sunday-Rocker2	0	0.40	0.47	0.39	0.40	0.44	0.29	0.50	0.46	0.24	0.00	0.00	0.00
Sunday-Rocker2	1	0.70	0.81	0.39	**0.90**	**0.93**	0.66	**0.90**	0.88	0.65	0.00	0.00	0.00
Sunday-Rocker2	2	0.10	0.07	0.23	0.00	0.00	0.11	0.30	0.31	0.17	0.00	0.00	0.00
Sunday-Rocker2	3	**0.80**	**0.88**	0.41	0.50	0.50	0.23	0.60	0.69	0.34	0.00	0.00	0.00
Sunday-Rocker2	4	0.30	0.28	0.31	0.30	0.37	0.25	0.40	0.30	0.18	0.00	0.00	0.00
NLPGroup-IISERB	0	0.00	0.00	0.02	**0.90**	**0.92**	0.30	**0.90**	0.92	0.33	0.00	0.00	0.00
BLUE	1	**0.80**	**0.88**	**0.54**	0.70	0.64	0.67	0.80	0.84	**0.74**	**0.80**	**0.86**	**0.72**

because we submitted only 682 runs. This was due to the inference for the 2 BERT models which was computational expensive. All 682 runs took 4 days and 4 hours.

5. Conclusions

In this paper, we presented the contributions of the Sunday Rockers team in the eRisk 2022 "Early Detection of Depression" shared task. We have used a variety of models and techniques, including Transformers, Voting Ensemble & other Machine Learning models using multiple linguistic features.

We described 5 solutions: 2 for chunk classification using deep models and 3 for cumulative texts using lighter models. Different from most previous methods, we handcrafted the statistical features and combined them with TF-IDFs.

Our maximum latency weighted, obtained with the second submission (Voting Text), is ranked on the 5^{th} place.

For future work, we aim to address the issue of severely imbalanced training data and small amount of positive samples by augmentation of texts inserting common sentences or even extracting them from various Depression subreddits from Reddit.

Moreover, there is a need to optimize the inference of high volume data, especially for deep models. This may be done by deploying the model inference on a special cloud with a high speed transfer. By overcoming this, we would also make use of previous hidden layers of BERT in order to monitor the evolution of the individual mindset.

Acknowledgments

We would like to thank our supervisor, Ana-Sabina Uban, for the guidance throughout the development of this project.

References

[1] S. Evans-Lacko, S. Aguilar-Gaxiola, A. Al-Hamzawi, J. Alonso, C. Benjet, R. Bruffaerts, W. Chiu, S. Florescu, G. de Girolamo, O. Gureje, J. M. Haro, Y. He, C. Hu, E. Karam,

N. Kawakami, S. Lee, C. Lund, V. Kovess, D. Levinson, G. Thornicroft, Socio-economic variations in the mental health treatment gap for people with anxiety, mood, and substance use disorders: Results from the who world mental health (wmh) surveys, Psychological Medicine 48 (2017) 1–12. doi:10.1017/S0033291717003336.

[2] A. Halfin, Depression: the benefits of early and appropriate treatment., The American journal of managed care 13 4 Suppl (2007) S92–7.

[3] M. D. Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: ICWSM, 2013.

[4] M. A. Moreno, L. Jelenchick, K. G. Egan, E. D. Cox, H. N. Young, K. Gannon, T. L. Becker, Feeling bad on facebook: depression disclosures by college students on a social networking site, Depression and Anxiety 28 (2011).

[5] F. Sadeque, T. Pedersen, T. Solorio, P. Shrestha, N. Rey-Villamizar, S. Bethard, Why do they leave: Modeling participation in online depression forums, in: SocialNLP@EMNLP, 2016.

[6] H. A. Schwartz, M. Sap, M. L. Kern, J. C. Eichstaedt, A. Kapelner, M. Agrawal, E. Blanco, L. Dziurzynski, G. J. Park, D. Stillwell, M. Kosinski, M. Seligman, L. H. Ungar, Predicting individual well-being through the language of social media, Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing 21 (2016) 516–27.

[7] A. Barrón-Cedeño, G. Martino, M. Esposti, F. Sebastiani, C. Macdonald, G. Pasi, A. Hanbury, M. Potthast, G. Faggioli, N. Ferro, Experimental ir meets multilinguality, multimodality, and interaction., 2022.

[8] S. W. Stirman, J. W. Pennebaker, Word use in the poetry of suicidal and nonsuicidal poets, Psychosomatic medicine 63 (2001) 517–522.

[9] J. Zinken, K. Zinken, J. C. Wilson, L. Butler, T. Skinner, Analysis of syntax and word use to predict successful participation in guided self-help for anxiety and depression, Psychiatry research 179 (2010) 181–186.

[10] G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, W. Zhu, Depression detection via harvesting social media: A multimodal dictionary learning solution., in: IJCAI, 2017, pp. 3838–3844.

[11] M. L. Joshi, N. Kanoongo, Depression detection using emotional artificial intelligence and machine learning: a closer review, Materials Today: Proceedings (2022).

[12] J. H. Shen, F. Rudzicz, Detecting anxiety through reddit, in: Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality, 2017, pp. 58–65.

[13] M. De Choudhury, E. Kiciman, M. Dredze, G. Coppersmith, M. Kumar, Discovering shifts to suicidal ideation from mental health content in social media, in: Proceedings of the 2016 CHI conference on human factors in computing systems, 2016, pp. 2098–2110.

[14] M. M. Tadesse, H. Lin, B. Xu, L. Yang, Detection of depression-related posts in reddit social media forum, IEEE Access 7 (2019) 44883–44893.

[15] D. E. Losada, F. A. Crestani, A test collection for research on depression and language use, in: CLEF, 2016.

[16] J. Parapar, P. Martín-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2021: Early risk prediction on the internet, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction: 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21–24, 2021, Proceedings, Springer-Verlag, Berlin, Heidel-

berg, 2021, p. 324–344. URL: https://doi.org/10.1007/978-3-030-85251-1_22. doi:10.1007/978-3-030-85251-1_22.

[17] F. Heylighen, J. Dewaele, L. Apostel, Formality of language: definition, measurement and behavioral determinants, 1999.

[18] A. Luca, M. Luca, C. Calandra, Sleep disorders and depression: brief review of the literature, case report, and nonpharmacologic interventions for depression, Clinical Interventions in Aging 8 (2013) 1033 – 1039.

[19] J. Nerbonne, The secret life of pronouns. what our words say about us, Lit. Linguistic Comput. 29 (2014) 139–142.

[20] T. Edwards, N. S. Holtzman, A meta-analysis of correlations between depression and first person singular pronoun use, Journal of Research in Personality 68 (2017) 63–68.

[21] D. L. Mowery, A. Park, C. J. Bryan, M. Conway, Towards automatically classifying depressive symptoms from twitter data for population health, in: PEOPLES@COLING, 2016.

[22] M. M. Greaves, A corpus linguistic analysis of public reddit and tumblr blog posts on non-suicidal self-injury, 2018.

[23] M. D. Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media, in: ICWSM, 2013.

[24] G. Colombo, P. Burnap, A. Hodorog, J. Scourfield, Analysing the connectivity and communication of suicidal users on twitter, Computer Communications 73 (2016) 291 – 300.

[25] V. Vasyliuk, Psycholinguistic text analysis for evaluation of person's emotional state, Abstracts of the 2nd International scientific and practical conference, Lviv (2019).

[26] J. M. Havigerová, J. Haviger, D. Kuera, P. Hoffmannová, Text-based detection of the risk of depression, Frontiers in Psychology 10 (2019).

[27] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, E. Cambria, Mentalbert: Publicly available pretrained language models for mental healthcare, CoRR abs/2110.15621 (2021). URL: https://arxiv.org/abs/2110.15621. arXiv:2110.15621.

[28] D. Hendrycks, K. Gimpel, Bridging nonlinearities and stochastic regularizers with gaussian error linear units, CoRR abs/1606.08415 (2016). URL: http://arxiv.org/abs/1606.08415. arXiv:1606.08415.

[29] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2019. arXiv:1711.05101.

[30] H. Zhang, M. Cissé, Y. Dauphin, D. Lopez-Paz, mixup: Beyond empirical risk minimization, ArXiv abs/1710.09412 (2018).