Towards transfer learning using BERT for early detection of self-harm of social media users

Qamar un Nisa [1] and Rafi Muhammad [1]

[1] *National University of Computer & Emerging Sciences - FAST, ST 4 Sector 17D Shah Latif Town Karachi 75030, Karachi, Pakistan*

Abstract

Online social media channels are everywhere and almost everyone uses them to post various parts of their lives that they are not comfortable sharing in real. Large amounts of data can be collected from these channels and used to study and development of systems that detect depression and tendency of self-harm in individuals on the internet. It is generally believed that social posts from a time order can be used to predict depression or signs of self-harm for a user. Deep learning models for Natural Language Processing tasks are already producing better results on language-based analysis of sentiment, offensive or depression detection for a collection of text. BERT and transformer-based architectures have proven to achieve state-of-the-art results for Natural Processing Tasks. We will apply transfer learning and supervised learning algorithm Logistic Regression on data to early detect signs of self-harm for a user. BERT is used to generate word embeddings of sentences, has a limitation of processing sentences only 512 tokens long, so we created custom algorithm to break sentences then generate embeddings then concatenate them for supervised learning. Depression leads to various forms of self-harm and our goal is to detect the tendency of self-harm in individuals as soon as possible to intervene.

Keywords

BERT transformers, Word Embeddings, Early Detection of Depression, Transfer Learning

1. Introduction

Globally, depression is a common illness affecting more than 264 million people [1]. If left unchecked, depression can get worse with time, resulting in self harm or worse; suicide, according to World Health Organization, one person commits suicide every 40 seconds [2] various helplines and awareness campaigns have been introduced to people in the hope of lowering suicide rate but not everyone has that awareness or access to connect with a professional. 15 to 29-year-old people are most affected by depression that leads to suicide, making it 2nd most subsequent cause of death for people in this age group. Technological evolution has changed habits of people due to advanced devices, social media networks and lifestyle, millennials prefer to text instead of talking to people on call or face to face. According to a survey by OpenMarket, which surveyed 500 millennials, "75% of millennials would rather lose the ability to talk versus text" [3]. With the rise of social media, there are various websites and channels where people openly talk about their struggles and battles with depression. There are support groups where people share their stories about depression, attempted suicide and other self-harm incidents. People prefer to text/write on social media about depression instead of seeing a professional for help, in some cases even preferring to talk to Bots. Because of social media channels

[1]CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
EMAIL: k180914@nu.edu.pk (A. 1); muhammad.rafi@nu.edu.pk (A. 2)

we now have some textual data that can be salvaged for early detection of depression. There are some hidden patterns and styles in textual data that can help identify an author or even gender. Shlomo Argamon et al [4] did research that stated that the gender of a writer can be identified based on how many pronouns, facts, and noun-specifiers are used by a writer. Female writers use more pronouns like "her, she, him" and male writers use more facts and noun specifiers like "the number of facts" [4]. Similarly depressed individuals can also be detected by using sophisticated natural language processing techniques. Technology is helping in all areas and with the help of new natural language processing methods and tools, depressed individuals hopefully can be identified using natural language processing techniques before their condition gets worse or they provide any self-harm.

Idea of Early Depression Detection (EDD) is to monitor texts of a user over time in chronological order and assess using different learning approaches whether a user is depressed, in early stages to intervene and get some help to user. Because of unavailability of different public datasets regarding early detection of depression in internet users based on their comments and texts on social media, there haven't been diverse solutions to the specific problem to check from other datasets. Researchers have used Natural Language approaches along with ensemble classifiers for EDD but very few have tried to solve this with state-of-the-art deep learning models. In this study, we hope to deal with the problem, looking at the previous techniques that yielded good results, work with them along with new models and result in a working solution. Granted, a machine learning model cannot be as effective as a professional counsellor or psychiatrist but not everyone has access to a professional and with the rise of social media and textual data, a sophisticated model can help internet users.

This paper describes an experiment to work on the early depression detection problem of classifying a user's text from reddit in chronological order and predicting signs of self-harm. The approach used in this experiment is using BERT to create word embeddings and then Logistic Regression used for prediction.

2. Related Work

Detection of depression through textual data is a new and challenging field of research as it has been identified 7-8 years back. Assessing depression online through series of text has immense impact, as a lot of people are suffering from depression and don't have access to mental health officers. There is lack of data publicly available for Early Depression Detection (EDD) due to many reasons, including that the data is collected from social media sites which prohibit distribution based on the GDPR laws. Team of engineering students classified anxiety using EEG signals, making use of IoT and Machine Learning but for that every person has to have the EEG headset [5], hence the detection of depression through social media is essential. Because of the incremental classification problem elevated by EDD, there aren't any baseline strategies to detect depression or modelling as well. There have been various approaches by different authors for early detection of depression, some using feature extraction, while others using ensemble of classifiers. Some authors have worked on the problem without taking user history and progress over time, while others have used time-aware methodology. David E. Losada (2016) collected data from different social sites, largely reddit to create a dataset for detection of depression, the dataset is collected from popular social forum "Reddit" [6]. Losada's dataset is used as baseline dataset for EDD at eRisk conference and other research projects. Losada also worked on creating a baseline method and evaluation metric called "Early Risk Detection Error" (ERDE) for models created to detect depression in texts over time. Different authors have used different types of machine learning, natural language processing or hybrid approaches to solve the problem.

2.1. Deep Learning

Deep Learning is a type of machine learning inspired by the structure and working of human brain, neurons connected to one another, in deep learning, it is called "Neural Network". Supervised Machine Learning algorithms need sets of features in data to be classified, but in deep learning features are identified and implicitly learned by the Neural Network itself. Deep learning techniques and frameworks like ULMfit, ELMo etc have been used to do word embeddings then passed on to text classifiers [7] or doing the text classification itself [8]. Elena Fano et al (2019) [7] used deep learning

framework ELMo to detect features, word embeddings and classified the output using multi-layer perceptrons.

M Trotzek et al. (2018) specified that language metadata is important in EDD text sequences, used Neural networks for word embeddings and classification [8]. Trotzek used pre-trained word vectors taken from Wikipedia and trained on fastText and GloVe. Trotken trained fastText model on reddit dataset that has 1.7 billion user texts and posts [9]. Trotken pre-processed the data that it also has punctuations, emoticons and also special character words since he believes based on linguistic research that a depressed user has certain textual attributes, and they are important to preserve too. Trotzek used Convolutional neural network for text classification, in addition to the reddit data, Trotzek also used metadata features as secondary input to the CNN.

Matero et al. (2019) pursued that depression detection for just one source of user writings is limited and proposed to use 3 different types of features or dimensions to predict instead of just textual data. He proposed to use Open Vocabulary features (word embeddings) from BERT, Theoretical dimensions like affect, intensity, valence, dominance etc and based on research said age and gender play a role too as the ratio of suicide is different in both genders. Third dimension he proposed is meta features of data like n-grams and statistics to get more inference about data. Matero proposed to use dual-context, one context from Reddit [6] writings and other features Meta or theoretical dimensions [10]. Matero didn't use time-aware classification methodology, so it is taking into account different features but not the user text history and used logistic regression to classify the risk of suicide. Matero evaluated his results with Accuracy and F1 score, no ERDE as there wasn't time aware methodology used.

3. Problem Statement – eRisk 2021 Task 2: Early Detection of Signs of Self-Harm

Consider a collection of social media posts of a user *U*, user posts several contents *P* in a time *n*. Posts of a user are stored in chronological order over time. The frequency of how much a user posts in a day, week etc. and the order of those posts in a time for example each day the posts *P* get more depressing and gradually have explicit words like 'I had a panic attack today', 'World would be better without me' etc. These type of analysis on the posts can be used to infer depressions and signs of self-harm early in a user.

$$U_{\text{posts}} = \{P_1, P_2, P_3 \ldots\ldots P_n\}$$

Idea is to propose a model that can identify a depressed user without the situation being critical that a user cause self-harm to themselves, through a sophisticated system, there can be intervention to help the user. Each post *P* of a user will be passed through the model sequentially and using time-aware methodology (considering past posts of user), user will be predicted by the model to be depressed or not depressed. Due to the nature of task, it is important that the model predicts user to be depressed in time, so not only does the model have to predict correctly but also in time as late prediction will not be useful in the given scenario.

We participated in Task 2 [11] for eRisk 2021, this task was also featured in 2019 [12] as Task 2 and in 2020 [26] as Task 1.

4. DataSet

The dataset used is following the same format as [6]. Dataset is a collection of a user's texts, comments on other people's posts, their own posts with titles, collected from social forum website Reddit. Dataset is anonymous and users are denoted by user numbers, additional data is the time and date of posts and comments. There are 2 classes of users "self-harm" and "not self-harm", denoted by 1 and 0 in data. The dataset released by CLEF eRisk 2021 [11] is used for the experiment and research.

Table 1
Statistics and details about the 2021 Task 2 Dataset [11].

	Training Data self-harm	Training Data not self-harm	Test Data self-harm	Test Data not self-harm
Number of Users	41	299	104	319
Number of Posts	6927	163506	11691	91136
Min. Posts per User	8	10	9	9
Max. Posts per User	997	1992	942	1990
Mean of Posts per User	168	546	112	285

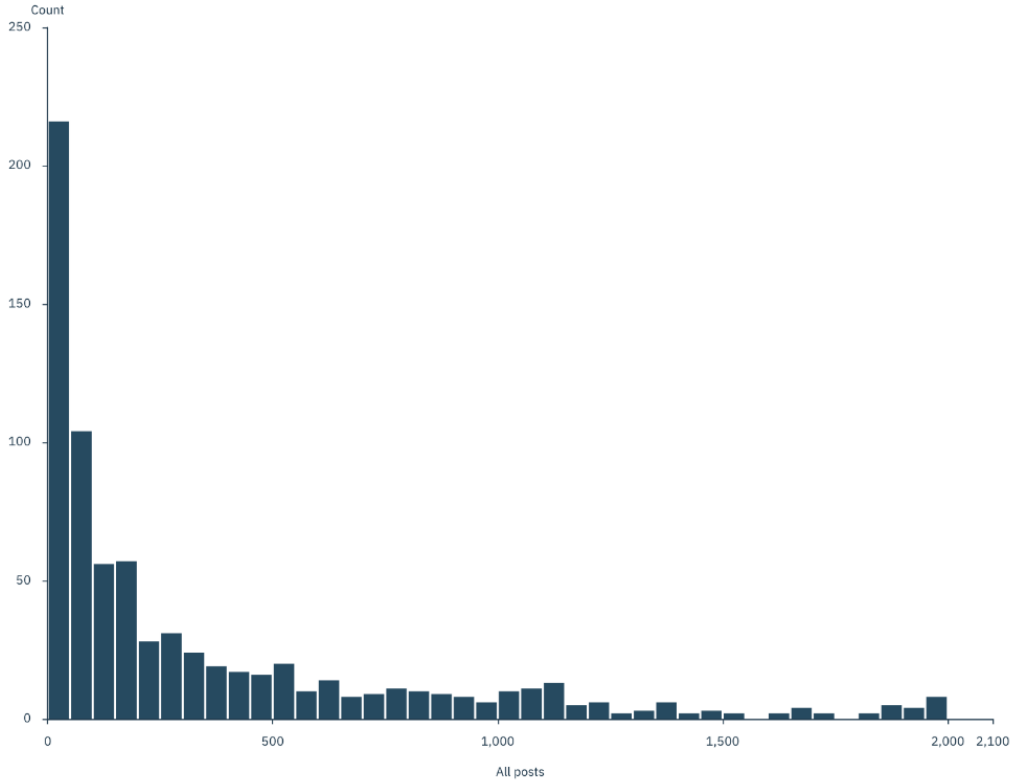

Figure 1: Histogram of all user writings (training and testing). Out of total 763 users, 376 users have writings from 8 to 150. This can also mean that 150 user writings should be processed in training stage if there is limitation of computing resources or time as majority users have writings no more than 150 and an optimal solution should give accurate results on first 150 user writings.

5. Methodology

Model that does additional natural language processing on data like bag of words, word embeddings then resulting data passed through classifier or ensemble classifiers has better outcome in past research. Word embeddings are the most important part of text classification tasks, there have been many pre-trained deep learning frameworks that provide state-of-the-art word embeddings. Elena Fano [7] used GLoVe and ELMo for word embeddings. BERT has limitation of processing only 512 tokens, hence not used by everyone and efficiently. Matero et al [10] used BERT similarly for word embeddings but they truncated the sentences in training due to the 512 token limitations, so all user writings weren't used to training. Trotzek [8] also experimented with BERT word embeddings but also stated the limitation of tokens and processing only first 512 tokens of sentences, [8] used fastText for word embeddings as they didn't have any limitation and provided better results. In this study we will use BERT [13] to generate word embeddings then use the extracted features to train supervised machine learning model, using Logistic Regression Classifier similar to [6].

5.1. Data Processing

The training data for Task 2 was collection of reddit user writings in XLM format, since the data was real collected from Reddit (a social networking forum) the data had to be cleaned, prepared to be processed to accommodate the 512-token limitation of BERT. First all URLs from each User Post is removed, then all words in the post are counted. An empty $String_S$ is initialized for each user to store user posts/writing. If the total words in post are less than 300 then we check further to see if concatenating user post with $String_S$ will result in more than 300 words or not. If not, then both are concatenated, and next post is processed and go through the same flow. If the post contains more than 300 words, then it is broken down into small sentences and split using punctuation marks, the same process of checking words again and concatenating with $String_s$ is applied then, if after splitting on punctuations the word count is still more than 300 then it means it's a long paragraph and is not being split using punctuations, we then split the long post using blank space and concatenate with $String_S$ till $String_S$ has 300 words. If $String_S$ and the new post's length after being concatenating is more than 300 then $String_S$ is saved in dataframe as a row and then emptied for next sentence. The whole flow of how the data was processed is shown in detail in Figure 2 as Flowchart. Each row in new created Data has UserId, newly processed sentences, Class Label of the user being depressed or not. Newly generated data of 340 users now has 16060 rows. Now prepared Data is then used to generate word embeddings using transfer learning on BERT.

The motivation behind this approach is that when a sentence is passed to BERT, if a sentence has 15 words it still outputs a vector of 786 dimension. Using this approach, smaller sentences in chronological order are concatenated instead of removing them as each post of a user is important. Other teams have truncated posts or removed small posts from data. Using this approach, we aren't losing any data and are also maintaining the chronological order of posts and the context of sentences.

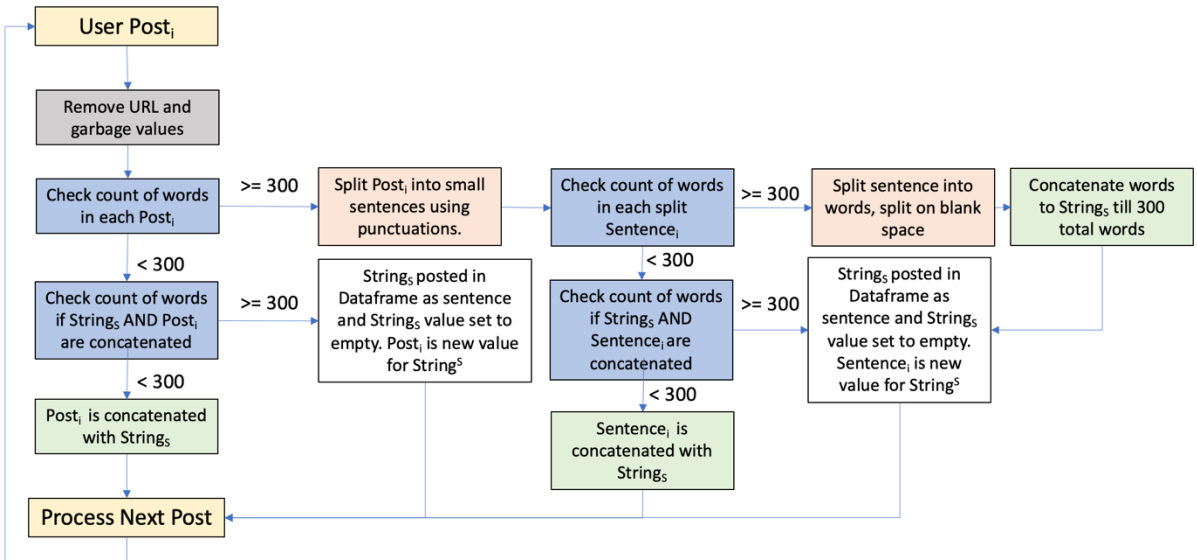

Figure 2: Flow-chart of Data Preparation Stage, breaking user posts into small sentences of 300 words. Data preparation is used for training classifier, not while testing as seen in Figure 3.

Table 2
Statistics and details about the 2021 T2 training Data, ready to be prepared to be processed on BERT.

	Data Stats	Comments

Total number of Users	340	-
Minimum Sentence per User	1	21 Users have just 1 sentence
Maximum Sentence per User	817	1 User has 817 sentences generated; 2nd highest is 491 sentences
Mean of Sentences	47	The average number of sentences is 47

5.2. BERT Word Embeddings

Bidirectional Encoder Representations from Transformers (BERT) was created by google [13] to out-perform various Natural Language Processing models like ELMo, ULMfit, Generative Pre-Training etc as they are all uni-directional and hardly bidirectional for downwards NLP streams. BERT is called as a method of pre-training language representations, we first pre-train a large general-purpose corpus then uses that for downstream NLP tasks that we are targeting like early detection in our study. BERT has pre-trained word embeddings of 30,000-word tokens, each word token having 768 features.

BERT pre-trained model is available via hugging face transformer library, BERT has different versions that vary in training parameters or cases like lower-case, upper-case, multilingual etc. We used *'bert-base-uncased'* for transfer learning to generate word embeddings for the sentences and user writings.

The prepared training data is passed to BERT in batches of 3000 rows to generate word embeddings. The word embeddings generated are of 16060 x 768 dimensions

Some more data processing is applied to the generated word embeddings to prepare for machine learning, all the rows of a single user are concatenated to have one row per user for classification. This resulted in 340 x 627456 dimensions. The columns were huge, so further dimensionality reduction was applied for faster processing and training of model. After reduction the size of columns is 1655.

5.3. Depression / Self-harm Classifier

Depression classifier is trained using processed word embeddings are then sampled on Logistic Regression using cross-validation to find the optimal weight and the optimal value of C for normalization. The optimal values are then applied on training data word embeddings and model is saved to be used for testing and prediction. We created another variation, where classifier is built using AutoML open-source library 'auto-sklearn', which applies different machine learning algorithms on data and does hyperparameter tuning based on the accuracy metric [17]. Logistic Regression is chosen as classifier to perform experiment on test-server data provided by CLEF eRisk team. The results on training data with different confidence score or threshold values are shown in Table 3. The confidence value of 0.5 on the logistic regression output gave the most optimal results in our experiment of predicting on test data set during training. The setting of CLEF eRisk states that once a prediction about a user is made, it cannot be changed later on so for first 5 user posts, we decided to use 0.75 confidence value. AutoML took longer time to train model and outputs bad scores as due to the sparsity of training data it is adding 0 in test data set, due to this AutoML model is predicting 0 unless the data is of same dimensions (1655) without the 0 padding at the end. AutoML model predicted just one sample to be positive but that was not true positive. AutoML model was going to be run2 but due to bad results in training we did not submit it and only submitted run of logistic regression model.

Table 3
Training results on evaluation metrics using Logistic Regression and AutoML classifier and different confidence scores as threshold values for prediction.

Classifier	Threshold Value	F1	P	R	$ERDE_5$	$ERDE_{50}$	$ERDE_{100}$
LR	0.5	0.387368	0.247978	0.884615	0.193439	0.190534	0.190534
LR	0.75	0.332203	0.256545	0.471154	0.215949	0.212559	0.212559
LR	0.8	0.253165	0.225564	0.288462	0.237447	0.234808	0.234808
AutoML	0.5	0.000000	0.000000	0.000000	0.246444	0.246444	0.246444

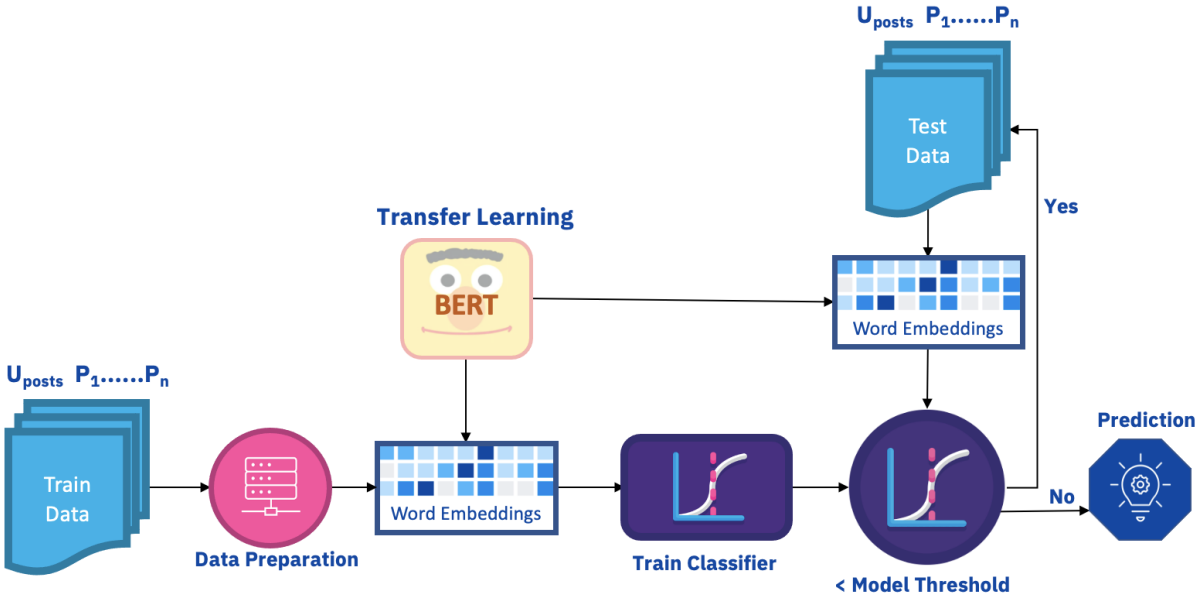

Figure 3: High Level Flow Diagram / Architecture Diagram of the experiment

5.4. Baseline Classifier

The research in Early Depression Detection [7] has been used as a benchmark by researchers in the domain [18]. Hence, for the experimental setup for this study, we have reproduced the experiment of [7] to be used as a Baseline Model, the experiment is implemented as follows:

CLEF eRisk Dataset has negative and positive users in the format discussed in Chapter 4.1. All negative and positive writings are appended to one corpus. Feature extraction on is performed by vectorizing the corpus *TfidfVectorizer*, having standard stop list in *English* language and minimum document frequency of 20 – words that appear in less than 20 user writings are removed. Vectorization is done with sklearn library in python. Resulting vector is stored in a sparse 486x12968 array.

Logistic regression is used for classification. Vector of ground truth values is created for 486 users, denoting depressed as 1 and not depressed as 0. Implementation of Logistic Regression is done using sklearn library of Logistic Regression, in this experiment logistic regression is taking 4 parameters. *Penalty* to specify the normalization which is 'L1' in our study. *C* is the penalty parameter accompanying with the error term of the optimisation of model, like in SVM's *C* should have smaller value for strong regularization. *class_weight* is the last parameter to specify weights for classes, this is useful for data which has class imbalance.

Values of *C* and *class_weight* is optimized using the standard protocol by Chih-Wei hsu [34], grid search with exponentially growing sequences of $C = 2^{-10}, 2^{-4}, ..., 29$ *and* $w = 20, 21, ..., 29$ is applied on parameters. 4-fold cross-validation is done on eRisk 2021 t2 training data with the above sequences to find the optimal value. The value of *C* and *w* that maximizes F1 score, precision and recall, in [7] is C=16 and *w*=4. But we didn't get that value while experimenting. Based on the 4-fold cross-validation the got two sets of values that optimized F1 score, C=256 & w=64 and C=128 & w=32. The respective values are used with other hyperparameters, and logistic regression is applied on training data to build a self-harm classifier. We created two Baseline classifiers with different hyperparameters, to compare with our models.

6. Evaluation Metrics

Several evaluation metrics are proposed and widely used for the specified task, shared in detail below as mentioned in [11]:

6.1. Early Risk Detection Error (ERDE)

In detecting a serious illness or disease that causes harm to an individual, time is of the essence and detecting at right time to take preventive measures before the condition worsens. ERDE is a metric [6] to measure the error or accuracy of the model, too early detection can lead to false positive classifications or too late detection can lead to serious harm so the prediction should be at an optimal time for intervention, ERDE measures that for the model.

The delay in model is measured by the total user posts depression classifier processed to make the classification, denoted by *k*. Since positive classes in the cases is low, we can suffer "class imbalance" problem, so each error is weighted differently in ERDE.

Here, EDD system has made a classification decision of a user being depressed or not, *d* at a given time *t*. The prediction *d* can lead one of the four golden truth judgements: True Positive, True Negative, False Positive, False Negative. ERDE measure based on these four cases, is defines as:

$$ERDE_o(d,k) = \begin{cases} c_{fp} & \text{if } d = \text{positive AND ground truth} = \text{negative (FP)} \\ c_{fn} & \text{if } d = \text{negative AND ground truth} = \text{positive (FN)} \\ lc_o(k) \cdot c_{tp} & \text{if } d = \text{positive AND ground truth} = \text{positive (TP)} \\ 0 & \text{if } d = \text{negative AND ground truth} = \text{negative (TN)} \end{cases}$$

In different studies, sometimes the positive class is majority or sometimes negative is majority, in our study the positive classes are in minority. To build a measure that takes into consideration the class imbalance, c*fn* here is set to 1 and it should be greater than c*fp*, c*fp* can be set to be some proportion of positive examples (e.g., if there are 3% positive examples then set c*fp* to 0.03). The third factor, latency cost lco(*k*) (∈ [0, 1]) encodes a penalty if true positives are identified with delay and is set as [6]

$$lc_o(k) = 1 - \frac{1}{1 + e^{k-o}} \quad (2)$$

This function is parametrized by *o*, which controls the point where the cost grows more quickly. The error is calculated by taking mean of all classification of ERDE values. [11]

$ERDE_o$ is a baseline proposed data stream classification metric where not only do we have to make prediction but also identify when to make it, which is crucial for Early detection of depression and self-harm domain.

6.2. F-latency

CLEF eRisk 2021. [11] proposed to use Flatency, in addition to ERDE metric, to get more interpretable evaluations. Flatency was proposed by Sadeque and colleagues [16]. F-latency is also called and referred to as Latency-weighted F1. Which is F1 score multiplied by speed, speed is overall factor where if a user is correctly identified and classified in their first post and it is close to 0 for slow systems that have to process hundred or more writings of a user to identify true positives.

6.3. Standard Classification Measures

Whenever machine learning models are mentioned for supervised learning, the Accuracy of the model is an important metric to look at, but for some cases the accuracy of the model doesn't specify or give a whole view of the model therefore other standard classification measures like Precision, Recall and F-measure are used to evaluate the model further. Precision, Recall and F1 Score classification metrics are used in CLEF eRisk tasks related to depression as well [11].

7. Result & Discussion

The experiment delivered good results for $ERDE_o$ [11] and Recall, but the accuracy and efficiency of model is not optimal and leaves room for improvement. The confidence value of 0.5 used as threshold by Logistic Regression Model is increasing F1 Score slightly, the recall is increased drastically as compared to other results of 0.75 and 0.8 confidence score threshold value. ERDE is decreased with 0.5 confidence score as well. Submission of CLEF eRisk 2021 Task 2 submission was done on 6 user writings, which is low number to see the performance of whole model, so we applied the model after the submission to test what the result would have been if more writings were processed and predicted. We also tested our model with Baseline model created using the approach in [7]. The $ERDE_o$ of our runs had better $ERDE_o$ than the baseline model, our proposed model had better recall in train stage too as compared to the result of baseline model on train data. Explanation of runs along with predictions and result on metrics are shared in Table 4.

Table 4
Result of classifier on test-server data predictions. LR-T2 run1 & run2 are the runs submitted for CLEF eRisk 2021 task2 - *Early Detection of Signs of Self-Harm*. LR-T2 test run1 & run2 are experiments performed using same model but run on more user posts after the task2 submission.

Classifier	Number of Posts	F1	P	R	$ERDE_5$	$ERDE_{50}$
LR – T2 run1	6	0.172	0.124	0.283	0.101	0.097
LR – T2 run2	6	0.172	0.124	0.283	0.101	0.097
LR – T2 test run1	104	0.185	0.103	0.848	0.097	0.096
LR – T2 test run2	107	0.187	0.105	0.855	0.096	0.095
Baseline-w-64[7]	All	0.545	0.419	0.778	0.141	0.122
Baseline-w-32[7]	All	0.558	0.441	0.759	0.139	0.12

Run 1 is not cleaning URLs, punctuation marks and bad formats like #8217 for encoding of apostrophe (') etc., on user posts processed, whereas Run 2 is processing user posts and cleaning data before generating word embeddings. Run 2 isn't much different from Run 1 as there is very slight change in results of them, both not being optimal for Precision, F1 score. The recall on model is good. Run1 classified 1241 users as positive, out of those only 129 are true positives and Run1 classified 1232 as positive, correctly identifying 130 users. Run1 has read 104 user writings to make all predictions,

instead of reading all, Run2 has read 107 user writings to make all predictions. In first 5 posts, 1130 users were classified in Run2 making the speed of experiment to be optimal, details shown in Figure 4.

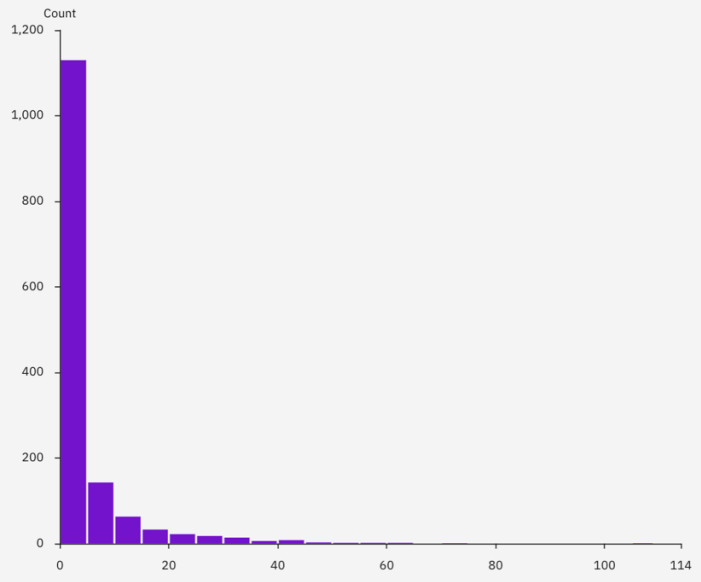

Figure 4: Histogram of user writings read for each user for classification of Run2.

7.1. Future work

The approach of Data Preparation is unique, due to the limitation of resources and time, we couldn't train BERT or RoBERTa on our generated data and used Transfer Learning instead. The proposed approach can be improved in future by following methodologies:

- The approach of Data Preparation is unique, due to the limitation of resources and time, we could not train BERT or RoBERTa on our generated data and used Transfer Learning instead. Own model on prepared data using our approach can be built from scratch or fine-tuned on BERT, RoBERTa to get more accurate results specific to the domain of Early Risk Prediction.
- Similar to team iLab from CLEF eRisk 2020 [14], another experiment can be done to process Test data similar to Train data in our experiment, pass-through classification model and the output probability averaged for all generated sentences and make decision if a user is depressed or not.
- Similar to team iLab from CLEF eRisk 2020 [14], another experiment can be done to process Test data similar to Train data in our experiment, pass-through classification model and the output probability averaged for all generated sentences and make decision if a user is at risk or not
- Instead of using BERT for extracting word embeddings, use BERT and its different architectures for classification and compare for better model.
- Data can be experimented and compared with different models of BERT like RoBERTa or more state-of-the-art models like GPT-3.

8. Acknowledgements

I would like to thank my university National University of Computers & Emerging Sciences, my research supervisor, Muhammd Rafi and my colleagues at IBM, Asna Javed, Mohammad Ali Khan and Fawaz Siddiqui for supporting me with my research and helping out with few technicalities.

9. References

[1] GBD 2017 Disease and Injury Incidence and Prevalence Collaborators. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990–2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet.

[2] World Health Organization, Mental Health and Substance use, 2016. URL: https://www.who.int/teams/mental-health-and-substance-use/suicide-data.

[3] J. Loechner, Text vs. Talk Gets Millennials' Attention, 2016. URL: https://www.mediapost.com/publications/article/275332/text-vs-talk-gets-millennials-attention.html.

[4] S. Argamon, J. Fine, Gender, Genre, and Writing Style in Formal Written Texts, Text - Interdisciplinary Journal for the Study of Discourse 23(3) (2003) doi:10.1515/text.2003.014.

[5] A. Arsalan, M. Majid, S. M. Anwar, Electroencephalography Based Machine Learning Framework for Anxiety Classification, (2019), in: Bajwa, I. Sarwar, Sibalija Et al., Intelligent Technologies and Applications, Second International Conference, INTAP 2019, Bahawalpur, Pakistan, 2020, pp. 187-197. doi:10.1007/978-981-15-5232-8_17.

[6] D. E. Losada, F. Crestani, A Test Collection for Research on Depression and Language Use, in: International Conference of the Cross-Language Evaluation Forum for European Languages (2016). doi:0.1007/978-3-319-44564-9_3

[7] E. Fano, J. Karlgren, N. Joakim, Uppsala University and Gavagai at CLEF eRISK: Comparing Word Embedding Models, in: Working Notes of CLEF 2019 - Conference and Labs of the Evaluation Forum, Lugano, Switzerland, September 9-12, 2019, CEUR-WS.org , 2019

[8] M. Trotzek, S. Koitka, and C. M. Friedrich, Utilizing Neural Networks and Linguistic Metadata for Early Detection of Depression Indications in Text Sequences, IEEE Transactions on Knowledge and Data Engineering 32(3) (2018) 588-601. doi:10.1109/TKDE.2018.2885515.

[9] Reddit, I have every publicly available Reddit comment for research. ~ 1.7 billion comments @ 250 GB compressed. Any interest in this?, 2016. URL: https://www.reddit.com/r/datasets/comments/3bxlg7/i_have_every_publicly_available_reddit_comment/

[10] M. Matero, A. Idnani, Y. Son, Et al., Suicide Risk Assessment with Multi-level Dual-Context Language and BERT, in: Proceedings of the Sixth Workshop on Computational Linguistics and Clinical Psychology (2019). doi:10.18653/v1/W19-3005.

[11] Parapar J., Martín-Rodilla P., Losada D.E., Crestani F. (2021) eRisk 2021: Pathological Gambling, Self-harm and Depression Challenges. In: Hiemstra D., Moens MF., Mothe J., Perego R., Potthast M., Sebastiani F. (eds) Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science, vol 12657.

[12] D. E. Losada, F. Crestani, J. Parapar, (2019) Overview of eRisk 2019 Early Risk Prediction on the Internet. in: Crestani, F., Braschler, M., Savoy, J., Rauber, A., Müller, H., Losada, D.E., Heinatz Bürki, G., Cappellato, L., Ferro, N. (eds) Experimental IR Meets Multilinguality, Multimodality, and Interaction. 10th

International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019.
[13] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, 2019. doi:10.18653/v1/N19-1423
[14] R. Martinez-Castano, A. Htait, L. Azzopardi, Y. Moshfeghi, (2020) Early risk detection of self-harm and depression severity using BERT-based transformers : iLab at CLEF eRisk 2020. In: Early Risk Prediction on the Internet, 2020-09-22 - 2020-09-25.doi
[15] J. M. Loyola, M. L. Errecalde, H. J. Escalante, M. Montes, Learning When to Classify for Early Text Classification, in: Computer Science – CACIC 2017 (pp.24-34) (2018). doi:10.1007/978-3-319-75214-3_3
[16] F. Sadeque, D. Xu, S. Bethard, Measuring the Latency of Depression Detection in Social Media, in: Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM 2018, pp. 495–503. ACM, 2018.
[17] M. Feurer, A. Klein, K. Eggensperger, J. Springberg, M. Blum, F. Hutter, Efficient and Robust Automated Machine Learning, in: Neural Information Processing Systems 28, NIPS 2015, pp. 2862-2970