NLP-IISERB@eRisk2022: Exploring the Potential of Bag of Words, Document Embeddings and Transformer Based Framework for Early Prediction of Eating Disorder, Depression and Pathological Gambling Over Social Media

Harshvardhan Srivastava[1], Lijin N S[2], Sruthi S[2] and Tanmay Basu[2]

[1]*Oracle India Private Limited, Bangalore, India*
[2]*Department of Data Science and Engineering, Indian Institute of Science Education and Research Bhopal, India*

Abstract

The eRisk lab at CLEF 2022 had released three different tasks based on the posts of different users over Reddit, a popular social media. The first task was early detection of signs of pathological gambling. The second task was the early prediction of depression. The third one was assessing the severity of eating disorders over social media posts. The BioNLP research group at the Indian Institute of Science Education and Research Bhopal (IISERB) participated in all three tasks and submitted five runs using five different text mining frameworks for task 1 and task 2 and four different runs for task 3. The methods involve different feature engineering schemes and text classification techniques. The performance of the classical bag of words model, paragraph embedding technique and transformer-based models were explored to identify significant features from the given corpora. Moreover, we have identified features based on the biomedical concepts for pathological gambling using Unified Medical Language Systems, a repository for biomedical vocabularies. Subsequently, we have explored the performance of different classifiers, e.g., logistic regression, random forest etc. using various such features generated from the given data. The official results on the test data of individual tasks show that the proposed frameworks achieve top scores in terms of some of the evaluation techniques, e.g., precision, F1 score, speed etc. for all three tasks. The paper describes the performance, value and validity of the proposed frameworks for individual tasks and the scopes for further improvement.

Keywords

information extraction, depression detection, identification of eating disorder, text classification, clinical text mining, biomedical NLP

1. Introduction

Early risk prediction is a new research area potentially applicable to various situations, such as identifying people with a risk of mental disorders, which have become a predominant issue today. Especially for the population of people living in conflicted-affected areas, the chance of

CLEF 2022 – Conference and Labs of the Evaluation Forum, September 05–08, 2022, Bologna, Italy
srivastavahv@gmail.com (H. Srivastava); lijin19@iiserb.ac.in (L. N. S); sruthi19@iiserb.ac.in (S. S); welcometanmay@gmail.com (T. Basu)
https://sites.google.com/view/tanmaybasu/ (T. Basu)
0000-0001-9536-8075 (T. Basu)

exposure to activities which can mentally affect them is very high. As mentioned in the study by Charlson et al. [1], the estimated prevalence of mental disorders (i.e., depression, anxiety, post-traumatic stress disorder, bipolar disorder, and schizophrenia) was 22·1% (95% UI 18·8–25·7) at any point in time when assessed in the conflict-affected populations. A strong correlation also exists between the income class of an individual and the mental illnesses associated with that person and low levels of household income are associated with several lifetime mental disorders and suicide attempts, and a reduction in household income is associated with increased risk for incident mental disorders [2].

With the advent of internet, online social platforms have become a regular media for almost all the people to share and express their thoughts and feelings freely and publicly with other people [3]. The information available over social media is a rich source for sentiment analysis or inferring mental health issues [4]. The CLEF eRisk 2022 shared task focuses on three tasks i.e., (i) early detection of signs of pathological gambling, (ii) early detection of depression and (iii) measuring the severity of the signs of eating disorders. The main goal of the eRisk 2022 challenge is to instigate discussion on the creation of reusable benchmarks for evaluating early risk detection algorithms by exploring issues of evaluation methodology, effectiveness metrics and other processes. Early identification advances can be utilized in various areas, especially those connected with the health and security of users interacting on the web and to identify the potential predators on the internet. The lab had organized three tasks this year and released different corpora for the individual tasks, which were developed using the postings of individual users over Reddit, a popular social media. We, the BioNLP group at IISERB, participated in all three tasks and performed reasonably well.

The performance of different feature engineering schemes and classification techniques was explored to identify pathological gambling, depression and eating disorder from the posts of the users over social media data released as part of individual shared tasks of the eRisk 2022. The proposed framework for task 1 and task 2 aims to train a machine learning classifier by using different types of features generated from the given training corpus to classify the documents of the test data. Note that the performance of a text classification technique is highly dependent on the important features of a corpus. Therefore the performance of different classifiers has been tested following different feature engineering schemes. The classical bag of words (BOW) model [5], paragraph embeddings [6] and transformer architecture based deep learning models were used to generate features from the given corpora. Two different term weighting schemes were used for the BOW model, viz., term frequency and inverse document frequency-based term weighting scheme [5] and entropy-based term weighting scheme [7]. Furthermore, four different attention layer-based deep learning models, namely, BERT (Bidirectional Encoder Representations from Transformers) [8], BioBERT[9], RoBERTa[10] and Longformer[11] were used to generate semantic features from the given training data.

Subsequently, the performance of ada boost [12], logistic regression [13], random forest [14] and support vector machine [15] classifiers have been reported using the BOW features and the paragraph embeddings based features individually on the training corpus following 10 fold cross-validation technique. Therefore, the best five frameworks were chosen based on their performance on the training corpus in terms of F1 score and subsequently, they have been implemented on the test corpus. Similarly, the features generated by a transformer-based architecture were used to train the classifier of the same architecture using the training data

following 10 fold cross-validation technique. Based on the decision-based results of task 1, the proposed Longformer model achieved the best score among all the submissions in terms of recall. The random forest classifier following the entropy-based term weighting scheme, achieved the top score in terms of recall, latency$_{TP}$ and speed among all the runs of task 1. The proposed entropy-based term weighting scheme using support vector (SVM) classifier outperforms the other runs in terms of F1 score and latency-weighted F1 score [16] for decision-based results of task 2. For task 3, the semantic similarity between a given question and the posts of the Reddit users were identified using different similarity measures, e.g., Jaro-Winkler distance [17], Cosine similarity [18] etc. The official results show that the proposed method using a pretrained BERT model and cosine similarity measure performed better than all the runs submitted by different teams in almost all the evaluation techniques used for task 3.

The paper is organized as follows. Section 2 describes the proposed frameworks for individual tasks. The experimental results are reported and analyzed in section 3. The conclusions and scopes of further works are presented in section 4.

2. Proposed Frameworks for Individual Tasks

Different text mining frameworks were proposed based on the requirements of individual tasks. The documents of the given corpora for individual tasks were released in XML format. Each XML document contains the postings of a Reddit user over a period of time with the corresponding dates. We extracted these postings from the XML documents and ignored the other entries. Therefore the corpus used for experiments in this article contains only the texts related to different posts on Reddit for individual users. The proposed frameworks for task 1 and task 2 include different feature engineering schemes and classification techniques. For task 3, the proposed framework uses various semantic similarity measures to identify the similarity between a given question and possible answers among the posts of individual Reddit users.

2.1. Feature Engineering Schemes for Task 1 and Task 2

2.1.1. Bag Of Words Features

The text documents are generally represented by the bag of words (BOW) model [5]. In this model, each document in a corpus is generally represented by a vector, whose length is equal to the number of unique terms, also known as vocabulary [5]. The conventional term weighting scheme is known as term frequency and inverse document frequency or *tf-idf*. Document frequency (df) is the number of documents in which a term appears. Inverse document frequency determines how frequently a term occurs in a corpus and it is defined as $idf_{term} = log(\frac{\#documents}{df_{term}})$. The weight of a term in a document, is determined by multiplying its term frequency with inverse document frequency. Moreover, the entropy based term weighting technique is used by many researchers to form term-document matrix from a text collection [7]. This method developed in the spirit that the more important term is the more frequent one that occurs in fewer documents, taking the distribution of the term over the corpus into account [7]. The weight of a term in a document, is determined by the entropy[1] of term frequency of the

[1] https://radimrehurek.com/gensim/models/logentropy_model.html

term in that document [7].

The BOW model generally creates sparse and high dimensional term-document matrices, which may affect the performance of the classifiers. Hence χ^2-statistic [19] based term selection technique was used to identify important terms from the term-document matrix, which is a widely used technique for term selection [19]. We have considered different number of terms generated by χ^2-statistic and evaluated the performance of individual classifiers using these set of terms on the training corpus. The best set of terms are used for experiments on the test data. These BOW features are used for the given data of task 1 and task 2.

2.1.2. Paragraph Embeddings Based Features

The unsupervised paragraph embeddings technique, also known as Doc2Vec model can express a document as a vector[6], which can identify semantic similarity between two documents by comparing the corresponding vectors. It was developed based on unsupervised Continuous Bag of Words (CBOW) and Skip-grams model, which expresses a word as a vector [20] using a given corpus and combines them to learn paragraph or document level embeddings [6]. The Doc2Vec model is trained on the individual training corpora of task 1 and task 2 to generate the embeddings from individual documents of these corpora. Therefore it was used to generate the features for individual documents of the test data for task1 and task 2. The number of such features was fixed by performing 10-fold cross validation technique on the training data.

2.1.3. UMLS Features

We have also considered the UMLS (Unified Medical Language System) [21] concepts extracted from the text as features for task 1 only. We could not find many such features for task 2 data and hence did not use it for task 2. UMLS is a comprehensive list of biomedical terms for developing automated systems capable of understanding the specialized vocabulary used in biomedicine and health care [21]. In UMLS there are 133 semantic categories[2] related to biomedicine and health. The semantic category of a term can be identified using MetaMap[3], a tool to recognize UMLS concepts in text data [22]. MetaMap first breaks the text into terms and then for each term it returns different semantic categories and ranked these categories according to a confidence score. It generates a Concept Unique Identifier (CUI) for each term belong to a particular semantic category [22]. We have used these CUIs as features and they are called UMLS features in this paper. UMLS features belong to five relevant semantic categories e.g., acquired abnormality, mental and behavioral dysfunction, etc. were considered for experiments for task 1.

2.2. Text Classification Techniques for Task 1 and Task 2

2.2.1. Classical Methods

Different text classification methods were used for task 1 and task 2 using BOW features, features generated by Doc2Vec model and UMLS features. The Adaptive Boosting (AB), Logistic

[2] https://mmtx.nlm.nih.gov/MMTx/semanticTypes.shtml
[3] https://metamap.nlm.nih.gov

Regression (LR), Random Forest (RF) and Support Vector Machine (SVM) classifiers were used for task 1 and task 2. The significant parameters of the individual classifiers were selected by using the grid search technique[4] following 10-fold cross validation model on the training data.

2.2.2. Transformer Architecture Based Embeddings

Multiple transformer architecture based models were used for task 1 to get the best embeddings for the given training corpus. The aim was to capture long range dependency and context of the conversations effectively. The first model that we explored is BERT (Bidirectional Encoder Representations from Transformers), which is a contextualized word representation model that is based on a masked language model and pre-trained using bidirectional transformers [8]. It was pre-trained on general domain corpora i.e., English Wikipedia and books [8]. We also explored two widely used extensions of BERT i.e., BioBERT [9], which is trained on PubMED articles and RoBERTa [10] that is trained on a news corpus by fixing some specific parameters and training strategies of BERT. Another alternative of BERT, the Longformer model has significant advantages over BERT to identify long term dependency in the given texts [11]. It presents a different attention mechanism that developed in conjunction with successive length of the document size using a sliding window technique [11]. We have used the pretrained models of BERT, BioBERT, RoBERTa and Longformer from the Hugging Face repository[5] and fine-tuned them individually on the given training corpus of task 1 and another Reddit data for pathological gambling [23].

2.3. Semantic Similarity Based Measures for Task 3

The objective of task 3 is to fill out a standard eating disorder questionnaire based on the evidence found in the history of postings of individual users. Hence the aim here is to find the contextual similarity between a given question and the posts of users for a period of time to generate a score between 0 to 6 to identify the severity of eating disorder. The performance of the following semantic similarity measures are explored in order to achieve this objective.

Jaccard similarity [18, 24, 25] is the ratio of common words between two sets of texts and the total unique words of these two sets. It ranges in [0,1], where 1 represents highest similarity and 0 represents no similarity between two sets of texts. Let X and Y be two sets of texts. The Jaccard similarity between X and Y can be defined as

$$\text{Jaccard (X,Y)} = \frac{|X \cap Y|}{|X \cup Y|}$$

Jaro-Winkler distance (JD) [17] is a string metric used for estimating the edit distance between two sets of texts. The lower the Jaro–Winkler distance for two strings is, the more similar the strings are. The score is normalized such that 1 means an exact match and 0 means there is no similarity. The Jaro–Winkler distance between X and Y is defined as follows:

$$\text{Jaro-Winkler (X,Y)} = 1 - \begin{cases} 0, & \text{if m = 0} \\ \frac{1}{3}\left(\frac{m}{|X|} + \frac{m}{|Y|} + \frac{m-t}{m}\right), & \text{otherwise} \end{cases}$$

[4] https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.GridSearchCV.html
[5] https://huggingface.co/

where m and t are respectively the number of common characters and number of transpositions between X and Y. Cahyono had shown that the Jaro-Winkler distance worked very well for plagiarism detection [26]. However, to our knowledge, this distance function has never used for identifying the severity of eating disorders or in any other such shared tasks of the earlier eRisk labs.

Cosine similarity between two documents [5] is measured as the similarity of the cosine of the angle between two document vectors. Cosine similarity between X and Y can be defined as

$$\text{Cos (X,Y)} = \frac{\vec{X} \cdot \vec{Y}}{|\vec{X}||\vec{Y}|}$$

Here $\vec{X}$ and $\vec{Y}$ are represented following the tf-idf weighting scheme of the BOW model. Cosine similarity [5] ranges in [0,1], where 1 indicates highest similarity and 0 indicates no similarity.

3. Experimental Evaluation

3.1. Datasets

The organizers released individual corpora for the given tasks using the postings of the users over Reddit for a given time period. The data were released in XML format with the identity, timestamp, title and postings of individual users.

3.1.1. Task 1:

The given training corpus of task 1 had two categories - pathological gambling and control group. In the training data, 164 users were marked as pathological gamblers and 2184 users were marked as control group, whereas in the test corpus, 81 users were marked as pathological gamblers, and 1998 users were marked as control group. The above statistics of the dataset clearly indicate that the users marked as pathological gamblers are observably smaller than the control group which during the training period creates preference by the models to treat the pathological gambling class as a stochastic error and created problems when generalising the values. In addition to the given training corpus, we have used two other Reddit corpora for pathological gambling [23][6][7] and added them to the pathological gambling class of the given training data to train different classifiers. Posts in these two external Reddit corpora are mostly related to gambling addiction [23]. We had done the experiments using both the given training data for task 1 and adding the external Reddit corpus to the given training data.

3.1.2. Task 2:

The given training corpus of task2 had two categories: the depression and control groups. In the training data, 214 users were marked as depressed users and 1493 users were marked as control group, whereas in the test corpus, 98 users were marked as depressed group and 1302 users were marked as control group. No external data was used to train the classifiers for task 2.

[6] https://www.reddit.com/r/GamblingAddiction/
[7] https://www.reddit.com/r/problemgambling/

3.1.3. Task 3:

The data set of task 3 comprises postings of individual users for a given period of time and a questionnaire having 22 different questions. The goal is to assess the degree of severity of eating disorders (scaled between 0 to 6) of a user for each of these questions based on Reddit postings. The ratings are an indication of the degree of agreement that the user has with the question, 0 meaning that the user in in disagreement with the hypothesis of the question and 6 meaning that the user is in maximum agreement to the hypothesis. Since no ground truths were provided for this data set, we used an anorexia data set of eRisk 2018 shared task 2 [27] to train the BERT model [8] in one of our runs submitted for this task.

3.2. Experimental Setup

We have submitted multiple runs following different frameworks for each of the tasks. For task 1 and task 2, we have evaluated the performance of different feature engineering techniques and the classifiers following 10 fold cross-validation method on the training corpus. We have chosen the five best frameworks to be tested on the test corpus. AB, LR, RF and SVM classifiers are implemented in Scikit-learn[8], a machine learning tool in Python. To overcome the effect of majority class over the classifiers, the balanced mode is used for each classifier, which automatically adjust weights of individual classes inversely proportional to the class frequencies in the training data[9]. Doc2Vec is implemented in Gensim[10], a deep learning library package in Python. We have used BERT, Bio-BERT, RoBERTa and Longformer models from the HuggingFace library[11].

The performance of the proposed frameworks using the training set were evaluated in terms of precision, recall and F1 score [28]. In addition to that, the organizers evaluated the performance of the runs in terms of $ERDE_5$ [29] $ERDE_{50}$[29], $latency_{TP}$ [30], speed [30] and latency-weighted F1 score [30].

The performance of the runs of task 3 was evaluated in terms of Mean Zero-One Error (MZOE), Mean Absolute Error (MAE), Macroaveraged Mean Absolute Error (MAE_{macro}), Restraint Subscale (RS), Eating Concern Subscale (ECS), Shape Concern Subscale (SCS), Weight Concern Subscale (WCS) and Global Eating Disorder (GED) [16]. These evaluation techniques are described in the overview paper of the eRisk 2022 shared task [16].

3.3. Analysis of Results

3.3.1. Task 1: Early Prediction of Pathological Gambling

Initially, we have implemented four classifiers using three different feature engineering schemes individually on the given training corpus. Moreover, we have used two relevant Reddit data

[8] http://scikit-learn.org/stable/supervised_learning.html
[9] https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html
[10] https://radimrehurek.com/gensim/models/doc2vec.html#gensim.models.doc2vec.Doc2Vec
[11] https://huggingface.co/

Table 1
Task1: Performance of Different Frameworks on the Training Corpus

Feature Types	Classifier	Precision	Recall	F1 Score
Entropy Based BOW Features (Using given training data)	AdaBoost	0.98	0.99	0.98
	Logistic Regression	0.91	0.95	0.93
	Random Forest	**1.00**	**1.00**	**1.00**
	Support Vector Machine	0.95	0.92	0.94
Entropy Based BOW Features (Using a Reddit data from another resource along with given training data)	AdaBoost	0.97	0.88	0.92
	Logistic Regression	0.81	0.95	0.86
	Random Forest	**0.98**	**0.97**	**0.97**
	Support Vector Machine	0.92	0.87	0.89
Entropy Based BOW and UMLS Features (Using given training data)	AdaBoost	**0.98**	**0.98**	**0.98**
	Logistic Regression	0.88	0.96	0.92
	Random Forest	0.96	0.96	0.96
	Support Vector Machine	0.89	0.94	0.91
TF-IDF Based BOW Features (Using given training data)	AdaBoost	0.98	0.98	0.98
	Logistic Regression	0.86	0.95	0.90
	Random Forest	**1.00**	0.95	**0.97**
	Support Vector Machine	0.93	0.95	0.94
Doc2Vec Based Features (Using given training data)	AdaBoost	0.92	0.89	0.90
	Logistic Regression	0.90	0.96	0.92
	Random Forest	0.98	0.86	0.91
	Support Vector Machine	0.89	0.95	0.92
Transformer Based Features (Using given training data)	BERT	0.98	0.77	0.84
	RoBERTa	0.98	0.74	0.82
	Longformer	0.94	0.89	**0.91**
	BioBERT	0.7	0.85	0.75

sets collected from different resources [23] and appended them to the pathological gambling category of the given training data. Subsequently, experiments conducted using all three feature engineering schemes and the classifiers on this appended data set. However, only entropy-based BOW features worked well on this appended data set and hence we reported these results in Table 1. Moreover, we reported the results by combining BOW and UMLS features following entropy-based term weighting scheme for all classifiers. We had implemented all the classifiers using just UMLS features following entropy-based term weighting scheme. However, none of the classifiers performed reasonably well, and hence we did not report these results in Table 1. The performances of these frameworks are reported in Table 1 in terms of precision, recall and F1-score. These results help to analyze the performance of proposed frameworks on the training set. Thereafter, the top five frameworks from Table 1 in terms of F1-score were selected and subsequently implemented on the given test corpus. Eventually, the performance of these five frameworks on the test corpus was communicated as official results of our team for task 1.

It can be seen from Table 1 that RF performs better than the other classifiers in terms of F1 score following the Entropy-based term weighting scheme of the BOW model using both the given training data and a relevant Reddit dataset collected from another resource [23].

Table 2
Task1: Decision-Based Results obtained on the test set

Runs	P	R	F1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	latency weighted F1
NLPGroup-IISERB 0 (BOW+TF-IDF+RF)	0.107	0.642	0.183	0.030	0.025	2.0	0.996	0.182
NLPGroup-IISERB 1 (BOW+Entropy+RF)	0.044	**1.000**	0.084	0.046	0.033	3.0	0.992	0.083
NLPGroup-IISERB 2 (BOW+Entropy+RF)†	0.043	**1.000**	0.083	0.041	0.034	**1.0**	**1.000**	0.083
NLPGroup-IISERB 3 (Longformer)	0.140	**1.000**	**0.246**	**0.025**	**0.014**	2.0	0.996	**0.245**
NLPGroup-IISERB 4 (UMLS+Entropy +AB)	**1.000**	0.074	0.138	0.038	0.037	41.5	0.843	0.116

† This model is trained using two Reddit data sets collected from two other resources

RF outperforms the other classifiers following the TF-IDF based term weighting scheme for BOW features in terms of F1 score. AB classifier beats other classifiers in terms of F1 score using the UMLS features following Entropy-based weighing scheme. The Longformer model performs better than the other transformer-based models based on the F1 score. Following their performance on the training corpus in terms of F1 score, these five frameworks had been implemented on the test corpus. For Doc2Vec based features, LR and AB classifiers beat the other classifiers based on F1 score, but these scores do not belong to the top five F1 scores in Table 1 and hence these models were not implemented on the test set.

The decision based results of the five runs on the test corpus in terms of precision, recall, F1 score, $ERDE_5$ [16], $ERDE_{50}$ [16], $Latency_{TP}$ [16] and speed [16], are reported in Table 2. It can be seen from this table that the NLPGroup-IISERB 4 run achieves the best precision score (1.0) among the precision scores of all 41 submissions for task 1 of the eRisk2022 challenge. The recall scores of NLPGroup-IISERB 1 (1.0), NLPGroup-IISERB 2 (1.0), and NLPGroup-IISERB 3 (1.0) runs are equal, and these are the best recall scores for task 1 among all submissions. The performance of NLPGroup-IISERB 2 run in terms of $latency_{TP}$ (1.0) and speed (1.0) performs better than all submissions for task 1. However, none of our submissions performs reasonably well in terms of F1 score, $ERDE_5$, $ERDE_{50}$, and latency-weighted F1 score.

The ranking based results of the five runs on the test corpus in terms of precision, recall, F1 score, $ERDE_5$ [16], $ERDE_{50}$ [16], $Latency_{TP}$ [16] and speed [16], are reported in Table 3. It can be seen from this table that none of our runs performs reasonably well in terms of all the evaluation metrics. We may consider this as one of the limitations of the proposed models for task 1, and we plan to investigate them further in future.

3.3.2. Task 2: Early Detection of Depression

We have implemented four different classifiers using three different feature engineering techniques individually on the training corpus. The performance of each of these models was

Table 3
Task1: Ranking Based Results on Test Set

Writings	Metrics	NLPGroup-IISERB0 (BOW+ TF-IDF+RF)	NLPGroup-IISERB1 (BOW+ Entropy+RF)	NLPGroup-IISERB2 (BOW+ Entropy +RF†)	NLPGroup-IISERB3 (Longformer)	NLPGroup-IISERB4 (UMLS+ Entropy +AB)
1	P@10	0.00	0.00	0.00	0.00	0.20
	NDCG@10	0.00	0.00	0.00	0.00	0.38
	NDCG@100	0.02	0.03	0.15	0.01	0.15
100	P@10	0.00	0.00	0.00	0.10	0.00
	NDCG@10	0.00	0.00	0.00	0.06	0.00
	NDCG@100	0.03	0.03	0.11	0.10	0.06
500	P@10	0.00	0.00	0.20	0.10	0.00
	NDCG@10	0.00	0.00	0.13	0.07	0.00
	NDCG@100	0.03	0.05	0.12	0.12	0.07
1000	P@10	0.00	0.00	0.00	0.10	0.00
	NDCG@10	0.00	0.00	0.00	0.07	0.00
	NDCG@100	0.03	0.03	0.08	0.12	0.07

† This model is trained using two Reddit data sets collected from two other resources

Table 4
Task2: Performance of Different Frameworks on the Training Corpus

Feature Types	Classifier	Precision	Recall	F1 Score
Entropy Based Features (Using given training data)	AdaBoost	0.59	0.7	**0.64**
	Logistic Regression	0.59	0.68	0.63
	Random Forest	0.66	0.57	0.61
	Support Vector Machine	0.65	0.68	**0.67**
TF-IDF Based Features (Using given training data)	AdaBoost	0.55	0.55	0.55
	Logistic Regression	0.47	0.67	0.55
	Random Forest	0.69	0.56	0.62
	Support Vector Machine	0.59	0.64	**0.62**
Doc2Vec Based Features (Using given training data)	Logistic Regression	0.38	0.82	0.52
	Random Forest	0.63	0.48	**0.54**
	Support Vector Machine	0.30	0.91	0.45

reported in Table 4 in terms of precision, recall and F1 score. These results were used to analyze the performance of the proposed models on the training set. Subsequently, the best five models from Table 4 in terms of F1-score had been selected and then implemented on the given test corpus. Finally, the performances of these five models on the test corpus were communicated as official results of our team.

It can be seen from Table 4 that Entropy-based BOW features yielded better results than TF-IDF and Doc2Vec based features for all the classifiers. The AB and SVM classifiers for entropy based BOW features performed better than all the other models in terms of F1 score. Table 4 shows that the performance of RF classifier using Entropy-based feature engineering scheme is reasonably well. Hence we have selected these three models to be implemented on the test

Table 5
Task2: Decision-Based Results obtained on the test set

Runs	P	R	F1	$ERDE_5$	$ERDE_{50}$	$latency_{TP}$	speed	latency weighted F1
NLPGroup-IISERB 0 (BOW+Entropy+SVM)	**0.682**	0.745	**0.712**	0.055	0.032	9.0	0.969	**0.690**
NLPGroup-IISERB 1 (BOW+TF-IDF+SVM)	0.385	0.857	0.532	0.062	0.032	18.0	0.934	0.496
NLPGroup-IISERB 2 (BOW+Entropy+RF)	0.662	0.459	0.542	0.069	0.058	62.0	0.766	0.416
NLPGroup-IISERB 3 (Doc2Vec+RF)	0.653	0.500	0.566	0.067	0.046	26.0	0.903	0.511
NLPGroup-IISERB 4 (BOW+Entropy+AB)	0.000	0.000	0.000	0.070	0.070	-	-	-

data. It may be noted from Table 4 that the LR classifier performed better than the RF classifier using BOW features following Entropy-based term weighting scheme in terms of F1 score. However, we did not select it to implement on the test corpus as LR often performs the same as of SVM. We also selected the best models of the TF-IDF based feature engineering scheme and the Doc2Vec based model to implement on the test data. Thus we have submitted a total of five runs using the test data for evaluation. Note that for Doc2Vec based model we could not implement the AB classifier within the deadline and hence this result is not reported in Table 4. Moreover, we could not implement the transformer based models for this task due to the limitation of time.

The decision based results of the five runs on the test corpus in terms of precision, recall, F1 score, $ERDE_5$ [16], $ERDE_{50}$ [16], $Latency_{TP}$ [16] and speed [16], are reported in Table 5. It may be noted that the NLPGroup-IISERB 0 run performed best in terms of F1 score (0.712) and latency weighted F1 score (0.690) among all the 62 runs submitted for task 2. Moreover, the performance of NLPGroup-IISERB 3 run performed second best in terms of F1 score (0.566) among all other runs. The precision scores of NLPGroup-IISERB 0 and NLPGroup-IISERB 2 runs respectively were the second (0.682) and third best (0.662) among all the submissions. The proposed models performed reasonably well in terms of other evaluation metrics for task 2, but could achieve a place in the top three positions. These results indicate the effectiveness of the proposed models.

Ranking based evaluation ranks the users in decreasing estimation of risk with the help of standard IR metrics, such as P@10 or Normalized Discounted Cumulative Gain (NDCG) [16]. Table 6 shows that the scores are not reasonably well for the first writing, except for the NLPGroup-IISERB 2 run. However, considering 100 writings, NLPGroup-IISERB 0, NLPGroup-IISERB 1 and NLPGroup-IISERB 4 runs outperform all other submissions in terms P@10 metric for task2. NLPGroup-IISERB 0 and NLPGroup-IISERB 4 runs performed the second best among all other runs in terms of NDCG@10 score, while the NLPGroup-IISERB 4 run performed second best among all submissions in terms of NDCG@100 score. For 500 writings, NLPGroup-IISERB 0 and NLPGroup-IISERB 4 runs perform better than all other submissions of the challenge in terms of P@10 and NDCG@10 metrics. Moreover, NLPGroup-IISERB 4 run achieves the second

Table 6
Task2: Performance of Different Frameworks on the ranking based evaluation on Test Set

Writings	Metrics	NLPGroup-IISERB0 (BOW+ Entropy+SVM)	NLPGroup-IISERB1 (BOW+ TF-IDF+SVM)	NLPGroup-IISERB2 (BOW+ Entropy+RF)	NLPGroup-IISERB3 (Doc2Vec+ RF)	NLPGroup-IISERB4 (BOW+ Entropy +AB)
1	P@10	0.00	0.30	0.70	0.00	0.00
	NDCG@10	0.00	0.32	0.79	0.00	0.00
	NDCG@100	0.02	0.13	0.24	0.06	0.04
100	P@10	**0.90**	**0.90**	0.00	0.10	**0.90**
	NDCG@10	0.92	0.81	0.00	0.19	0.93
	NDCG@100	0.30	0.27	0.00	0.06	0.66
500	P@10	**0.90**	0.80	0.00	0.00	**0.90**
	NDCG@10	**0.92**	0.84	0.00	0.00	**0.92**
	NDCG@100	0.33	0.33	0.00	0.02	0.69

Table 7
Task3: Performance of the proposed frameworks on the test set

Runs	MZOE	MAE	MAE_{macro}	GED_5	RS	ECS	SCS	WCS
NLPGroup-IISERB 1 (Tokenised Text+Jaccard Similarity)	0.92	2.58	2.09	2.04	2.16	1.89	2.74	2.33
NLPGroup-IISERB 2 (BERT+Cosine Similarity)	0.92	**2.18**	**1.76**	**1.74**	**2.00**	**1.73**	**2.03**	**1.92**
NLPGroup-IISERB 3 (BERT+Cosine Similarity)⋆	0.93	2.60	2.10	2.04	2.13	1.90	2.74	2.35
NLPGroup-IISERB 4 (Tokenised Text+Jaro-Winkler Distance)	**0.81**	3.36	2.96	3.68	3.69	3.18	4.28	3.82

⋆ This model is pre-trained using anorexia dataset from eRisk shared task 2 [27]

best score in terms of NDCG@100 score among all the runs. We could not submit the results for 1000 writings for task 2 within the given deadline and hence we could not achieve any score for the same.

It may be noted from Table 4 and Table 5 that the proposed frameworks using SVM classifier have high recall scores, but random forest classifier based models achieved high precision scores. Moreover, the SVM classifier using the BOW features following entropy based term weighting scheme consistently performed the best in terms of most of the decision based and ranking based evaluation metrics. Hence we may conclude that proposed model using entropy based BOW features and SVM classifier is an effective and robust model for early prediction of depression over social media.

3.3.3. Task 3: Measuring the severity of the signs of Eating Disorders

The performance of the four runs on the test corpus in terms of different evaluation measures [16] as described in section 3.2 are reported in Table 7. It can be seen from this table that the NLPGroup-IISERB 2 run, which is a combination of cosine similarity and BERT model fine-

tuned on anorexia dataset from eRisk 2018 shared task 2 [27] performed the best among all the other runs for task 3 in terms of all the evaluation metrics except MZOE metric. The proposed models performed well in terms of GED score indicate that they identify eating disorder and its side-effects reasonably well. The reason is that GED is indicative of the overall score of the 4 metrics RS, ECS, SCS and WCS respectively, which relate to restraint, eating, shape and weight concerns that are further associated with psychological effects of eating disorder. Moreover, NLPGroup-IISERB 1 and NLPGroup-IISERB 3 runs, respectively, achieved the second best and third best scores among all the other submissions for task 2 in terms of all metrics except the MZOE metric. Being unsupervised in nature, the proposed models for task 3 performed reasonably well in measuring the severity of eating disorders. These results indicate the value and validity of the proposed models for task 3.

4. Conclusion

The eRisk 2022 shared task highlights various challenges for early detection of depression and pathological gambling using the data of different users over Reddit for a given time period. We have proposed various text mining frameworks using different features from the given corpora to accomplish the given tasks. It has been observed from the empirical analysis that the classical BOW model performs better than all the deep learning-based models on the given data except the longformer model. Note that the embeddings were generated following the Doc2Vec model and transformer-based architecture using the given training corpus of the individual tasks, which have a reasonably low number of documents compared to the other pre-trained deep learning-based embeddings e.g., fasttext, which were trained on huge text collections. Consequently, these deep learning models cannot correctly represent the semantic interpretations of the given documents, and hence their performances are not as good as the classical BOW model. The Longformer model performed as good as the BOW model for Task1, but we could not explore its performance for task2 owing to time limitations. In the future, we plan to build a large training corpus by collecting data from Reddit and similar forums for early prediction of risks of different mental illnesses to develop pretrained longformer based embeddings to further improve the performance.

Acknowledgements

Tanmay Basu acknowledges the support of the seed funding (PPW/R&D/2010006) provided by Indian Institute of Science Education and Research Bhopal, India.

References

[1] F. Charlson, M. van Ommeren, A. Flaxman, J. Cornett, H. Whiteford, S. Saxena, New WHO prevalence estimates of mental disorders in conflict settings: a systematic review and meta-analysis, Lancet 394 (2019) 240–248.

[2] J. Sareen, T. O. Afifi, K. A. McMillan, G. J. Asmundson, Relationship between household income and mental disorders: findings from a population-based longitudinal study, Arch Gen Psychiatry 68 (2011) 419–427.
[3] M. De Choudhury, M. Gamon, S. Counts, E. Horvitz, Predicting depression via social media (2013) 1–10.
[4] M. De Choudhury, S. Counts, E. Horvitz, Social media as a measurement tool of depression in populations (2013) 47–56.
[5] C. D. Manning, P. Raghavan, H. Schutze, Introduction to information retrieval (2008).
[6] Q. Le, T. Mikolov, Distributed representations of sentences and documents (2014) 1188–1196.
[7] T. Sabbah, A. Selamat, M. H. Selamat, F. S. Al-Anzi, E. H. Viedma, O. Krejcar, H. Fujita, Modified frequency-based term weighting schemes for text classification, Applied Soft Computing 58 (2017) 193–206.
[8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).
[9] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, Biobert: a pre-trained biomedical language representation model for biomedical text mining, Bioinformatics 36 (2020) 1234–1240.
[10] 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).
[11] I. Beltagy, M. E. Peters, A. Cohan, Longformer: The long-document transformer, arXiv preprint arXiv:2004.05150 (2020).
[12] Y. Freund, R. Schapire, N. Abe, A short introduction to boosting, Journal-Japanese Society For Artificial Intelligence 14 (1999) 1612.
[13] A. Genkin, D. D. Lewis, D. Madigan, Large-scale bayesian logistic regression for text categorization, Technometrics 49 (2007) 291–304.
[14] B. Xu, X. Guo, Y. Ye, J. Cheng, An improved random forest classifier for text categorization., JCP 7 (2012) 2913–2920.
[15] S. Tong, D. Koller, Support vector machine active learning with applications to text classification, Journal of machine learning research 2 (2001) 45–66.
[16] J. Parapar, P. Martin-Rodilla, D. E. Losada, F. Crestani, Overview of erisk 2022: Early risk prediction on the internet (2022).
[17] W. Cohen, P. Ravikumar, S. Fienberg, A comparison of string metrics for matching names and records, in: Proceedings of Kdd workshop on data cleaning and object consolidation, volume 3, 2003, pp. 73–78.
[18] A. Huang, et al., Similarity measures for text document clustering 4 (2008) 9–56.
[19] T. Basu, C. Murthy, A supervised term selection technique for effective text categorization, International Journal of Machine Learning and Cybernetics 7 (2016) 877–892.
[20] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality (2013) 3111–3119.
[21] O. Bodenreider, The unified medical language system (umls): integrating biomedical terminology, Nucleic acids research 32 (2004) D267–D270.
[22] A. R. Aronson, F. M. Lang, An overview of metamap: historical perspective and recent

advances, Journal of the American Medical Informatics Association 17 (2010) 229–236.
[23] R. Martínez-Castaño, A. Htait, L. Azzopardi, Y. Moshfeghi, Early risk detection of self-harm and depression severity using bert-based transformers: ilab at clef erisk 2020, Early Risk Prediction on the Internet (2020).
[24] T. Basu, S. Kumar, A. Kalyan, P. Jayaswal, P. Goyal, S. Pettifer, S. R. Jonnalagadda, A novel framework to expedite systematic reviews by automatically building information extraction training corpora, arXiv preprint arXiv:1606.06424 (2016).
[25] S. Chattopadhyay, T. Basu, A. K. Das, K. Ghosh, L. C. Murthy, Towards effective discovery of natural communities in complex networks and implications in e-commerce, Electronic Commerce Research 21 (2021) 917–954.
[26] S. Cahyono, Comparison of document similarity measurements in scientific writing using jaro-winkler distance method and paragraph vector method 662 (2019).
[27] S. Paul, S. K. Jandhyala, T. Basu, Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. (2018).
[28] T. Basu, S. Goldsworthy, G. V. Gkoutos, A sentence classification framework to identify geometric errors in radiation therapy from relevant literature, Information 12 (2021) 139.
[29] D. E. Losada, F. Crestani, A test collection for research on depression and language use (2016) 28–39.
[30] D. E. Losada, P. Martin-Rodilla, F. Crestani, J. Parapar, Overview of erisk 2021: Early risk prediction on the internet (2021).