Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks

Sayanta Paul, Jandhyala Sree Kalyani* and Tanmay Basu**

Ramakrishna Mission Vivekananda Educational and Research Institute
Belur Math, Howrah, West Bengal, India
(sayanta95, sree.kalyani95, welcometanmay)@gmail.com

Abstract. The CLEF eRisk 2018 challenge focuses on early detection of signs of depression or anorexia using posts or comments over social media. The eRisk lab has organized two tasks this year and released two different corpora for the individual tasks. The corpora are developed using the posts and comments over Reddit, a popular social media. The machine learning group at Ramakrishna Mission Vivekananda Educational and Research Institute (RKMVERI), India has participated in this challenge and individually submitted five results to accomplish the objectives of these two tasks. The paper presents different machine learning techniques and analyze their performance for early risk prediction of anorexia or depression. The techniques involve various classifiers and feature engineering schemes. The simple bag of words model has been used to perform ada boost, random forest, logistic regression and support vector machine classifiers to identify documents related to anorexia or depression in the individual corpora. We have also extracted the terms related to anorexia or depression using metamap, a tool to extract biomedical concepts. Theerefore, the classifiers have been implemented using bag of words features and metamap features individually and subsequently combining these features. The performance of the recurrent neural network is also reported using GloVe and Fasttext word embeddings. Glove and Fasttext are pre-trained word vectors developed using specific corpora e.g., Wikipedia. The experimental analysis on the training set shows that the ada boost classifier using bag of words model outperforms the other methods for task1 and it achieves best score on the test set in terms of precision over all the runs in the challenge. Support vector machine classifier using bag of words model outperforms the other methods in terms of fmeasure for task2. The results on the test set submitted to the challenge suggest that these framework achieve reasonably good performance.

Keywords: identification of depression, anorexia detection, text classification, information extraction, machine learning

* Sayanta Paul and Jandhyala Sree Kalyani have equal contribution in this work
** Corresponding author

1 Introduction

Early risk prediction is a new research area potentially applicable to a wide variety of situations such as identifying people with mental illness over social media. Online social platforms allow people to share and express their thoughts and feelings freely and publicly with other people [1]. The information available over social media is a rich source for sentiment analysis or inferring mental health issues [2]. The CLEF eRisk 2018 challenge focuses on early prediction of risks related to mental disorder using the social media. The main goal of eRisk 2018 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 related to the creation of test collections for early detection of depression [3]. It has organized two tasks this year and released two different corpora for the individual tasks and these corpora are developed using the posts and comments over Reddit, a popular social media [3]. The first task is early risk prediction of depression using the posts and comments on Reddit. The other task is a pilot task and the aim of the task is to identify the signs of anorexia using the given corpus of comments and posts over Reddit.

Depression is a common illness that negatively affects feelings, thoughts and behaviors and can harm regular activities like sleeping. It is a leading cause of disability and many other diseases [1]. According to WHO (World Health Organization)[1] statistics, more than 300 million people over the world are affected in depression and in each country at least 10% are provided treatment. Poor recognition and treatment of depression may aggravate heart failure symptoms, precipitate functional decline, disrupt social and occupational functioning, and lead to an increased risk of mortality [4]. Early detection of depression is thus necessary. Unfortunately the rates of detecting and treating depression among those with medical illness are quite low [5]. To be diagnosed with depression, there must be proper resources to detect depression. Many research works have been done in the last few years to examine the potential of social media as a tool for early detection of depression or mental illness [1, 6, 7]. The first task of this challenge is mainly concerned about evaluating the performance of different machine learning frameworks for potential information extraction from the given corpus of Reddit posts regarding the symptoms of depression [3]. A set of posts over Reditt of a particular person is considered as a single document. The corpus is divided into training and test set. The training set is further divided into two categories i.e., depression and control group i.e., non-depression. Therefore 10 chunks of the test set were released over ten weeks with each chunk per week. Each test chunk contains the posts of a particular person. The task is to identify whether the posts of a particular person in a chunk belong to depression category.

Anorexia is a serious psychiatric disorder distinguished by a refusal to maintain a minimally normal body weight, intense fear of weight gain, and disturbances

[1] www.who.int/mental_health/management/depression/en/

in the perception of body shape and weight [8]. Anorexia has severe physical side effects and may be associated with disturbances in multiple organ systems [8]. According to National Eating Disorder Association, USA, 70 million people of all ages suffer from anorexia[2]. A survey of WHO considers severe anorexia as one of the most burdensome diseases in the world [9]. Moreover, anorexia can adversely affect chronic health conditions, such as cardiovascular disease, cancer, diabetes and obesity. An individual suffering from anorexia may reveal one or several signs such as rapidly losing weight or being significantly thin, depressed or lethargic and so on [8]. The motivation behind the second task is that if anorexic symptoms are properly identified on time, then, professionals could intervene before anorexia progresses. The objective of the second task is to develop effective machine learning frameworks to detect the signs of anorexia using the given corpus. The corpus is divided into training and test set. The training set is divided into two categories - anorexia, and non-anorexia i.e., control group [3]. The task consists of identifying whether the posts of a particular person in the test set belong to the anorexia category.

In this paper, different machine learning frameworks have been proposed to accomplish the given tasks. The aim is to train a machine learning classifier using the training set to identify anorexia or depression of the individual documents of the test sets of these tasks. The performance of a text classification technique is highly dependent on the potential features of a corpus. Therefore the performance of different classifiers have been tested using both text features and biomedical features extracted from the given corpus. In general, each unique term of a corpus is considered as a feature and therefore the frequency of the individual terms are considered to form the document vectors [10]. This is known as bag of words (BOW) model. However, the term document matrix of a corpus becomes sparse and high dimensional following the BOW model. The same may deviate the performance of the classifiers. Hence we have used MetaMap[3], a tool to extract UMLS concepts in free text [11]. UMLS stands for Unified Medical Language System and it can identify semantic types of a term in free text that belong to different pre-defined biomedical categories [12]. Here we have considered only those terms that belong to the semantic categories related to depression or anorexia depending upon the tasks. We have implemented Metamap for individual corpora of the given tasks and extracted the UMLS features. Subsequently, ada boost [13], logistic regression [14], random forest [15], support vector machine [16] classifiers have been implemented using only BOW features, only UMLS features and combining BOW and UMLS features to categorize the documents of the test set of individual tasks. Moreover, for the first task recurrent neural network is implemented using fasttext, a pretrained word vectors developed over crawling the web [17, 18]. For the second task, the recurrent neural network is implemented using GloVe [19], a pretrained word vectors developed using a Wikipedia and a Twitter corpus.

[2] https://www.nationaleatingdisorders.org/CollegiateSurveyProject
[3] https://metamap.nlm.nih.gov

The empirical results for the first task demonstrate that the ada boost, random forest and support vector machine classifiers using BOW features outperform the other frameworks using UMLS features and combining BOW and UMLS features. Furthermore, ada boost classifier using BOW features outperforms the other methods and it achieves best score on the test set in terms of precision over all the submissions in the eRisk 2018 challenge. For the second task, the experimental results show that the support vector machine classifier using BOW features outperforms the other frameworks using both UMLS features and combining BOW and UMLS features. The results on the test set submitted to the challenge suggest that these frameworks for task2 achieve reasonably good performance. However, there are some submissions for this pilot task, which beat the performance of this framework.

The paper is organized as follows. The proposed machine learning frameworks are explained in section 2. Section 3 describes the experimental evaluation. The conclusion is presented in section 4.

2 Proposed Methodologies

Various machine learning techniques have been proposed here to identify the documents related to anorexia from the given corpus, which is released in XML format. Each XML document contains the posts or comments of a Reddit user over a period of time with the corresponding dates and titles. We have extracted the posts or comments from the XML documents and ignored the other entries. Therefore the corpus used for experiments in this article contain only the free texts related to different posts over Reddit for individual users. Different types of features are considered to build the proposed frameworks to identify anorexia or depression of the individual documents using state of the art classifiers.

2.1 Feature Engineering Techniques

Different feature engineering techniques exist in the literature of text mining. We have considered both raw text features and semantic features in the proposed methods.

2.1.1 Bag Of Words (BOW) Features

The text documents are generally represented by the bag of words (BOW) model [20][10]. 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 [21].

Let us denote the number of documents of the corpus and the number of terms of the vocabulary by N and n respectively. The number of times the i^{th} term t_i occurs in the j^{th} document is denoted by $tf_{ij},\ i = 1, 2, ..., n;\ j = 1, 2, ..., N$.

Document frequency df_i is the number of documents in which a particular term appears. Inverse document frequency determines how frequently a term occurs in a corpus and it is defined as $idf_i = log(\frac{N}{df_i})$. The weight of the i^{th} term in the j^{th} document, denoted by w_{ij}, is determined by combining the term frequency with the inverse document frequency as follows:

$$w_{ij} = tf_{ij} \times idf_i = tf_{ij} \times log(\frac{N}{df_i}), \quad \forall\, i = 1, 2, ..., n \text{ and } \forall\, j = 1, 2, ..., N$$

This weighting scheme is known as *tf-idf* weighting scheme. The documents can be efficiently represented using the vector space model in most of the text mining algorithms [22]. In this model each document d_j is considered to be a vector $\boldsymbol{d_j}$, where the i^{th} component of the vector is w_{ij}, i.e., $\boldsymbol{d_j} = (w_{1j}, w_{2j}, ..., w_{nj})$. The document vectors are often sparse as most of the terms do not occur in a particular document and the vectors are also high dimensional. However, this *tf-idf* weighting scheme is used to represent document vectors throughout this paper.

2.1.2 UMLS Features

We have also considered the UMLS concepts extracted from the text as features. The UMLS stands for Unified Medical Language System and it is a comprehensive list of biomedical terms for developing automated systems capable of understanding the specialized vocabulary used in biomedicine and health care [23]. In UMLS there are 133[4] semantic categories related to biomedicine and health. The semantic category of a term can be identified using MetaMap[5], a tool to recognize UMLS concepts in free-text [24]. MetaMap first breaks the text into phrases and then for each phrase it returns different semantic categories of a term 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 [11]. These CUIs are considered as features and they are called as UMLS features in this article.

For the first task we have retained only those terms related to some manually selected semantic categories related to depression, namely, mental health and behavioral dysfunctions, abnormalities, diagnostic procedures, signs and symptoms, and findings. For the second task, the terms belonging to the UMLS concepts, namely, Protein, Activity, Disease, Food, Individual Behavior, Social Behavior, and Vitamin are considered in the experiments as the other semantic categories in UMLS are not related to eating habits or eating disorders.

MetaMap also normalizes the identified concepts of a term and provides a concept unique identifier (CUI) for each of the concepts [11]. We have generated features corresponding to the CUIs and these features are called as UMLS features throughout this paper.

[4] https://mmtx.nlm.nih.gov/MMTx/semanticTypes.shtml
[5] https://metamap.nlm.nih.gov

2.2 Text Classification Techniques

Different text classification methods have been implemented to identify depression or anorexia in the given corpus using the BOW features and UMLS features individually and by combining them. The proposed frameworks are developed using ada boost, logistic regression (LR), Random Forest (RF), Support Vector Machine (SVM) and recurrent neural network (RNN) classifiers.

SVM is widely used for text categorization [16]. The linear kernel is recommended for text categorization as the linear kernel performs nicely when there is a lot of features [25]. Hence linear SVM is used in the experiments.

Random Forest is an ensemble of decision tree classifiers, which is trained with the bagging method. The general idea of the bagging method is that a combination of learning models increases the overall result. It has shown good results for two class text classification problems [15]. We have used random forest classifier using Gini index as the measure of the quality of a split.

Logistic regression performs well for binary class classification problem [14]. We have implemented logistic regression using liblinear, a library for large scale linear classification [25].

The Ada boost algorithm is an ensemble technique, which can combine many weak classifiers into one strong classifier [13]. This has been widely used for binary class classification problems [26].

RNN is an useful classifier for sequential data because each neuron or unit can use its internal memory to maintain information about the previous input. This allows the network to gain a deeper understanding of the statement. In principle, RNN can handle context from the beginning of the sentence which will allow more accurate predictions of a word at the end of a sentence [17]. For the first task, RNN is implemented using Fasttext embeddings, a pre-trained word vector on 600 billion tokens, 2 million vocabulary and 300 dimensional vectors generated from a corpus of Wikipedia [27]. For task 2, RNN is implemented using GloVe embeddings, a pre-trained word embeddings on 840 billion tokens, 2.2 million vocabulary and 300 dimensional vectors generated from a corpus of Wikipedia and Twitter [19].

3 Experimental Evaluation

3.1 Description of Data

3.1.1 Task1

The corpus released as part of the first task is a collection of posts or comments from a set of users over Reddit [3]. The corpus is divided into two categories - the posts of the users who are suffering from depression, and the posts of the

other users belong to the control group or non-depression category i.e., the users who are not diagnosed with depression [3]. For each user, the collection contains a sequence of writings in chronological order. For each user, the collection of writings has been divided into 10 chunks. The first chunk contains the oldest 10% of the posts, the second chunk contains the second oldest 10% posts, and so forth [3]. The overview of the corpus is presented in Table 1. As the corpus

Table 1. Overview of the Corpus for Task1

	Training Set		Test Set	
	Depressed	Control Group	Depressed	Control Group
No. of subjects	135	72	79	741
No. of submissions (posts and comments)	49,557	481,837	40,665	504,523
Average no. of submissions per subject	367.1	640.7	514.7	680.9
Average no. of days from first to last submission	586.43	625.0	786.9	702.5
Average no. of words per submission	27.4	21.8	27.6	23.7

consists of posts and comments over Reddit, we cannot rule out the possibility of having some individuals who are suffering from depression in the control group (non-depression), and vice-versa. The fundamental issue is how to determine a set of posts that indicates depression. Hence it is necessary to have adequate knowledge about the corpus. The corpus contains 1,076,582 posts or comments from 1027 unique users, of which the posts of 486 users are considered as training set, and rest 820 are used as test set. The most important factor is that the data is unbalanced.

3.1.2 Task2

The corpus released as part of task2 is also a collection of posts or comments from a set of users over Reddit [3]. The data is different from the data of task1, however, both of the corpora are generated from Reddit posts. This corpus is also divided into two categories - the posts of the users who are suffering from anorexia, and the posts of the other users belong to the control group or non-anorexia category i.e., the users who are not diagnosed with anorexia [3]. The corpus contains a series of posts in sequential manner for each user and it is divided into 10 chunks for each user. The first chunk contains the oldest 10% of the posts, the second chunk contains the second oldest 10% posts and so on [3]. The overview of the corpus is presented in Table 2. The corpus contains 2,53,752 posts or comments from 472 unique users, of which the posts of 152 users are considered as training set, and rest 320 are used as test set. This indicates that the corpus is unbalanced. The objective is to identify the posts in the test set that belong to anorexia category.

Table 2. Overview of the Corpus for Task2

	Training Set		Test Set	
	Anorexic	Control Group	Anorexic	Control Group
No. of subjects	20	132	41	279
No. of submissions (posts and comments)	7452	77,514	17,422	151,364
Avgerage no. of submissions per subject	372.6	587.2	424.9	542.5
Avgerage no. of days from first to last submission	803.3	641.5	798.9	670.6
Avgerage no. of words per submission	41.2	20.9	35.7	20.9

3.2 Experimental Setup

The term-document matrices are generally sparse and high dimensional. The same may have adverse affect on the quality of the classifiers. Hence the significant terms related to different categories of a corpus is to be determined. Many term selection techniques are available in the literature. The term selection methods rank the terms in the vocabulary according to different criterion function and then a fixed number of top terms forms the resultant set of features. A widely used term selection technique is χ^2-statistic [10] and this is used in the experiments. We have considered different number of top terms generated by χ^2-statistic and evaluated the performance of different classifiers using these set of terms from the training set. Eventually we have considered the best feature subset for individual classifiers.

Ada boost, LR, RF and SVM classifiers are implemented in Scikit-learn[6], a machine learning tool in Python [28]. RNN is implemented in Keras[7], a deep learning tool in Python. The other experimental settings for the individual tasks are mentioned below.

3.2.1 Task1

The data of the same challenge in 2017 has been released as the training set for this task. The corpus of the 2017 challenge was divided into training set and test set. The ground truths were available for both training and test set. We have used this training set to train different classifiers of the proposed frameworks in this article. The parameters of different classifiers are tuned using 10-fold cross validation technique on this training set of 2017 challenge. The test data of 2017 challenge is used as the validation set to evaluate the performance of the classifiers of the proposed frameworks using the ground truths. The classifiers using a particular type of features that perform the best on the validation set are chosen for implementation on the test set of this year. Subsequently, the results of the proposed frameworks on this test set have been submitted to the eRisk 2018 challenge.

[6] http://scikit-learn.org/stable/supervised_learning.html
[7] https://keras.io

3.2.2 Task2

The given training set of second task is further divided into two parts namely, training set and validation set. The new training set is build by randomly choosing 80% documents individually from anorexia and non-anorexia categories. Similarly the rest 20% of these categories form the validation set. The parameters of different classifiers are tuned using 10-fold cross validation technique on the newly formed training set and therefore the performance of these classifiers are tested on the validation set. The classifiers using a particular type of features that had shown better results than other such frameworks are submitted to the challenge.

3.3 Evaluation Measures

The performance of the proposed method and the state of the art classifiers are evaluated by using the standard precision, recall and fmeasure and ERDE. The precision and recall for two class classification problem can be computed as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP+FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP+FN}}$$

Here TP stands for *true positive* and it counts the number of data points correctly predicted to the positive class. FP stands for *false positive* and it counts the number of data points that actually belong to the negative class, but predicted as positive (i.e., *falsely predicted as positive*). FN stands for *false negative* and it counts the number of data points that actually belong to the positive class, but predicted as negative (i.e., *falsely predicted as negative*). TN stands for *true negative* and it counts the number of data points correctly predicted to the negative class. The fmeasure combines recall and precision with an equal weight in the following form:

$$\text{Fmeasure} = \frac{2 \times \text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

The closer the values of precision and recall, the higher is the fmeasure. Fmeasure becomes 1 when the values of precision and recall are 1 and it becomes 0 when precision is 0, or recall is 0, or both are 0. Thus fmeasure lies between 0 and 1 [29]. A high fmeasure value is desirable for good classification [29].

The organizers of this challenge introduced early risk detection error (ERDE), which checks the correctness of the decision made and the delay to make such decision [30]. The delay was measured by counting the number (k) of distinct textual items seen before giving the answer. The threshold of ERDE was set to 5

to 50 posts which was represented by $ERDE_5$ and $ERDE_{50}$. The correctness of each emitting decision and the delay taken by the system to make the decision has to be calculated. The delay is measured here by counting the number (k) of individual documents seen before giving the answer. Another fundamental issue is that, the corpus used in this task is unbalanced. Consider a binary decision d taken by a system with delay k. The prediction d can be either of TP, TN, FP or FN. Given these four cases ERDE [30] can be defined as

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

The values of c_{fp} and c_{fn} depend on the application domain and the implications of FP and FN decisions. The function $lc_o(k)$ is a monotonically increasing function of k, which is parameterized by o. The minimum value of o is considered as 5 and the maximum value as 50. Note that ERDE lies in range [0, 1]. A low value of ERDE is desirable as this is a measure to find error in the system [30].

3.4 Analysis of Results

3.4.1 Task1

We have reported the performance of Ada Boost, LR, RF and SVM classifiers on the validation set using BOW features, UMLS features and the combination of BOW and UMLS features respectively in Table 3, Table 4 and Table 5. Note that the validation set is the test set of the same challenge in 2017. The performance of these classifiers are measured in terms of fmeasure in these tables. These results are useful to analyze the performance of different proposed frameworks. Eventually, the best frameworks have been implemented on the given test set of eRisk 2018 challenge and subsequently the results are communicated.

Table 3 shows that the performance of Ada Boost is better than the other classifiers in terms of precision, recall and fmeasure. It can be seen from Table 4 that the performance of Ada Boost is best among all other classifiers in terms of recall and fmeasure. It is observed from Table 5 that LR, RF outperforms the other classifiers in terms of precision, recall and fmeasure respectively.

It may be noted from Table 3 and Table 4 that the performance of all the classifiers using BOW features are better than the same using UMLS features. Moreover, Table 3 and Table 5 show that all the classifiers using the BOW features perform better than the same using the combination of BOW and UMLS features. This indicates that UMLS features have little influence on the performance of the classifiers. It is manually checked that the number of UMLS features are too small and there are absence of biomedical terms related to depression in the documents. This may be the reason of poor performance. Consequently,

Table 3. Performance of Different Classifiers Using BOW Features

Classifiers	Precision	Recall	Fmeasure
Ada Boost	**0.75**	**0.76**	**0.75**
Logistic Regression	**0.75**	0.73	0.74
Support Vector Machine	0.72	0.71	0.72
Random Forest	0.71	0.74	0.73

Table 4. Performance of Different Classifiers Using UMLS Features

Classifiers	Precision	Recall	Fmeasure
Ada Boost	0.41	**0.50**	**0.45**
Logistic Regression	0.46	0.43	0.37
Support Vector Machine	**0.48**	0.46	0.41
Random Forest	0.46	0.43	0.40

Table 5. Performance of Different Classifiers Using the Combination of BOW and UMLS Features

Classifiers	Precision	Recall	Fmeasure
Ada Boost	0.61	0.62	0.61
Logistic Regression	**0.64**	0.63	**0.63**
Support Vector Machine	0.62	0.63	0.62
Random Forest	0.63	**0.64**	**0.63**

we have submitted the results of Ada Boost, LR, RF and SVM classifiers using BOW features on the test set to the challenge.

We have also submitted a result of RNN using Fasttext embedding, as RNN has been widely used for text categorization in recent years. However, the performance of RNN on the validation set is not as good as the other classifiers using BOW features. The precision, recall and fmeasure of the same is 0.64, 0.60 and 0.62 respectively. Note that we have fixed the sequence length of each sentence considered by RNN as 150 due to the limitation in the resources. The results of RNN may be improved by increasing the sequence length in the model, which is beyond the scope of this article.

The results of Ada Boost, LR, RF and SVM classifiers using BOW features and RNN classifier using Fasttext embedding on the given test set in terms of $ERDE_5$, $ERDE_{50}$, precision, recall and fmeasure are reported in Table 6. RKMVERIA, RKMVERIB, RKMVERIC, RKMVERID indicate the results of LR, SVM, Ada boost, and RF classifiers respectively using BOW features. RKMVERIE indicates the result of RNN classifier using fasttext embedding. Table 6 shows that precision of the RKMVERIC framework is better than the precision of the other RKMVERI frameworks and RKMVERIC achives the best score in terms of the precision of 45 submissions in the eRisk 2018 challenge. It can be seen from Table 6 RKMVERID performs better than other RKMVERI

Table 6. The Performance of Various Classifiers using Different Evaluation Measures

Methods	$ERDE_5$	$ERDE_{50}$	Fmeasure	Precision	Recall
RKMVERIA (LR using BOW)	10.14%	8.68%	0.52	0.49	0.54
RKMVERIB (SVM using BOW)	10.66%	9.07%	0.47	0.37	**0.65**
RKMVERIC (Ada Boost using BOW)	**9.81%**	9.08%	0.48	**0.67**	0.38
RKMVERID (RF using BOW)	9.97%	**8.63%**	**0.58**	0.60	0.56
RKMVERIE (RNN using Fasttext)	9.89%	9.28%	0.21	0.35	0.15

frameworks in terms of fmeasure and the same is the fourth best fmeasure in the competition.

3.4.2 Task2

We have reported the performance of Ada Boost, LR, RF and SVM classifiers on the validation set using BOW features, UMLS features and the combination of BOW and UMLS features respectively in Table 7, Table 8 and Table 9. The performance of these classifiers are measured in terms of fmeasure in these tables.

It can be seen from Table 7 that the performance of SVM is better than the other classifiers in terms of precision recall and fmeasure. Table 8 shows that the performance of SVM is the best among all other classifiers in terms of fmeasure. It can be observed from Table 9 that Ada Boost classifier outperforms other classifiers in terms of fmeasure.

It may be noted from Table 7 and Table 8 that the performance of all the

Table 7. Performance of Different Classifiers Using BOW Features

Text Classifiers	Precision	Recall	Fmeasure
AdaBoost	0.91	0.93	0.91
Logistic Regression	0.96	0.97	0.97
Random Forest	0.98	0.92	0.95
Support Vector Machine	0.97	0.98	**0.98**

Table 8. Performance of Different Classifiers Using UMLS Features

Text Classifiers	Precision	Recall	Fmeasure
Ada Boost	0.54	0.52	0.46
Logistic Regression	0.56	0.51	0.47
Random Forest	0.47	0.49	0.16
Support Vector Machine	0.58	0.49	**0.55**

classifiers using BOW features are better than the same using UMLS features.

Table 9. Performance of Different Classifiers Using the Combination of BOW and UMLS Features

Text Classifiers	Precision	Recall	Fmeasure
Ada Boost	0.43	0.51	**0.47**
Logistic Regression	0.57	0.51	0.14
Random Forest	0.47	0.49	0.16
Support Vector Machine	0.46	0.48	0.16

Table 10. The Performance of Various Classifiers using Different Evaluation Measures

Methods	$ERDE_5$	$ERDE_{50}$	Fmeasure	Precision	Recall
RKMVERIA (SVM using BOW)	**12.17**%	**8.63**%	**0.67**	0.82	**0.56**
RKMVERIB (LR using BOW)	12.93%	12.31%	0.46	0.81	0.32
RKMVERIC (RF using BOW)	12.85%	12.85%	0.25	**0.86**	0.15
RKMVERID (RNN using GloVe)	12.89%	12.89%	0.31	0.80	0.20
RKMVERIE (AdaBoost using BOW)	12.93%	12.31%	0.46	0.81	0.32

Moreover, Table 7 and Table 9 show that all the classifiers using the BOW features perform better than the same using the combination of BOW and UMLS features. This indicates that UMLS features have little influence on the performance of the classifiers. We have manually checked that the number of UMLS features are too small, which may be a reason of poor performance. Consequently, we have submitted the results of Ada Boost, LR, RF and SVM classifiers using BOW features on the test set to the challenge. We have also submitted a result of RNN using GloVe embedding, as RNN has been widely used for text categorization. However, the performance of RNN on the validation set is not as good as the other classifiers using BOW features. The fmeasure of the same is 0.56.

The results of Ada Boost, LR, RF and SVM classifiers using BOW features and RNN classifier using GloVe embedding on the given test set in terms of $ERDE_5$, $ERDE_{50}$, precision, recall and fmeasure are reported in Table 10. RKMVERIA, RKMVERIB, RKMVERIC, RKMVERIE indicate the results of SVM, LR, RF, and Ada Boost classifiers respectively using BOW features. RKMVERID indicates the result of RNN classifier using GloVe embedding. Table 10 shows that precision of the RKMVERIC framework is better than the precision of the other RKMVERI frameworks and it is the fourth best score among the precision of 35 submissions in the eRisk 2018 challenge. RKMVERIA performs better than other RKMVERI frameworks in terms of $ERDE_5$, $ERDE_{50}$, recall and fmeasure.

4 Conclusion

The eRisk 2018 shared task highlights a variety of challenges for early detection of depression and anorexia using the data over social forums. Depression is a type of mental disorder that has adverse affects on feelings, thoughts and behav-

iors and can harm regular activities like sleeping, working etc. Anorexia is also a mental disorder distinguished by a refusal to maintain a normal body weight, intense fear of weight gain and disturbance in the perception of body shape and weight. However, it is generally difficult to identify depression or anorexia from different symptoms. The treatment for these diseases can be started on time, if the alarming symptoms are diagnosed properly. The aim of this challenge is to detect signs of such diseases from the posts or comments of individuals over social media. Various machine learning frameworks have been developed using different types of features from the free text to accomplish this task. We have examined the performance of both bag of words features and UMLS features using different classifiers to identify depression. However, it is observed that a few UMLS features exist in the corpus. Hence the proposed methodologies relied on the BOW features. The experimental results show that the performance of these methodologies are reasonably good. We have also implemented the RNN classifier using the Fasttext and GloVe word embeddings. However, the performance of these RNN models are not so good as we have to fix the sequence length of each sentence as 150 only due to limitation of resources. In future, we can implement RNN using higher length of word embeddings for better performance.

References

1. M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of ICWSM*, 2013, pp. 1–10.
2. M. De Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *Proceedings of the Annual ACM Web Science Conference*, 2013, pp. 47–56.
3. D. E. Losada, F. Crestani, and J. Parapar, "Overview of eRisk – early risk prediction on the internet," in *Proceedings of the Ninth International Conference of the CLEF Association*, Avignon, France, 2018.
4. J. A. Cully, D. E. Jimenez, T. A. Ledoux, and A. Deswal, "Recognition and treatment of depression and anxiety symptoms in heart failure," *Primary Care Companion to the Journal of Clinical Psychiatry*, vol. 11, no. 3, pp. 103–109, 2009.
5. L. E. Egede, "Failure to recognize depression in primary care: issues and challenges," 2007.
6. S. C. Guntuku, D. B. Yaden, M. L. Kern, L. H. Ungar, and J. C. Eichstaedt, "Detecting depression and mental illness on social media: an integrative review," *Current Opinion in Behavioral Sciences*, vol. 18, pp. 43–49, 2017.
7. G. Shen, J. Jia, L. Nie, F. Feng, C. Zhang, T. Hu, T.-S. Chua, and W. Zhu, "Depression detection via harvesting social media: A multimodal dictionary learning solution," in *Proceedings of the International Joint Conference on Artificial Intelligence*, 2017, pp. 3838–3844.
8. E. A. Danila Musante, "Anorexia nervosa: Role of the primary care physician," *JCOM*, vol. 15, no. 9, 2008.
9. B. Duthey, "Priority medicines for europe and the world: A public health approach to innovation," *WHO Background paper*, vol. 6, 2013.
10. T. Basu and C. Murthy, "A supervised term selection technique for effective text categorization," *International Journal of Machine Learning and Cybernetics*, vol. 7, no. 5, pp. 877–892, 2016.

11. A. R. Aronson and F. M. Lang, "An overview of metamap: Historical perspective and recent advances," *Journal of the American Medical Informatics Association*, vol. 17, no. 3, pp. 229–236, 2010.
12. A. T. McCray and S. J. Nelson, "The representation of meaning in the UMLS," *Methods of Information in Medicine*, vol. 34, no. 01/02, pp. 193–201, 1995.
13. Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society for Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999.
14. A. Genkin, D. D. Lewis, and D. Madigan, "Large-scale bayesian logistic regression for text categorization," *Technometrics*, vol. 49, no. 3, pp. 291–304, 2007.
15. B. Xu, X. Guo, Y. Ye, and J. Cheng, "An improved random forest classifier for text categorization." *JCP*, vol. 7, no. 12, pp. 2913–2920, 2012.
16. S. Tong and D. Koller, "Support vector machine active learning with applications to text classification," *Journal of Machine Learning Research*, vol. 2, no. Nov, pp. 45–66, 2001.
17. P. Liu, X. Qiu, and X. Huang, "Recurrent neural network for text classification with multi-task learning," *arXiv preprint arXiv:1605.05101*, 2016.
18. E. Grave, P. Bojanowski, P. Gupta, A. Joulin, and T. Mikolov, "Learning word vectors for 157 languages," *arXiv preprint arXiv:1802.06893*, 2018.
19. J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of EMNLP*, 2014, pp. 1532–1543.
20. C. D. Manning, P. Raghavan, and H. Schutze, *Introduction to Information Retrieval*. Cambridge University Press, New York, 2008.
21. T. Basu and C. A. Murthy, "A similarity based supervised decision rule for qualitative improvement of text categorization," *Fundamenta Informaticae*, vol. 141, no. 4, pp. 275–295, 2015.
22. G. Salton and M. J. McGill, *Introduction to Modern Information Retrieval*. McGraw Hill, 1983.
23. O. Bodenreider, "The unified medical language system (UMLS): Integrating biomedical terminology," *Nucleic Acids Research*, vol. 32, pp. D267–D270, 2004.
24. A. R. Aronson, "Effective mapping of biomedical text to the UMLS metathesaurus: The metamap program," in *Proceedings of AMIA Symposium*, 2001, pp. 17–21.
25. R. E. Fan, K. W. Chang, C. J. Hsieh, X. R. Wang, and C. J. Lin, "Liblinear: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, no. Aug, pp. 1871–1874, 2008.
26. R. E. Schapire, Y. Singer, and A. Singhal, "Boosting and rocchio applied to text filtering," in *Proceedings of SIGIR conference*, 1998, pp. 215–223.
27. T. Mikolov, E. Grave, P. Bojanowski, C. Puhrsch, and A. Joulin, "Advances in pre-training distributed word representations," in *Proceedings of the International Conference on Language Resources and Evaluation*, 2018.
28. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
29. T. Basu and C. A. Murthy, "A feature selection method for improved document classification," in *Proceedings of the International Conference on Advanced Data Mining and Applications*, 2012, pp. 296–305.
30. D. E. Losada and F. Crestani, "A test collection for research on depression and language use," in *International Conference of the Cross Language Evaluation Forum for European Languages*. Springer, 2016, pp. 28–39.