# IISERB@LT-EDI-ACL2022: A Bag of Words and Document Embeddings Based Framework to Identify Severity of Depression Over Social Media

**Tanmay Basu**
Department of Data Science and Engineering
Indian Institute of Science Education and Research Bhopal, India
tanmay@iiserb.ac.in

## Abstract

The DepSign-LT-EDI-ACL2022 shared task focuses on early prediction of severity of depression over social media posts. The BioNLP group at Department of Data Science and Engineering in Indian Institute of Science Education and Research Bhopal (IISERB) has participated in this challenge and submitted three runs based on three different text mining models. The severity of depression were categorized into three classes, viz., no depression, moderate, and severe and the data to build models were released as part of this shared task. The objective of this work is to identify relevant features from the given social media texts for effective text classification. As part of our investigation, we explored features derived from text data using document embeddings technique and simple bag of words model following different weighting schemes. Subsequently, adaptive boosting, logistic regression, random forest and support vector machine (SVM) classifiers were used to identify the scale of depression from the given texts. The experimental analysis on the given validation data show that the SVM classifier using the bag of words model following term frequency and inverse document frequency weighting scheme outperforms the other models for identifying depression. However, this framework could not achieve a place among the top ten runs of the shared task. This paper describes the potential of the proposed framework as well as the possible reasons behind mediocre performance on the given data.

## 1 Introduction

Early prediction of mental illness over social media is a new research area potentially applicable to a wide variety of situations such as identifying people having anxiety and depression over social media (Basu and Gkoutos, 2021). Depression is a common mental illness that involves sadness and lack of interest in day to day activities (Kayalvizhi and Thenmozhi, 2022; Sampath et al., 2022). Poor recognition and late treatment of depression may have serious consequences like heart failure (Cully et al., 2009). Early detection of depression is thus necessary. The information available over social media is a rich source for sentiment analysis or inferring mental health issues (Basu and Gkoutos, 2021). 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 (De Choudhury et al., 2013; Hovy and Spruit, 2016; Benton et al., 2017; Paul et al., 2018; Basu and Gkoutos, 2021).

In the last few years, eRisk group, has organized a series of NLP shared-task for early prediction of different types of mental illnesses (Losada and Crestani, 2016; Losada et al., 2018, 2019, 2020, 2021). As part of these shared tasks, many machine learning based frameworks have been proposed for early prediction of depression using social media posts. Oliveira (Oliveira, 2020) proposed a model using SVM classifier with different types of hand-crafted features (i.e. bag of words, lexicons and behavioural patterns) to estimate the level of depressin using posts over Reddit. Alhuzali et al. used different pre-trained language models and random forest classifier for early detection of depression over Reddit posts (Alhuzali et al., 2021). Guangyao Shen proposed a new multimodal depressive dictionary learning model to detect depressed users on Twitter, and compared their solution with Naive Bayes classifier and multiple social Networking Learning (Shen et al., 2017).

The DepSign-LT-EDI-ACL2022 shared-task is focused on early prediction of severity of depression using Reddit posts, a popular social media (Sampath et al., 2022). The organizers released the Reddit posts of a set of users and the ground truths based on the severity of depression of a portion of the data were also released (Sampath et al., 2022; Kayalvizhi and Thenmozhi, 2022). There are three

*Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 234 - 238
May 27, 2022 ©2022 Association for Computational Linguistics

levels of severity, viz., no depression, moderate, and severe. The data with ground truths were further divided into training and validation set to build the model. The rest of the data without grounds truths were used as test set to evaluate the performance of the model.

We developed a generic text classification framework to identify the severity of depression from the given data without having any additional inputs from the clinical experts. The contributions of this paper are (a) explore the performance of different feature engineering schemes to derive relevant features from given Reddit posts, and (b) presenting a generic text classification framework to generate potential features from the given data to help improve the quality of severity of classification of depression.

## 2 Methodology

The text data contain the chats of different social media users over a period of time. The proposed framework relies on deriving textual features from the given data, and consists of two major steps as described below.

### 2.1 Feature Engineering

We explored different feature engineering schemes to identify relevant features from text data. The classical bag of words model and document embedding based features generated from the given text data were used in the proposed framework.

#### 2.1.1 Bag of Words Model

Initially, the unigrams, bigrams and trigrams were generated following the bag of words (BOW) model. Unigrams, bigrams and trigrams generated from sentences were used as features with the SVM classifier in the experimental analysis. A unigram considers all unique words in a sentence as features (Manning et al., 2008). On the other hand, a bigram or a trigram, considers only two or three consecutive words as a feature respectively (Manning et al., 2008). Both bigrams and trigrams were used in this framework since there are many terms in the training corpus e.g., severe depression, social anxiety, developing drug addiction etc. Such words should be conjoined for better analysis. Subsequently the document vectors were generated based on the following two different term weighting schemes.

1) Term Frequency and Inverse Document Frequency (TF-IDF[1]) (Basu and Murthy, 2016) of the unigrams, bigrams and trigrams generated from the given text data were used as weight of such features. The weight of the $i^{th}$ term in the $j^{th}$ document, denoted by $W_{ij}$, is determined by multiplying the term frequency $(tf_{ij})$ with the inverse document frequency $(idf_i)$ as follows:

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

where n be number of documents, m be the number of terms combining unigrams, bigrams and trigrams in the given training data and $df_i$ is the document frequency i.e., the number of documents where the $i^{th}$ term occurs.

b) Entropy[2] of the term frequency (Basu and Gkoutos, 2021; Sabbah et al., 2017) of individual unigrams, bigrams and trigrams generated from the given training data was also considered as the term weight. In this method, the weight of a term $t_i$ in the $j^{th}$ document, denoted by $W_{ij}$, is determined as follows:

$$W_{ij} = \log(tf_{ij} + 1) \times \left(1 + \frac{\sum_{j=1}^{n} P_{ij} \log P_{ij}}{\log(n+1)}\right),$$

where $P_{ij} = \frac{tf_{ij}}{\sum_{j=1}^{n} tf_{ij}}$

Now it may be noted that the number of BOW features are generally high which makes the document vectors sparse. Therefore chi-square statistics (Basu and Murthy, 2016) were used on the set of BOW features and subsequently the best set of features were extracted by applying a predefined threshold on chi-square statistics score. We had done the experiments with different numbers of this threshold for the chi-square statistics on the training set using 10-fold cross validation technique. The threshold which generates the best performance on the training data was used to run on the given test data.

#### 2.1.2 Document Embeddings

Furthermore, we have generated features using paragraph embeddings technique from the given

[1] Scikit-learn TFIDF Transformer

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

Table 1: Overview of Different Runs

| Runs | Model | # Features |
|---|---|---|
| IISERB 1 | Doc2Vec + RF | 70 |
| IISERB 2 | Entropy Based BOW + LR | 10000 |
| IISERB 3 | TF-IDF Based BOW + SVM | 10000 |

data, which is also known as document embeddings or Doc2Vec model (Le and Mikolov, 2014). It was developed based on unsupervised Continuous Bag of Words (CBOW) and Skip-grams model, which expresses a word as a vector (Mikolov et al., 2013) using a given corpus. Doc2Vec model is an extension of CBOW and Skip-grams model and basically combines them to learn paragraph or document level embeddings (Le and Mikolov, 2014). It is implemented in Gensim[3], a Python library. Here the model was built by training it using the given training corpus and a similar data released as part of the second shared task of eRisk 2021 (Losada et al., 2021). Therefore this model was used to generate the features for individual documents of the given validation and test data. The number of such features was fixed by performing 10-fold cross validation technique on the training data.

### 2.2 Text Classification Framework

Different text classification techniques viz., Adaptive Boosting (AB), Logistic Regression (LR), Random Forest (RF) and SVM were implemented to identify severity of depression in the given data. Each of these classifiers was implemented using three different types of features namely, Entropy based BOW features, TF-IDF based BOW features and Doc2Vec based features.

AB classification algorithm is an ensemble technique, which can combine many weak classifiers into one strong classifier (Freund et al., 1999). Linear SVM is widely used for text classification (Paul et al., 2018). SVM with linear kernel is recommended for text classification as the linear kernel performs well when there is a lot of features (Fan et al., 2008). Hence linear SVM was used in the experiments. RF is a popular classification method based on an ensemble of bootstrapped classification trees (Xu et al., 2012). The multinomial LR algorithm using LibLinear, a library for large-scale linear classification generally perfroms well for data with large features (Genkin et al., 2007).

## 3 Experimental Evaluation

### 3.1 Experimental Setup

We have submitted three runs following three different models. The overview of the runs are given in Table 1. Initially, the combination of different classifiers and feature selection schemes were individually trained on the given training data and their performance were tested on the validation data. Three best models were chosen out of all the models executed on the validation set and these models were implemented on the test data and the results were submitted. The parameters of different classifiers are chosen following 10 fold cross validation method on the training corpus. The performance of these models were evaluated by using macro-averaged precision, recall and f-measure scores (Paul et al., 2018). AB, LR, RF and SVM classifiers were implemented in Scikit-learn[4], a machine learning tool in Python (Basu and Gkoutos, 2021).

### 3.2 Analysis of Results

We used three different types of features to evaluate the performance of four different classifiers on the validation set. The best result of each type of feature engineering scheme and for each classifier is reported in Table 2 in terms of macro-averaged precision, recall and f-measure. These results are useful to analyze the performance of different models. Thereafter, the best classifier for each type of feature in terms of f-measure in Table 2 was selected and subsequently implemented on the given test data. Eventually the performance of these three models on the test data were submitted as official results of our team.

It can be seen from Table 2 that LR performs better than all other classifiers for entropy based BOW features and SVM outperforms the other classifiers for TF-IDF based BOW features in terms of macro-averaged f-measure. Moreover for Doc2Vec based features, RF classifier performs better than all other classifiers. Therefore these three models were chosen based on their performance in Table 2 and ran them on the test corpus. The results of three runs on the test corpus in terms of macro-averaged precision, recall and f-measure are reported in Table 3. It may be noted here that the performance of these three runs are not reasonably well and hence none of these frameworks achieve a place in the top ten

[3] https://radimrehurek.com/gensim/models/doc2vec.html

[4] http://scikit-learn.org/stable/supervised_learning.html

Table 2: Performance of Different Models on the Validation Data

| Feature Type and Classifier | PR* | RL* | FM* |
|---|---|---|---|
| BOW (Entropy) + AB | 0.51 | 0.48 | 0.49 |
| BOW (Entropy) + LR | 0.55 | 0.52 | **0.53** |
| BOW (Entropy) + RF | 0.44 | 0.48 | 0.46 |
| BOW (Entropy) + SVM | 0.54 | 0.51 | 0.52 |
| BOW (TF-IDF) + AB | 0.50 | 0.47 | 0.48 |
| BOW (TF-IDF) + LR | 0.51 | 0.49 | 0.49 |
| BOW (TF-IDF) + RF | 0.46 | 0.51 | 0.48 |
| BOW (TF-IDF) + SVM | 0.54 | 0.50 | **0.51** |
| Doc2Vec + AB | 0.37 | 0.38 | 0.37 |
| Doc2Vec + LR | 0.46 | 0.47 | 0.46 |
| Doc2Vec + RF | 0.48 | 0.47 | **0.47** |
| Doc2Vec + SVM | 0.38 | 0.35 | 0.36 |

*Here PR, RL and FM stands for precision, recall and f-measure. Macro-averaged precision and recall are computed for each model and then the f-measure is computed using these macro-averaged precision and recall scores.

results of the shared-task.

The LR and SVM classifiers using BOW features generally perform well for text data. However, the BOW model can not achieve the semantic interpretation of the words in texts. Hence it could not perform very well for text data of social media posts as they contain irregular texts of diverse meaning. The document embedding model can get rid of this situation and therefore we used the Doc2Vec model to capture the semantic interpretation of the online texts. It can be observed from Table 2 that Doc2Vec based model could not perform well on the test corpus like the BOW model. The deep learning based models works well when trained on large corpora (Basu and Gkoutos, 2021). Here the Doc2Vec based model performs poorly as it was trained on the given training corpus and the corpus released as part of eRisk 2021 shared-task for prediction of self-harm over social media (Losada et al., 2021), which are reasonably small in size.

## 4 Conclusion

The proposed framework relied on extracting different types of relevant text features, including unigrams, bigrams, trigrams and document embeddings to identify the scale of depression. The LR classifier using BOW features following TF-IDF based term weighting scheme achieved the best performance on validation and test data in terms of macro-averaged f-measure. The performance of different models indicate that the combinations of different types of features are important rather than using a single type of feature set. It has been observed from the experimental results that the conventional BOW model performs better than the document embeddings on the test data. Note that we have developed the document embeddings based on the given training corpus, which has reasonably low number of documents in compare to the other pretrained deep learning based word embeddings e.g., Glove, which were trained on huge text collections. As a result the Doc2Vec model cannot properly identify the semantic interpretations of the given data and hence its performance is not as good as the BOW model. In future we plan to develop some pretrained transformer based embeddings for depression and other mental disorders by collecting documents over social media, Wikipedia and other relevant resources to improve the performance of such classification tasks.

Table 3: Performance of Three Runs on the Test Data

| Runs | Precision* | Recall* | F-measure* |
|---|---|---|---|
| IISERB 0 | 0.416 | 0.444 | 0.414 |
| IISERB 1 | **0.430** | 0.465 | 0.437 |
| IISERB 2 | 0.427 | **0.481** | **0.438** |

*Macro-averaged precision and recall are computed for each run and then the f-measure is computed using these macro-averaged precision and recall scores.

## Acknowledgements

This work is done as part of the seed funding (PPW/R&D/2010006) provided by Indian Institute of Science Education and Research Bhopal, India.

## References

Hassan Alhuzali, Tianlin Zhang, and Sophia Ananiadou. 2021. Predicting sign of depression via using frozen pre-trained models and random forest classifier. In *CLEF (Working Notes)*.

Tanmay Basu and Georgios V Gkoutos. 2021. Exploring the performance of baseline text mining frameworks for early prediction of self harm over social media. In *In Proceedings of CLEF Working Notes*. Springer.

Tanmay Basu and CA Murthy. 2016. A supervised term selection technique for effective text categorization. *International Journal of Machine Learning and Cybernetics*, 7(5):877–892.

Adrian Benton, Margaret Mitchell, and Dirk Hovy. 2017. Multi-task learning for mental health using social media text. *arXiv preprint arXiv:1712.03538*.

Jeffrey A Cully, Daniel E Jimenez, Tracey A Ledoux, and Anita Deswal. 2009. Recognition and treatment of depression and anxiety symptoms in heart failure. *Primary care companion to the Journal of clinical psychiatry*, 11(3):103.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. *ICWSM*, 13:1–10.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874.

Yoav Freund, Robert Schapire, and Naoki Abe. 1999. A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780):1612.

Alexander Genkin, David D Lewis, and David Madigan. 2007. Large-scale bayesian logistic regression for text categorization. *Technometrics*, 49(3):291–304.

Dirk Hovy and Shannon L Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598.

S Kayalvizhi and D Thenmozhi. 2022. Data set creation and empirical analysis for detecting signs of depression from social media postings. *arXiv preprint arXiv:2202.03047*.

Q. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of International Conference on Machine Learning*, pages 1188–1196.

David E Losada and Fabio Crestani. 2016. A test collection for research on depression and language use. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 28–39. Springer.

David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk – Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Ninth International Conference of the CLEF Association (CLEF 2018)*, Avignon, France.

David E Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of erisk 2019: Early risk prediction on the internet. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 340–357. Springer.

David E Losada, Fabio Crestani, and Javier Parapar. 2020. Overview of erisk 2020: Early risk prediction on the internet. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*, pages 272–287. Springer.

David E Losada, Patricia Martin-Rodilla, Fabio Crestani, and Javier Parapar. 2021. Overview of erisk 2021: Early risk prediction on the internet. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer.

C. D. Manning, P. Raghavan, and H. Schutze. 2008. *Introduction to Information Retrieval*. Cambridge University Press, New York.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of Advances in Neural Information Processing Systems*, pages 3111–3119.

Luıs Oliveira. 2020. Bioinfo@ uavr at erisk 2020: on the use of psycholinguistics features and machine learning for the classification and quantification of mental diseases. In *Proceedings of the CEUR Workshop Proceedings, Thessaloniki, Greece*, pages 22–25.

Sayanta Paul, Sree Kalyani Jandhyala, and Tanmay Basu. 2018. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In *In Proceedings of CLEF Working Notes*.

Thabit Sabbah, Ali Selamat, Md Hafiz Selamat, Fawaz S Al-Anzi, Enrique Herrera Viedma, Ondrej Krejcar, and Hamido Fujita. 2017. Modified frequency-based term weighting schemes for text classification. *Applied Soft Computing*, 58:193–206.

Kayalvizhi Sampath, Thenmozhi Durairaj, Bharathi Raja Chakravarthi, and Jerin Mahibha C. 2022. Findings of the shared task on Detecting Signs of Depression from Social Media. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Association for Computational Linguistics.

Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, Tat-Seng Chua, and Wenwu Zhu. 2017. Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI*, pages 3838–3844.

Baoxun Xu, Xiufeng Guo, Yunming Ye, and Jiefeng Cheng. 2012. An improved random forest classifier for text categorization. *JCP*, 7(12):2913–2920.