# Analysis of Global Word Representations for Depression Detection

Niveditha Sekar[a], S Chandrakala[a*] and G Prakash[b*]

[a] Intelligent Systems Lab, School of Computing, SASTRA Deemed to be University, Thanjavur, India
[b] Department of CSE, Amrita School of Engineering, Bengaluru, Amrita Vishwa Vidyapeetham, India

**Abstract**

Social media such as Twitter, Facebook, Google plus, Reddit, Tumblr have been a widely used platform for people to communicate, share views and feelings with others freely. The information obtained from this short text messages helps in predicting their emotions, views, sentiment, opinion and it is applied in different fields like marketing, election, product review, sentiment analysis, emotion detection etc. Behavioral analysis from text data is another widely popular field. This paper gives an analysis of global word representations and overview of the work done on depression detection related tasks. Major steps such as pre-processing of data, feature extraction, representation and classification methods are summarized.

**Keywords**

Social media, depression detection, behavioral analysis, emotion detection, GloVe representation, deep learning

## 1. Introduction

Behavioral analysis is the study of human behavior. It involves observing the behavior, identifying the mental state, analyzing and understanding the change in human behavior. Behavioral analysis is also called as emotional/sentimental analysis. Among several emotions, the crucial ones are with negative emotions. Some negative emotions are stress, depression, frustration, hate, envy, anger, anxiety, boredom and panic. These emotions may affect the mental health as well as physical health of a person. In which, depression is a persistent mood disorder and in the worst case, it can be a life-threatening one. So it is essential to identify the people at the risk of depression. Face-to-face interviews and a set of questionnaire are used by Psychiatrist, to understand the behavioral health of the person. It provides a more accurate result, but few people are not aware of abnormalities in their mental health to consult a Psychiatrist. In order to address this, the Depression can be detected from the social media data of the users itself [1]. Since most of the people around the world are using social media like Facebook, Twitter, Instagram etc. Depression can be detected from their text messages, status updates, posts they are sharing, self-reported surveys and the communities or pages they are following [2-4].

This analysis can be done from text data, speech/audio data and visual data [5], [6]. The data for this analysis can be collected from any social media. Since most of the user prefers to share short text messages on the events happening around them or information about them, it is more informative to analyze the social media text data. This sentimental analysis is very popular since it is needed in wide application areas of marketing, artificial intelligence, political science, human-computer interaction, psychology, stock market prediction etc. Figure 1 shows the flow diagram of depression detection system.

ISIC'21: International Semantic Intelligence Conference, February 25-27, 2021, New Delhi, India
EMAIL: nivedithasekarit@gmail.com (Niveditha Sekar); chandrakala@cse.sastra.edu (S Chandrakala); gprakas_74@rediffmail.com (G Prakash)
* Corresponding Author

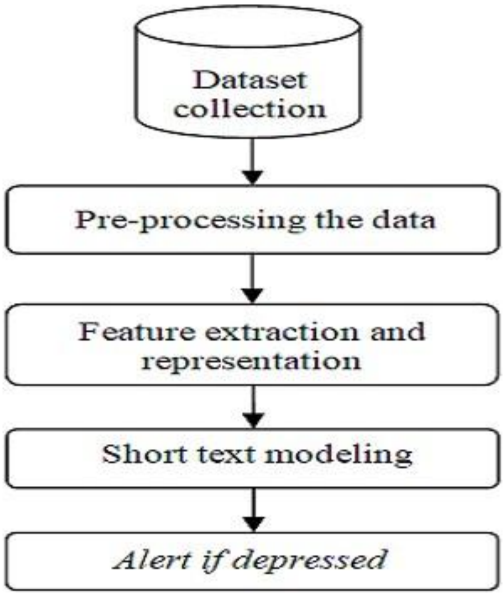


**Figure 1**: Flow diagram of Depression detection system

## 2. Challenges in short text data analysis

Text data which is collected from any social media does not have a structure. Each user expresses his/her view in different ways and their text includes new words, short form of words, errors in the spelling of the words etc. [7], [8]. It is difficult to detect depression from a single tweet of a user. Thus, we need to observe a history of tweets of a particular user [9]. Also, there is a word limit for twitter tweets. Within 140 characters it is hard to express one's feelings and also it is hard for the analyst to interpret their feelings [10]. In order to identify their emotion, they have to analyze the comment and retweets about that particular tweet. This is a long chain process, to detect the emotion of a particular user. A large collection of tweets, from the history of that particular user have to be taken into account and also the comments, retweets for each tweet by the user have to be considered for this emotion detection.

A typical social media user used to share information about them in any of this form text messages, photos or videos. They share information in a consistent manner. The opposite of this is also true i.e. users who are under stress or depression are not much interested to have communication on social media [5], [10]. This low activeness in social media results in lesser tweets and thereby it is difficult to identify the emotion of the user with accuracy. The main task in the emotion analysis is to understand the semantic nature of the short text messages. Most of the features identified from the short text or tweet are sparse features. It is really challenging to detect the emotion from those sparse features since they contribute very less value in the detection of emotion [11]. In a word-level representation, most of the words identified are ambiguous, and they also contain stop words. Hence, it is difficult to identify their emotion class label by a classifier [11]. It is also difficult to identify the original meaning of the sentence, when it has a sarcastic tone. Since the sentences may sound joyful, but they actually express sadness. It leads to false positive in the result [8], [12].

## 3. Short text datasets for depression detection

The short text dataset can be collected through the Twitter public API [5], [13], [14] or through the short text datasets, which are already available [15-17]. Twitter public API provides a means to access the twitter software platform. Several software libraries are available for each programming language namely tweepy for Python and rtweet for R. Twitter API is of two types, they are Twitter REST API and Twitter Streaming API. Twitter Streaming API [18], [19] will provide live tweets until you stop, whereas REST API will provide historical data. Table 1 lists few short text datasets, which are used in depression detection literature.

**Table 1**
Summary of few Short Text Datasets for Depression Detection

| Ref. | Dataset name | Description |
|---|---|---|
| [7], [9], [15] | CLPsych dataset | 1,746 twitter users examples, in which 246 are PTSD users and 327 are depressed users. |
| [9] | BellLetsTalk campaign dataset | All tweets with #BellLetsTalk hashtag are collected, in which 95 people disclosed that they are depressed. |
| [16] | CLEF/eRisk 2017 dataset | 887 Reddit users examples, in which 135 are depressed. |
| [10] | Sina weibo dataset | 23,304 users tweets are crawled, in which 11,074 users are stressed. |
| [17] | LiveJournal dataset | This dataset consists of 2,132 posts. In which, 758 are depressed posts. |
| [11] | SemEval 2007 dataset | This dataset consists of 1,250 news headlines. They are labelled into 6 emotions. |
| [11] | ISEAR dataset | This dataset contains 7,666 sentences. They are labelled into 7 emotions. |

## 4. Pre-processing the short text data

Before feature selection, the short text data is pre-processed to refine the unstructured and noisy data. Pre-processing phase is an important phase, as it helps in improving the performance.

- In the pre-processing phase, all the non-ASCII, non-English characters, URLs and @username are removed. Since they are not contributing any valued information to the depression detection system.
- All the acronyms are expanded to its full form like “idk” as “I don’t know”.
- This phase performs tokenizing, stemming and removing stop words [17], [20], [21]. Tokenizing process will split the texts into sequence of tokens. Stemming process will reduce the length of a word, by reducing the word to its word stem like “rained”, ”raining” as “rain”. Stop words are removed, some of them are “a”, “the”, “and” etc.
- In each word, if a letter is appearing continuously more than twice then it is replaced with its appropriate word [22], [23] like “Noooooo” as “No”.
- Also negative references are replaced by their full words i.e. “can’t” is replaced by “cannot”.
- Emoticons and emojis are replaced with their words.

## 5. Feature extraction and representation

### 5.1. Feature extraction

From the pre-processed data, the features are extracted, represented and are given as input to the classification methods. There are several features or attributes involved in the process of depression detection. Some of the features used for this depression detection are user-level feature, tweet-level feature, temporal feature, non-temporal feature, social interaction feature, content feature, posting behavior feature, term frequency feature, Bag-Of-Words (BOW) feature, hashtags, negation, LIWC feature, word N-gram feature, Part-of-speech (POS) feature, topic, tweet frequency, RT [24] etc. Several feature extraction techniques are available as built-in commands in R language, SciPy, Numpy etc.

The tweet-level attributes will give information from the tweet, image, retweets, comments and likes. The user level attributes will provide more information on the emotion of the user; it includes the behavior of the user from their social interaction and from their posts. The social interaction attributes have information about the content and the structure in which the user communicates with his friends [5], [10]. Tweets are classified in time series for temporal feature, whereas history of tweets is used in non-temporal feature. Term frequency feature gives the frequency count of individual word or n-gram of words. POS feature finds the adjectives since they provide more information. Negation feature gives the actual opinion orientation like “not happy” is equivalent to “sad” [25]. Bag-Of-Words will provide the occurrence of each word in a document. Word N-grams feature is similar to Bag-Of-Words. N-gram includes phonemes, syllables, letters, words [16]. To reduce dimension or attributes Principal Component Analysis (PCA) is used [26].

## 5.2. Representation

There are several feature representation models are available. Some of the representation models are Word2Vec representation, FastText, Global vector for word representation (GloVe) model, word N-gram feature representation, twitter specific feature representation, word sentiment polarity score representation, word representation features, temporal feature vector, non-temporal feature vector etc.

Word2Vec representation uses continuous skip gram and BOW features. Based on non-temporal feature, overall emotion score is calculated. For temporal feature, if a user did not tweet anything for a day its score is taken as zero. In such a way, emotion score vector is calculated [27]. In word embedding, all the words are mapped into a multi-dimensional vector, where semantically related words are neighbors. The word sentiment polarity score representation, finds either the word has a strong relationship with positive sentiment or non-positive sentiment. To identify this, it uses the lexicon based sentiment feature and Senti-wordnet. FastText is similar to skip-gram representation, where each ngram has its vector. Vector representation helps to improve the performance, as it provides the hidden details [36]. The GloVe model is a regression model which will map words with similar context into a feature vector [28]. The GloVe representation model proves to be effective and is showing improved performance when combined with Deep Convolutional Neural Network, than the state-of-art approaches [28], [40].

# 6. Depression detection methods

The extracted features and derived representations are fed as input for further modeling. Depression can be detected from the short text data with the help of various modeling methods, such as Discriminative model based methods, Ensemble model based methods, Probabilistic model based methods, ANN based methods, Deep learning based methods and Unsupervised learning based methods.

## 6.1. Discriminative model based methods

SVM is a discriminative classifier. SVM is most suited for text data, because of the sparse nature of the text. Text data can be categorized into two categories. They are user-level attributes and tweet-level attributes. In tweet-level category, first the features are extracted, next the features are segregated into different classes like depressed words, non-depressed words, polarity words, stop words etc. In user-level category, the user tweet history is considered. All the tweets of the user are considered like a single tweet and then tweet-level detection is performed. It uses (BOW) to get the vocabulary. Then it is trained using SVM in original dataset, dataset balanced by under-sampling and dataset balanced by over-sampling. It is observed that user-level classification gives high performance with respect to recall measure in comparison with the tweet-level classification even for the limited number of feature. Also it is difficult to detect whether the user is depressed or not from a single tweet/post, hence user-level category is used [9]. It is also observed that when Linear SVM is applied on BOW feature, it provides good performance in terms of Recall measure [15]. SVM gives good accuracy when compared with Naïve Bayes and Logistic regression methods [29]. Table 2 gives the summary of few discriminative modeling.

## 6.2. Ensemble model based methods

Random Forest (RF) classifier is an ensemble classifier. It is a multitude of decision tree, for more accurate results. To detect depression from the text data, temporal feature and non-temporal feature are used. Feature vector from non-temporal feature is referred as EMO. EMO, LIWC and combination of EMO+LIWC feature sets are given as input to Random Forest classifier. It is observed that RF gives high precision and recall than SVM [30]; also it provides more information with temporal features [27]. RF classifier is also used to classify the online post and communities into depressive and non-depressive. On top of the extracted LIWC feature, RF is applied to classify them. Hierarchical HMM is used for determining the degree of depression in the social communities. RF, Logistic Regression, and Gaussian NB are applied with different representation methods such as Word2Vec, FastText with Skip-gram, and GloVe. RF provides better performance than the other models when combined with FastText [36]. Table 3 gives the summary of few ensemble modeling.

## 6.3. Probabilistic model based methods

Naïve Bayes is a probability based classifier. Naïve Bayes algorithm has assumes that each feature is independent. Bag-Of-Words (BOW) approach will provide the words with its occurrence frequency. BOW feature is given as input to different classification algorithms like DT, NB, Linear SVM and Logistic Regression. Each tweet is treated as a document. Here Bag-Of-Words finds the occurrence frequency of words related to depression. Decision tree will provide results for most of the cases, but it may be unstable when there is a change in data. Linear SVM is also used for this purpose, where a straight line is used to differentiate classes. It uses a maximum-margin hyperplane to perform this identification of classes. Logistic Regression uses the probability of words belonging to a particular class and curve is drawn to identify the best fit for the depression case. Here Naïve Bayes theorem shows better performance with respect to accuracy when compared with other classifier algorithms. When evaluating with respect to Precision and F1-score Logistic Regression gives good performance [15]. Also, Naïve Bayes is the best classification approach when compared with BP neural network and Decision tree. Also, Naïve Bayes gives high precision and recall value [26]. Table 4 gives the summary of few probabilistic modeling.

## 6.4. Artificial Neural Network (ANN) based methods

Artificial Neural Network (ANN) is combined with several unsupervised learning model to detect depression from social media text data. Some of the unsupervised learning models are Biterm Topic model, Word2vec, Replicated Softmax Machine. BTM identifies words that appear together. It will identify two words that appear together, if the size of window is given as two. BTM uses topic to represent the hidden aspects of the document. Word2vec is a word embedding process, identifies both semantic and syntactic regularities in the sentence. And it will group them in clusters, if the vectors have similar semantic meanings i.e. it computes the association with words and groups them together. RSM is similar to term frequency counter, it will count the occurrences of a particular word in the vocabulary collected. RSM also identifies the hidden topical structure. On top of this unsupervised learning model, Stochastic Gradient Descent (SGD) model is applied. SGD acts as transfer learning approach, as this will transfer the high-level semantic features to ANN. In order to filter the noisy feature and to maintain the stability of this model, Sparse Encoding method is applied. The transfer learning approach used in this Hybrid Neural Network is called as Latent Semantic Machine (LSM). It accepts the raw features from the unsupervised learning models and derives them into high level semantic feature mixture, which will be fed into the Neural network. It is observed that HNN+BTM with one LSM and HNN+BTM with two LSM performed better in terms of F1 measure, than HNN with other unsupervised learning models. It is also observed that HNN+RSM and

HNN+Word2vec with sparse encoding give better performance than HNN+RSM and HNN+Word2vec without sparse encoding. The selection of the unsupervised learning models for extracting the source features added more value to this HNN model [11].

Feed Forward (FF) is a type of ANN. The Reddit dataset is pre-processed and fed to FF neural network method. This FF modeling is used for multiclass classification, which involves "selfharm", "suicidewatch", "anxiety", "depression", etc. It is observed that FF classifier gives more accurate results when compared with SVM and linear regression [31]. Table 5 gives the summary of few ANN based modeling.

## 6.5. Deep learning based methods

### 6.5.1. Convolutional Neural Networks (CNN)

*CNN with global max pooling layer.* The pre-processing of twitter data provides a vocabulary for further phases. The words are encoded into a sequence of fixed length and occurrence of the word is limited to two times in that sequence. Then unsupervised training models are used to transform the encoded words into a low dimensional vector. For this many models are available like Skip-gram and CBOW. Skip-gram concentrates on the contextual words and is able to detect rare words whereas the CBOW concentrates on the current words and is much of a continuous Skip-gram. Skip-gram and CBOW are the two layers of Word2Vec model. This unsupervised training model is performed with different sense and it involves two tasks. They are predicting word and sense from the input. For this, it first identifies the word that occurs together, example "happy" it can come with words like "journey", "morning", "birthday". Then Rectified Liner Unit (ReLU) is used, this will identify the label for the missing data and sense of the sentence, thereby produce the label output.

On top of these embeddings, variants of CNN are applied. **CNNWithMAX** means Convolution with 250 layers is applied and then the global max pooling layer is applied to extract the global information. In **MultiChannelCNN**, three times CNN is applied, with the filter of length 3, 4, and 5. **MultiChannelPoolingCNN** is same as MultiChannelCNN but with two different max-pooling sizes 2 and 5. MutiChannel CNN and bi-directional GRU combined to give more accuracy than CNN [38]. These CNN variants are compared with the RNN model and it is observed that CNN with global max pooling layer gives high performance than RNN based model by providing the highest precision and recall [7].

*CNN with Factor Graph Model (FGM).* CNN is combined with the Factor Graph Model (FGM) to extract more tweet level and user level information. In this approach, CNN method is applied on the dataset along with the Cross Auto Encoders (CAE). CNN will provide the user-level attributes, which is obtained from tweet-level. Then this will be given as input to the next phase FGM. FGM considers three factors and three aspects of this attributes to map this into states. The three factors are attribute factor, dynamic factor and social factor. To depict the correlation of the stress state and time with attribute, attribute factor is used. Dynamic factor is used to give correlation of the stress state and dynamic time. Social factor is used to depict the correlation between the stress state and time with polarity comments. The three main aspects, which FGM is taking into account, are the following user-level attributes: posting behavior, content, and social interaction. Based on these factors and aspects the user-level attributes are mapped with the respective stress state level. This CNN+FGM give better performance, by providing the highest precision and recall, when compared with the traditional methods like SVM, RF, LR [10].

*DCNN with Global vector for word representation (GloVe) model.* DCNN method helps to identify whether the tweets express positive or non-positive emotion. Before applying Deep Convolutional Neural network, the tweets are preprocessed, features are extracted and represented into feature vector using GloVe model. The GloVe model is a regression model which combines the following two methods local context window and global matrix factorization. Deep Convolutional Neural Networks (DCNN) is applied on the vector, generated by GloVe model. The twitter specific feature vector, unigram and bigram feature vector, word sentiment polarity score feature vector are combined into a single feature vector. In the first Convolutional layer, on top of the combined feature vector, Convolutional filter is applied to get new vector. The vector is mapped to a fixed length vector. Again convolutional layer is applied to get new vector. This GloVe+DCNN model uses three k-max pooling layer and three convolutional layers to give the probability of positive or negative sentiment in the tweet. It is observed that GloVe+DCNN provide high precision and recall when compared with BoW/GloVe with SVM or LR [28].

### 6.5.2. Recurrent Neural Networks (RNN)

RNN is widely used in NLP. Word2Vec model is used to represent the vocabulary. It also helps to determine or predict the word and sense from the input. Rectified Liner Unit (ReLU) is also used. ReLU helps in identifying the missing label for the data and also identifies the sense of the sentence. This embedding is given as input to RNN model. RNN is applied with Bidirectional LSTM and context-aware attention. LSTM prevents error from exploding and vanishing gradient problems. Bidirectional RNN connects the output from two hidden layers of opposite direction to the same output. Bidirectional LSTM helps to concatenate both forward and backward representation. Context-aware attention provides the weighted sum of all words in a sequence and also it helps to focus on the more important words. It is observed that optimized embedding performed better, than the trainable random embedding for RNN. Also, when compared with CNN based models, RNN shows low performance with respect to precision and recall [7].

*LSTM and Gate Recurrent Unit (GRU).* LSTM and GRU are best suited for predicting long-term data involving delay. Combining GRU with LSTM helps in handling the difficulties in LSTM, which is the training speed. GloVe representation is used to utilize both local and global details of the data. Among RNN, LSTM, GRU and LSTM-GRU, LSTM-GRU provides better performance [37]. Table 6 gives the summary of few deep learning based modeling.

## 6.6. Unsupervised learning based methods

K-means is an unsupervised learning method. Before applying k-means to the observations, the collected data is pre-processed. Then the data is analyzed, by calculating the word frequency. Words in the vocabulary are represented into vector, using one-hot encoding or word embedding process. Word2Vec model can also be used to generate vectors. Then k-means clustering is applied, words with similar meaning are grouped together in clusters. Based on cosine similarity it is easy to accumulate semantically similar words in the clusters [32]. For Latent Dirichlet Allocation (LDA) method, the extracted N-gram features are fed as input. LDA is applied on term-document matrix and gives output as topic-document matrix, which is fed into Multilayer Perceptron (MLP). MLP works with 30 topics as input and two hidden layer of 60 and 30 units. It gives comparatively moderate performance with respect to precision and recall, which is due to the unsupervised nature of the topic extraction [33]. Table 7 gives the summary of few unsupervised learning based modeling.

**Table 2**
Summary of few Discriminative Model Based Methods

| Ref. | Dataset | Feature | Classifier | Performance |
|---|---|---|---|---|
| [9] | CLPsych2015, BellLetsTalk 2015 | Bag-Of-Words, User-level, tweet-level feature | SVM | Precision-0.58, Recall-0.77 |
| [15] | CLPsych2015 | Bag-Of-Words | SVM | Precision-0.83, Recall-0.83 |

| Ref. | Dataset | Feature | Classifier | Performance |
|---|---|---|---|---|
| [29] | Twitter, 20newsgroups | N-gram , negation, Part-Of-Speech (POS) | SVM | Precision-0.83, Recall-0.79 |

**Table 3**
Summary of few Ensemble Model Based Methods

| Ref. | Dataset | Feature | Classifier | Performance |
|---|---|---|---|---|
| [27] | Twitter Streaming API | Non-Temporal(EMO), Temporal(EMO-TS), LIWC feature | RF | Precision-0.90, Recall-0.86 |
| [30] | LiveJournal | LIWC feature | RF | Precision-0.89, Recall-0.90 |
| [36] | Twitter dataset | Skip-gram | RF | Precision-0.82, Recall-0.81 |

**Table 4**
Summary of few Probabilistic Model Based Methods

| Ref. | Dataset | Feature | Classifier | Performance |
|---|---|---|---|---|
| [15] | CLPsych2015 | Bag-Of-Words | Naïve Bayes | Precision-0.82, Recall – 0.82 |
| [26] | Questionnaire data | Words, pronoun and punctuation feature | Naïve Bayes | Precision-0.75, Recall-0.62 |

**Table 5**
Summary of few Artificial Neural Network (ANN) Based Methods

| Ref. | Dataset | Feature | Classifier | Performance |
|---|---|---|---|---|
| [31] | Reddit dataset | Linguistic features | Feed Forward (FF) | Precision-0.92, Recall-0.88 |
| [11] | SemEval and ISEAR dataset, SinaNews dataset | words feature | ANN with unsupervised learning model | F1-0.60 |

**Table 6**
Summary of Few Deep Learning Based Methods

| Ref. | Dataset | Feature | Classifier | Performance |
|---|---|---|---|---|
| [7] | CLPsych2015, BellLetsTalk2015 | Continuous Bag- Of-Words, Skip gram | CNN with global max pooling layer | Precision-0.87, Recall-0.87 |
| [10] | Sina Weibo's REST APIs, Tencent Weibo | User-level feature, Tweet-level feature | CNN with FGM | Precision-0.90, Recall-0.96 |
| [28] | Stanford Twitter Sentiment Test dataset, SemEval 2014 Task9 dataset, Stanford Twitter Sentiment Gold dataset, Sentiment Evaluation Dataset, Sentiment Strength Twitter dataset | Word N-gram feature, emoticons, hashtags, negation, Part- Of-Speech (POS) and capitalized words. | DCNN with GloVe model | Precision-0.88, Recall-0.87 |
| [7] | CLPsych2015, BellLetsTalk2015 | Continuous Bag- Of-Words, Skip gram | RNN with bidirectional LSTM | Precision-0.63, Recall-0.65 |
| [37] | IMDB dataset | words feature | LSTM and GRU | F1-0.86 |

**Table 7**
Summary of few Unsupervised Learning Based Methods

| Ref. | Dataset | Feature | Classifier | Performance |
|---|---|---|---|---|
| [32] | Twitter data | Word frequency | K-means | Cosine similarity helps in easy clustering |
| [33] | eRisk 2018 pilot task dataset | N-gram features | Latent Dirichlet Allocation (LDA) | Precision-0.32, Recall-0.62 |

# 7. Analysis of Global word representations

## 7.1. Dataset overview

The dataset used in the following experimental analyses is "CLEF/eRisk 2018 dataset". The aim of the CLEF eRisk is to identify the people, liable to depression from the data available on the Internet. It paved a way of interdisciplinary research in the field of depression related problems. The people under depression can be alerted when early signs of depression are found. eRisk 2017 dataset focussed on the early risk prediction with multiple actors (Ex: Children sexual abuse) and with single actors (Ex: Depression, bipolar disorder, teenage distress) from online text data. eRisk 2018 dataset is formed with the 2017 dataset, it involved in the early prediction of Depression and Anorexia among the social media users. Both eRisk 2017 and eRisk 2018 uses the same source of data, i.e. it collects the social media texts from a particular collection of users. The data is arranged in chronological order of 10 chunks from oldest to newest of each user. It provides data for both training and testing. The training data is divided into depressed and control groups i.e., non-depressed. The eRisk 2017 dataset is a collection of writings from 887 social media users, where 135 are depressed. The eRisk 2018 dataset is an extended collection of 2017 dataset, which consists of writings from 1,707 users, where 214 users are depressed.

## 7.2. Methodologies used

**Analysis with TF-IDF representation and LDA**. The eRisk dataset is pre-processed as an initial step. The TF-IDF vectorizer is well suited for text dataset. As this will provide the unique list of words used in the dataset, along with their frequency of occurrence. It helps in classifying the words under a particular set of topics. The TF-IDF vectorizer of Scikit-learn converts the writings of social media users into a matrix of TF-IDF features. The terms extracted using the TF-IDF vectorizer is formed as a matrix and given as input to Latent Dirichlet Allocation (LDA). The output of LDA is the topic matrix. As each document is composed of different topics or attributes. And each topic is composed of different words. This topic matrix is given as input to the MLP model, it consists of two intermediate layers of 50 & 20 units. By this approach, each user is labeled as depressed or not. The performance of the TF-IDF and LDA model is depicted in Table 8.

**Analysis with GloVe and RNN.** GloVe combines the advantage of methods local context window and global matrix factorization, to provide meaningful word insights. The GloVe model is providing promising results for text classification. It is combined with the RNN model, as RNN model widely used for text classification and Natural Language Processing (NLP). The eRisk dataset is pre-processed and tokenized and then given as input to the GloVe representation model. To provide meaningful statistics, GloVe forms the word-to-word co-occurrence matrix. The Resultant of GloVe is given as input to the RNN model. It involves two hidden layers of varying units. The output layer of the RNN helps in labeling the users as a depressed person or non-depressed person. The performance of GloVe and RNN model is depicted in Table 8.

**Analysis with GloVe and CNN.** The GloVe model is observed to be effective for sentiment analysis from text data mining [28], the GloVe representation model is combined with CNN to analyze the result. The dataset taken for this analysis consists of a few empty writings, which are ignored. Then the dataset is pre-processed while preserving the emoticons and symbols since they provide valuable information. Each user's writing in each chunk is analyzed and formed a matrix of words with a pre-trained set of word embeddings. This pre-processed tokenized input is given to a single Convolutional layer of 100 filters with CReLU activation. A Single max pooling layer is applied to classify each user as depressed or not. The GloVe model is combined with different layers of CNN and LSTM network and the performance is observed high for GloVe with the **multiple layers of CNN and bi-LSTM** [39]. The performance of GloVe and CNN model is depicted in Table 8.

## 7.3. Performance analysis

The Classification report and Confusion matrix is used to analyze the performance of the above methodologies. The following Table 8 shows Precision, Recall and F1 of these three methods.

The TF-IDF representation focuses mainly on the frequency of word occurrence in documents. Then maps the word into an appropriate topic, thereby it classifies them. In this case, whenever some word related to depression comes, it classifies them as depressed, which is not the ideal way. The Global vector for word representation considers the frequency of word occurrence and the frequency of co-occurrence of words thereby provides more valuable information in classifying them. The GloVe representation founds to be significantly better than the most commonly used word representation TF-IDF. The RNN and CNN classifier works well with text representation and its performances are analyzed with GloVe representation. From the table, it is found that GloVe representation is better than TF-IDF representation. Also, GloVe representation performs well with CNN than RNN. This is because the RNN model gives better result with word embeddings of higher length. From the analysis, it is found that GloVe representation with CNN classifier provides comparatively better results.

**Table 8**
Summary of Performance Analysis

| Methodology | Precision | Recall | F1 |
|---|---|---|---|
| TF-IDF representation with LDA | 0.62 | 0.28 | 0.49 |
| GloVe with RNN [34] | 0.80 | 0.20 | 0.31 |
| GloVe with CNN [35] | 0.42 | 0.66 | 0.51 |
| GloVe with multiple CNN and bi-LSTM [39] | 0.60 | 0.54 | 0.55 |

## 8. Future directions

The performance of depression detection system can be improved or made more meaningful with the following directions for future research.

- The depression detection task can also be done by extracting the emotions from the speech data.
- It can also be extended by grouping users, based on gender, age, locations and other demographic attributes.
- The Spatiotemporal features from video data can also be included, as they contribute more information.
- Daily variation of a user's depression can also be monitored.
- It can be extended by including the medical context, so the clinical depression can be detected from social media data.

## 9. Conclusion

This paper provided an overview of the depression detection system and the analysis of global word representations from the short text data. Datasets and Machine learning methods, used in recent years for the depression detection are summarized. The global word representations model proved to be effective is analyzed with different classifiers. Various challenges and future directions are summarized for future research.

## References


[1] Juyoung Song, Tae Min Song, Dong-Chul Seo, and Jae Hyun Jin.: "Data Mining of Web-Based Documents on Social Networking Sites that Included Suicide-Related Words among Korean Adolescents". Journal of Adolescent Health, 59(6):668–673 2016.

[2] Budhaditya Saha, Thin Nguyen, Dinh Phung, and Svetha Venkatesh.: "A Framework for Classifying Online Mental Health-Related Communities with an Interest in Depression". IEEE Journal of Biomedical and Health Informatics, 20(4):1008–1015 2016.

[3] Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt.: "Detecting Depression and Mental Illness on Social Media: An Integrative Review". Current Opinion in Behavioral Sciences, 18:43–49 2017.

[4] Elizabeth M Seabrook, Margaret L Kern, Ben D Fulcher, and Nikki S Rickard.: "Predicting Depression from Language-Based Emotion Dynamics: Longitudinal Analysis of Facebook and Twitter Status Updates". Journal of Medical Internet research, 20(5) 2018.

[5] Guangyao Shen, Jia Jia, Liqiang Nie, Fuli Feng, Cunjun Zhang, Tianrui Hu, TatSeng Chua, and Wenwu Zhu.: "Depression Detection via Harvesting Social Media: A Multimodal Dictionary Learning Solution". In Proceedings of the

Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI-17), pages 3838–3844, 2017.
[6] Andrew G Reece and Christopher M Danforth.: "Instagram Photos Reveal Predictive Markers of Depression". EPJ Data Science, 6(1):15 2017.
[7] Ahmed Husseini Orabi, Prasadith Buddhitha, Mahmoud Husseini Orabi, and Diana Inkpen.: "Deep Learning for Depression Detection of Twitter Users". In Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic, pages 88–97 2018.
[8] Harshali P Patil and Mohammad Atique.: "Sentiment Analysis for Social Media: A Survey". In Information Science and Security (ICISS), 2015 2nd International Conference on, pages 1–4 2015.
[9] Zunaira Jamil.: "Monitoring Tweets for Depression to Detect At-risk Users", 2017.
[10] Huijie Lin, Jia Jia, Jiezhong Qiu, Yongfeng Zhang, Guangyao Shen, Lexing Xie, Jie Tang, Ling Feng, and Tat-Seng Chua.: "Detecting Stress Based on Social Interactions in Social Networks". IEEE Transactions on Knowledge and Data Engineering, 29(9):1820–1833 2017.
[11] Xiangsheng Li, Yanghui Rao, Haoran Xie, Raymond Yiu Keung Lau, Jian Yin, and Fu Lee Wang.: "Bootstrapping Social Emotion Classification with Semantically Rich Hybrid Neural Networks". IEEE Transactions on Affective Computing, 8(4):428–442 2017.
[12] Monireh Ebrahimi, Amir Hossein Yazdavar, and Amit Sheth.: "Challenges of Sentiment Analysis for Dynamic Events". IEEE Intelligent Systems, 32(5):70–75 2017.
[13] Zhaoxia Wang, Chee Seng Chong, Landy Lan, Yinping Yang, Seng Beng Ho, and Joo Chuan Tong.: "Fine-Grained Sentiment Analysis of Social Media with Emotion Sensing". In Future Technologies Conference (FTC), pages 1361–1364 2016.
[14] Sara Rosenthal, Noura Farra, and Preslav Nakov.: "SemEval-2017 task 4: Sentiment Analysis in Twitter". In Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017), pages 502–518 2017.
[15] Moin Nadeem.: "Identifying Depression on Twitter". arXiv preprint arXiv:1607.07384 2016.
[16] Maxim Stankevich, Vadim Isakov, Dmitry Devyatkin, and Ivan Smirnov.: "Feature Engineering for Depression Detection in Social Media" 2018.
[17] Maryam Mohammed Aldarwish and Hafiz Farooq Ahmad.: "Predicting Depression Levels Using Social Media Posts". In Autonomous Decentralized System (ISADS), 2017 IEEE 13th International Symposium on, pages 277–280 2017.
[18] Felix Ming Fai Wong, Chee Wei Tan, Soumya Sen, and Mung Chiang.: "Quantifying Political Leaning from Tweets, Retweets, and Retweeters". IEEE Transactions on Knowledge and Data Engineering, 28(8):2158–2172 2016.
[19] Zhenhua Zhang, Qing He, Jing Gao, and Ming Ni.: "A Deep Learning Approach for Detecting Traffic Accidents from Social Media Data". Transportation Research Part C: Emerging Technologies, 86:580–596 2018.
[20] C Sindhu, Dyawanapally Veda Vyas, and Kommareddy Pradyoth.: "Sentiment Analysis Based Product Rating Using Textual Reviews". In Electronics, Communication and Aerospace Technology (ICECA), 2017 International conference of, volume 2, pages 727–731.
[21] Anukarsh G Prasad, S Sanjana, Skanda M Bhat, and B S Harish.: "Sentiment Analysis for Sarcasm Detection on Streaming Short Text Data". In Knowledge Engineering and Applications (ICKEA), 2017 2nd International Conference on, pages 1–5 2017.
[22] Sonia Xylina Mashal and Kavita Asnani.: "Emotion Intensity Detection for Social Media Data". In Computing Methodologies and Communication (ICCMC), 2017 International Conference on, pages 155–158 2017.
[23] Ahmed Al-Saffar, Suryanti Awang, Hai Tao, Nazlia Omar, Wafaa Al-Saiagh, and Mohammed Al-bared.: "Malay Sentiment Analysis Based on Combined Classification Approaches and Senti-Lexicon Algorithm". PloS One, 13(4):e0194852 2018.
[24] Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh,

and Hiroyuki Ohsaki.: “Recognizing Depression from Twitter Activity”. In Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems, pages 3187–3196 2015.

[25] Amit G Shirbhate and Sachin N Deshmukh.: “Feature Extraction for Sentiment Classification on Twitter Data”. International Journal of Science and Research (IJSR) ISSN (Online), pages 2319–7064 2016.

[26] Changye Zhu, Baobin Li, Ang Li, and Tingshao Zhu.: “Predicting Depression from Internet Behaviors by Time-Frequency Features”. In Web Intelligence (WI), 2016 IEEE/WIC/ACM International Conference on, pages 383–390 2016.

[27] Xuetong Chen, Martin D Sykora, Thomas W Jackson, and Suzanne Elayan.: “What About Mood Swings: Identifying Depression on Twitter with Temporal Measures of Emotions”. In Companion of the The Web Conference 2018 on The Web Conference 2018, pages 1653–1660 2018.

[28] Zhao Jianqiang, Gui Xiaolin, and Zhang Xuejun.: “Deep Convolution Neural Networks for Twitter Sentiment Analysis”. IEEE Access, 6:23253–23260 2018.

[29] Anees Ul Hassan, Jamil Hussain, Musarrat Hussain, Muhammad Sadiq, and Sungyoung Lee.: “Sentiment Analysis of Social Networking Sites (SNS) Data Using Machine Learning Approach for the Measurement of Depression”. In Information and Communication Technology Convergence (ICTC), 2017 International Conference on, pages 138–140 2017.

[30] Iram Fatima, Hamid Mukhtar, Hafiz Farooq Ahmad, and Kashif Rajpoot.: “Analysis of User-Generated Content from Online Social Communities to Characterise and Predict Depression Degree”. Journal of Information Science, page 0165551517740835 2017.

[31] George Gkotsis, Anika Oellrich, Sumithra Velupillai, Maria Liakata, Tim JP Hubbard, Richard JB Dobson, and Rina Dutta.: “Characterisation of Mental Health Conditions in Social Media using Informed Deep Learning”. Scientific reports, 7:45141 2017.

[32] Long Ma, Zhibo Wang, and Yanqing Zhang.: “Extracting Depression Symptoms from Social Networks and Web Blogs via Text Mining”. In International Symposium on Bioinformatics Research and Applications, pages 325–330 2017.

[33] Maupome´.: “Using Topic Extraction on Social Media Content for the Early Detection of Depression” 2018.

[34] Paul, S., Kalyani, J.S. and Basu, T.,: “Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks”. 2018.

[35] Trotzek, M., Koitka, S. and Friedrich, C.M.,: “Word Embeddings and Linguistic Metadata at the CLEF 2018 Tasks for Early Detection of Depression and Anorexia” 2018.

[36] Kaibi, Ibrahim, and Hassan Satori.: “A comparative evaluation of word embeddings techniques for twitter sentiment analysis”. In 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), pp. 1-4. IEEE 2019.

[37] Ni, Ru, and Huan Cao.: “Sentiment Analysis based on GloVe and LSTM-GRU”. In 2020 39th Chinese Control Conference (CCC), pp. 7492-7497. IEEE 2020.

[38] Cheng, Yan, Leibo Yao, Guoxiong Xiang, Guanghe Zhang, Tianwei Tang, and Linhui Zhong.: “Text sentiment orientation analysis based on multi-channel CNN and bidirectional GRU with attention mechanism”. IEEE Access 8: 134964-134975 2020.

[39] Goularas, Dionysis, and Sani Kamis.: “Evaluation of deep learning techniques in sentiment analysis from Twitter data”. In 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), pp. 12-17. IEEE 2019.

[40] Abid, Fazeel, Muhammad Alam, and Adnan Abid.: “Representation of Words Over Vectors in Recurrent Convolutional Attention Architecture for Sentiment Analysis”. In 2019 International Conference on Innovative Computing (ICIC), pp. 1-8. IEEE 2019.