Extracting information from social media platform for early detection of depression among individuals

by

Mirza Abdullah Al Noman
19301244
Waleed Bin Habib Khan
19301178
Pratick Roy Chowdhuruy
19301065

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science & Engineering

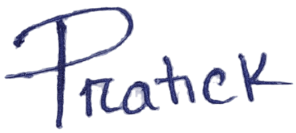
Department of Computer Science and Engineering
School of Data & Sciences
Brac University
May 2023

Signatures

Student's Full Name & Signature:

Mirza Abdullah Al Noman
19301244

Waleed Bin Habib Khan
19301178



Pratick Roy Chowdhuruy
19301065

Supervisor's Full Name & Signature:

Farig Yousuf Sadeque, PhD
Assistant Professor
Department of Computer Science and Engineering
School of Data & Sciences
Brac University

Approval

The thesis/project titled "Extracting information from social media platform for early detection of depression among individuals" submitted by

1. Mirza Abdullah Al Noman(19301244)
2. Waleed Bin Habib Khan(19301178)
3. Pratick Roy Chowdhuruy(19301065)

has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science & Engineering on May , 2023.

Examining Committee:

Supervisor:
(Member)

Farig Yousuf Sadeque, PhD
Assistant Professor
Department of Computer Science and Engineering
School of Data & Sciences
Brac University

Thesis Co-ordinator:
(Member)

Md. Golam Rabiul Alam, PhD
Professor
Department of Computer Science and Engineering
School of Data & Sciences
Brac University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
School of Data & Sciences
Brac University

Abstract

In our world, 3.8% of the total population is suffering from depression and it is the fourth major cause of death in 15-29 year olds. It is estimated that more than 75% of people suffering from it in low and middle income countries receive no treatment. Also, in these countries, so many people live with such conditions without even recognizing it because of the lack of proper diagnosis and mental health facilities. However, a huge chunk of the population is connected and active on different social media platforms. Detection of depression from social media activities can help in recognizing the problems in an individual level and in a public health level to know its prevalence in different demographics. The early prediction of such can help us to work on the problem before the onset. In our work, we propose to use state of the art machine learning and deep learning models to provide an efficient early detection model for diagnosis of such. We hope that it would help individuals and relevant authorities to find out the illness and its severity for the betterment of global and regional mental health.

Keywords: Natural Language Processing; Neural Network; Depression; Major Depressive Disorder; Early Risk Detection; Transformer; Psycholinguistics

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Depression: Severity and Effects

The The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) is the standard diagnostic manual for mental disorders in the United States. The manual defines depression as a mood disorder that is characterized by persistent sadness, loss of interest or pleasure, changes in appetite or weight, changes in sleep patterns, fatigue, difficulty concentrating, and thoughts of death or suicide [18].
According to Global Health Data Exchange (GHDx) [57] prevalence of depression among adults is 5.0%, and for people older than 60 years old it is 5.7%.Globally estimated 280 million are suffering from depression. Women are roughly twice as likely to experience depression than men as the female adult population has 2% more prevalence than the male population [57].
According to estimates, high-income nations like France (21%) and the United States (19%) have the highest rates of depression [22]. Highly developed nations have significantly higher odds of diagnosing depression since their health care systems are far better able to recognize and treat mental diseases. Because of this, less developed nations do not necessarily have lower rates of depression; rather, the treatment of mental diseases is frequently neglected in favor of more pressing issues like starvation, disease, and hygienic conditions. According to the World Health Organization, more than 75% of those with mental problems in low- and middle-income nations do not have access to the appropriate care [44].
Depression frequently exists together with various diseases and health issues such as Cancer, Strokes, Heart attacks, Diabetes and HIV. Depression affects 25% of people with cancer , 10–27% of stroke victims, one in three heart attack survivors and diabetic patients [6]. Also, among HIV-positive people, depression ranks as the second most prevalent mental health disorder[8].
Looking at the mass prevalence, the effects of this are detrimental to both individual and public health. Although it is evident that depressive disorders are common and can have disastrous effects on the population, efforts for diagnosing and treating such disorders are not enough [9].

1.2 Effect of emotion on language

Emotions and mental states of the mind can greatly impact the language of a person. Language is not only a tool for communication but can also be used to construct and

regulate emotions. In accordance with the psychological constructionist view, language helps to categorize emotion and construction of meaningful emotional states [32]. Emotions are not predetermined natural instincts, it is the result of basic elements of psychology and conceptual information that language can provide [15], [19]. Emotion can play a part in development of language in children through acquisition of emotion words in different stages of childhood [24]. Emotion can serve as a medium for communication helping a person manifest and understand others emotions and behaviors with minimum cognitive effort [3]. Emotions can have a decisive effect on the choice of words and overall tone during communication [21]. The choice of words making up the information inside the language can deliver insight into a person's thoughts and feelings which can serve as a window to a person's emotional state.

Emotion lexicon composed of emotional categories in language helps in shaping emotions during communication. Because particular emotion words can help to portray particular emotions and manage them [23]. Varieties of cultures can have differences in categorization of emotions. Consequently, imperfect translations can occur among different languages due to the specification of emotion words. Thus, it may be more accurate to refer to the emotional categories expressed by language as 'cognitive types', 'nominal kinds', or situation-specific ideas that may affect the way speakers of a language interpret meaningful experiences or perspectives at the moment [7], [25], [14].

Emotions having an impact on language is clear. Speakers of different native languages and cultures literally "see" different emotions of the face [28] or understand illustrations of the same category of emotion differently [10]. Emotions categories of all cultures and languages cannot be the same as English emotion categories to describe everyday emotional experiences and perceptions. Therefore, information can sometimes be misportrayed using a different language or interpreting a different language.

Emotions can have an influence on language on different levels of its structure. Starting from the phonological pattern of the language, it can also greatly influence grammar and lexicon during communication beyond how it might be perceived by others [17]. The definite way of arranging words to portray different types of emotions can be observed and learned. The distinct way of word arrangement in discourse can affect the emotions of others. Researching the expanding field of "sentiment analysis" and inspecting the use of natural language during discourse, the subjective state of individuals can be identified [21], [29].

The correlation between linguistic usage and depression was highlighted by Albert Ellis[1] as early as in 1977. Bernard et al. [34] showed proof of depression having a severe effect on speech and language processing. They showed that language used by people with depression tended to be more self-centered and use more first person singular pronouns. Moreover Jarrold et al. [13] claimed that self-centered language usage and increase in negative thinking are effects of depression on language. Trifu et al. [41] provided a more evidence based analysis of linguistic features among depressed people to find out morphological, syntactic and lexical characteristics. They confirmed that the previously mentioned characteristics do persist in such groups. Furthermore, they established that repetition and reverse word order are two markers of such linguistic data.

1.3 Computational learning to detect depression & the difference with traditional classification

The adverse effect of depression on an individual level and public health level is discussed and how much language is impacted by emotions is also shown. This gives us an opportunity to diagnose depression with automated detection using the linguistic inputs from individuals. Although getting enough linguistic data from every one could be an obstacle, social media creates an ample opportunity in this case as a huge chunk of the global population is connected with social media and regularly generates large personal text data. Computational learning techniques and advances in sentiment analysis give us the possibility to create models to detect depression from this textual data. Several pioneering studies have shown the efficacy of computational learning models being used to extract data from social media to predict depressive disorders among individuals [20], [16].
But if depression is identified at a level where it is already causing damaging impacts it might not be that helpful. The linguistic differences already start being created before the onset of depression. This gives us the prospect of early risk detection and consequently early intervention to save valuable lives and also decrease the negative impacts in psychological, social and economic domain. Losada and Crestini [36]were motivated by this idea to create the CLEF eRisk lab task on the early detection of depression. This is different from traditional classification problems as it not only requires to point an example to a class only but also to categorize it early. This calls for new evaluation metrics which were proposed by them [36] and later some other researchers [50]. 3 iterations of the lab and also other independent researchers have created a baseline to work on this problem but the state-of-the-art not being perfect leaves room for more improvement [61].

Chapter 2

Problem Statement

Early risk detection of depression through textual data of social media has been researched to both increase accuracy and decrease latency[20], [50]. Modern transformer based architectures have revolutionized natural language understanding and sentiment analysis[46]. These models have the potential to further increase accuracy and detect depression earlier[62]. As mental illnesses and disorders are stigmatized and there is a huge gap between the prevalence of depression and diagnosis and treatment, automatic early detection can go a long way to address the severity and diagnosis of depression. So, we propose a representational deep learning sequential model to textually detect depression that is on-par with state of the art models for such.

Chapter 3

Research Objectives

This study will investigate ways to detect early risk of depression using computational learning models. To obtain this goal, the research objectives include:

- To understand and compare between different feature extraction techniques to increase correlation of the features to the classification problem
- To compare sequential and non-sequential modeling performance for the task
- To create ways to handle the imbalance of depression corpus

Chapter 4

Literature Review

4.1 Gathering data for experimentation

Data collection from social media for early risk detection of depression is not a particularly easy task. The two main problems are accessibility of data and labeling of the data. Facebook, one of the largest ones, does not allow any such collection of user posted data without user consent. Other than that, twitter restricts the collection of posts per user and also the duration of the collection which limits the context that can be gathered. However, reddit allows some more flexibility here where data can be accessed more flexibly. One other problem is finding out the people who are clinically diagnosed as depressed. Getting a large number of people to testify their clinical diagnosis of depression and share their social media posts in a single common media is thus a hard job to do. A publicly available gold standard data set of social media textual content with depression categorization is yet to be implemented[58].

To our knowledge, a gold standard set was curated by Choudhury et al. [20] which used crowdsourcing to gather a population of a medically diagnosed group of depressed people. Twitter was their source of collection. They used crowdsourcing to test people by both CES D questionnaire[2] and Beck Depression Inventory. Then with their consensual sharing they collected the depressed set of tweets and also collected another control set of tweets.

Other data collection methods have tried to do manual screening by figuring out self-declaration and a second layer of checking by experts. Shared tasks and research competitions have made use of such data. Two of the well known ones are CLPsych 2015 and CLEF eRISk 2017,2018 and 2022. CLPsych 2015 used the dataset curated by Coppersmith et al. [31]. They collected data using twitter API and classified depression, bipolar, PTSD and SAD affected users and a control group by extracting tweets regarding confirmation of being affected and manually screening the affected group by reading if the tweets included testimony of such [27]. Losada and Crestini [36] have collected data from Reddit and labeled it by extraction of self declaration wordings and manual checking which has been the data set used by the CLEF eRisk labs.

Kayalvizhi and Thenmozhi [63] created another such data set by categorizing users in depressed, moderately depressed and not depressed by collecting data from depression related sub reddits and manually classifying them with expert opinions.

So the prevailing trend here is to classify data by looking for self-declaration of users

of being diagnosed with depression and checking by experts. CLEF eRisk labs have curated data in this way and ran research competitions for 5 years culminating in a large body of work using this as a standard.

4.2 Feature extraction techniques

The task of detecting depression from social media texts can be boiled down to a text classification task. For such tasks different statistical, syntactic structure based and representational feature extraction techniques are curated. Moreover, for tasks like this which are close to sentiment analysis, social engagement and writing behavior are also considered.

4.2.1 Social Engagement and Writing Behavior

In one of pioneering works in this field, Choudhury et al.[20] curated many social engagement and writing behavior based features for classification. They considered the qualitative and quantitative nature of engagement of users in social media to create different measures of social media engagements. Other than that, how writing behavior is changed at the onset of depression is also of interest as there is clear correlation between the change and depression. Number of posts each day, ratio of reply posts each day, fraction of retweets each day, fraction of links (urls) shared each day, proportion of question-centric posts from each day are some of the engagement related features extracted by the work. Time of posting and diurnal activities are also investigated by some to indicate engagement tendencies [59]. Textual spreading, time gap between posts, [45] follower/friend ratio, number of followers [38] are some more such features experimented by different researchers. Syntactic structures as features for textual classification is a very common approach. These are somewhat similar to writing behaviors and indicate different grammatical and structural differences in the writing of the two groups. Decrease of third person pronoun usage, increase of first person pronoun usage[20], relation to different parts of speech usage, increase in past tense verb usage, and negation words [38], [35], [37] are the most common such features used in the works related to this.

4.2.2 Statistical Features

Different machine learning algorithms make use of statistical features to perform better in classification tasks. These algorithms make use of the presence of specific words in writing to calculate the measures. Temporal frequency and frequency with respect to different classes of documents are also considered in some of the measures. BoW[37], TVT[47] and TF-IDF [51] are some of the most commonly used ones which have also found their application here. Other than that n-gram based features are very popular as different lexicons are created which perform pretty well in such tasks. LIWC [40], ANEW [20], AFINN [38], LabMT [35] are some of the most popular ones which perform well in sentiment analysis tasks and consequently in depression detection. Other than these, some [20] have created their own depression lexicon to get better results. UMLS metathesaurus is another such statistical tool based on a huge pool of medical records which maps texts to biomedical concepts, which has produced better results than other lexicons [40].

4.2.3 Representational Features

Representational embeddings learned from different machine learning algorithms have dominated recent research in different natural language processing tasks. These algorithms create vector representations of texts which depict the different semantic properties of the text. Word2Vec is one such algorithm which creates vector representations for words based on their meaning. Other than these Doc2Vec is another such embedding algorithm which preserves better contextual information as it produces a vector for a document. Both of the embedding algorithms have shown remarkable success in representing texts for classification tasks. As a result, they were experimented with in this field with success [63]. Certain other embedding techniques are GLoVeText, fastText which can be used to embed textual data and is shown to have good performance by Trotzek et al. [52]. Transformer based architectures have recently revolutionized the field of representation of texts. BERT is one such model which is trained on a large corpus of text, variations of which such as RoBERTa, DeBERTa, DistilBERT are used in depression detection research although with not such great results because of the lack of good amount of data [64].

4.2.4 Other Features

Other less common algorithms used in feature extraction include Latent Semantic Analysis [64] which tries to identify the hidden semantic pattern in a document and Latent Dirichlet Allocation [33] which is used in topic modeling based on a distribution of words.

4.3 Prediction models

Diverse models of statistical machine learning and neural network architectures are implemented throughout the literature that aim to solve the problem of early detection of depression. Even rule based classification modeling has found its way in this task. Sequential Incremental Classification with TVT was a model experimented in 2018's CLEF eRisk Lab [47]. Statistical learning based classifiers have been very common in creating models for depression classification. Naive bayes [47], support vector machine [40], logistic regression [40], random forest [37], k-nearest neighbors [43], all the common classifier models have been used with moderate amounts of success in different experimentations. Moreover, to handle the imbalanced nature of the data set and to get better results from weak learners, adaptive boosting, extreme gradient boosting have also been tried. Moreover, ensemble learning methods were used which garnered results a bit better than single classifier models. Neural network architecture with deep learning capabilities have dominated the field of natural language processing in recent years. Sequential modeling of posts from social media helps to get the continuous stream of indications from previous and current posts. In this way, the continuous process towards depression can be modeled to get better results. Thus, LSTM [49] and GRU [40] have been used many times with good results. CNN has also been applied that generated results better than recurrent neural network based ones in a certain study [48]. Although the number of posts to train or run on, can be an obstacle as these models can not process or retain the

memory of too long sequences. Concatenation strategies and threshold scheduling are used to tackle this problem. Pre trained transformers have also been really good in language related tasks which can be downstreamed to classification tasks as well. Different variations of BERT such as RoBERTa, DeBERTa, BioBERT and DistilBERT [64] have been experimented with, although the lack of enough data did not result in much greater successes. Other than these, some custom architectures have been developed for this task which have produced good results. Some of them are Time-Aware Affective Memories Network [60] and t-SS3 [53].

4.4 Evaluation Criteria

To measure the goodness of any classification model an evaluation metrics is needed by which different models can be compared. The task at hand has two main indices of evaluation. How accurately the model predicts depression before the onset of it and how fast it can predict so. For the first one, a precision or accuracy can seem to be enough but it fails to depict how much the model predicted correctly of the correct cases and how many negative cases were predicted as positive. To solve this issue, F1 score is a very good measure which embodies these features.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} = \frac{2 * TP}{2 * TP + FP + FN} \tag{4.4}$$

Losada et al. [36] suggested a new metric named ERDE(Early Risk Detection Error) specifically for early risk detection tasks.

$$ERDE_0(d, k) = \begin{cases} cf_p & \text{if d = positive AND ground truth = negative} \\ cf_n & \text{if d = negative AND ground truth = positive} \\ lc_o(k).ct_p & \text{if d = positive AND ground truth = positive} \\ 0 & \text{if if d = negative AND ground truth = negative} \end{cases} \tag{4.5}$$

where $lc_o(k)$ is the latency cost function,

$$lc_o(k) = 1 - \frac{1}{1 + \exp^{k-o}} \tag{4.6}$$

This function was only used in true positives to penalize latency because the cost of latency only matters in cases where the predicted result was correct that the person was depressed.

Sadeque et al. [50] suggested a new metric called latency weighted F1 score which better penalizes the latency and embodies the precision and recall dichotomy for prediction model scores.

$$\text{latency}(U, \text{sys}) = \text{median}\,\{t \mid u \in U \wedge \text{ref}(u) = +, \text{time}(\text{sys}, u)\} \tag{4.7}$$

$$P_{latency}(u, \text{sys}) = -1 + \frac{2}{1 + e^{-p \cdot (\text{time}(u,\text{sys}) - 1)}} \tag{4.8}$$

$$F_{latency}(U, \text{sys}) = F_1(U, \text{sys}) \cdot (1 - \text{median}\,\{\text{Platency}(u, \text{sys}) \mid u \in U \wedge \text{ref}(u) = +\}) \tag{4.9}$$

where U is the set of persons, ref(u) is the correct category (positive or negative) assigned to the person, sys(u) is the model's earliest confirmed prediction and p defines how quickly the penalty should increase.

4.5 Leveraging Advances in NLP

Certain advances in natural language understanding and language modeling can help us in our task. We hope to leverage these advances in our experimentation to create a better performing model.

4.5.1 Bidirectional approach

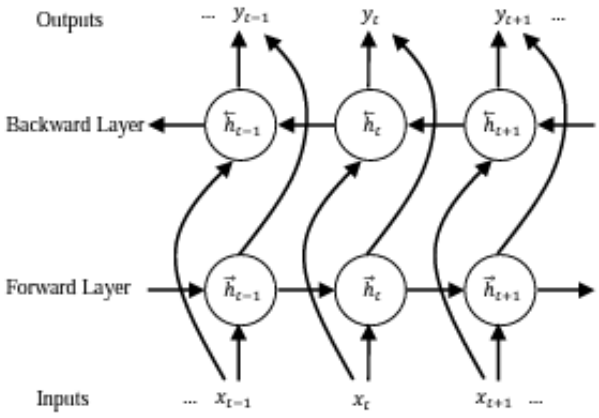


Figure 4.1: Bidirectional approach [39]

Natural language understanding has approached understanding the semantics and structure of a sentence from start to finish approach before. Now, the bidirectional approach to language modeling is becoming more common as it gives the model more information about language understanding [39]. Bidirectional recurrent modeling is dominating the field now, with newer architectures which process the whole sentence as a whole proving to be more powerful [42]. This allows the model to understand the relationship between the words in a sentence and better capture the contextual information.

4.5.2 Attention mechanism & transformers

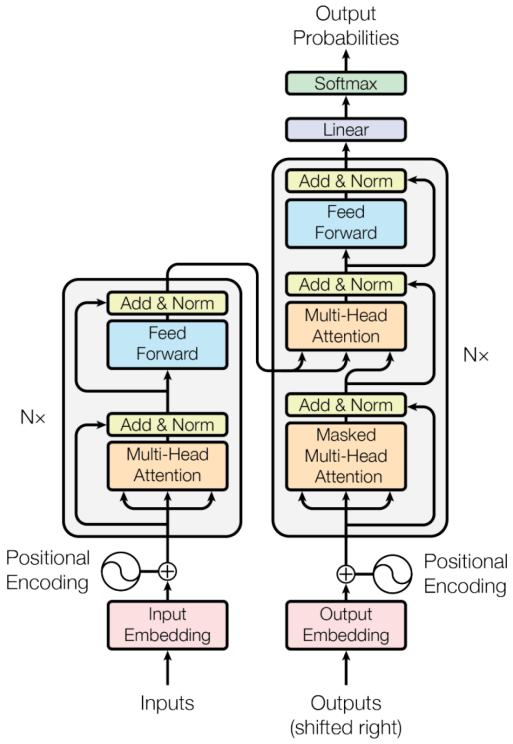


Figure 4.2: Architecture of Transfomer[42]

Attention mechanism is a system to give different parts of the input different weights to better represent their importance in the meaning. This mechanism has brought about a revolution in natural language understanding as it can better retain and give importance to contextual information and local and global dependencies. Self-attention is a type of attention that allows the system to attend to different parts of the same input sequence and as a result capture their interdependencies. Self-attention from different input sequences are organized into different sets and then concatenated and averaged to get the dependencies within the different sets. This mechanism is called multi-head attention. Along with these, positional embedding is applied in transformers that accounts for the positions of textual elements in their meaning.

4.5.3 Pretraining and transfer learning

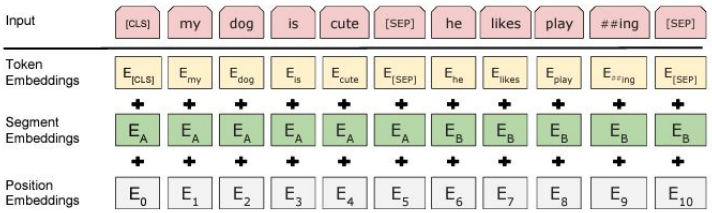


Figure 4.3: Input representation in BERT[46]

These natural language models, although continuously performing better than ever before, require large amounts of data to train on to achieve these results. This makes it infeasible for researchers all over the world to make use of such advances. But pre-training and transfer learning provide a solution in this regard. Organizations with large data retaining and huge training capabilities can train these architectures and save weights so that the same models with the same weight can be used by other researchers without having to redo the whole process. BERT [46] is one such pre-trained model which has achieved close to state-of-the-art benchmarks on many natural language tasks. Using this through transfer learning can help in our research.

4.5.4 Semantic similarity matching

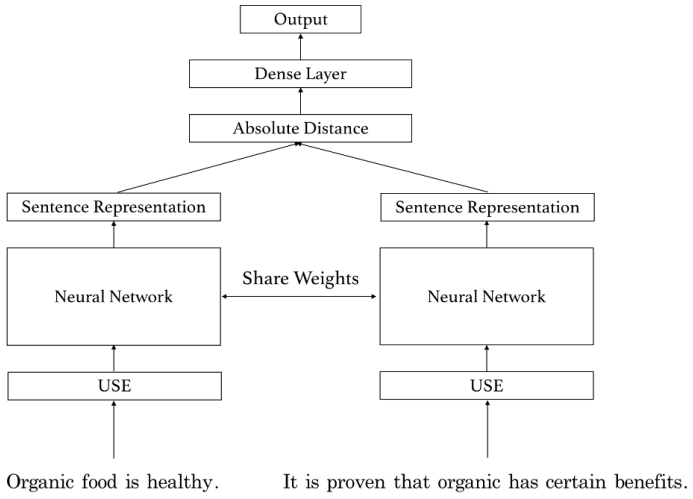


Figure 4.4: Architecture of Siamese network

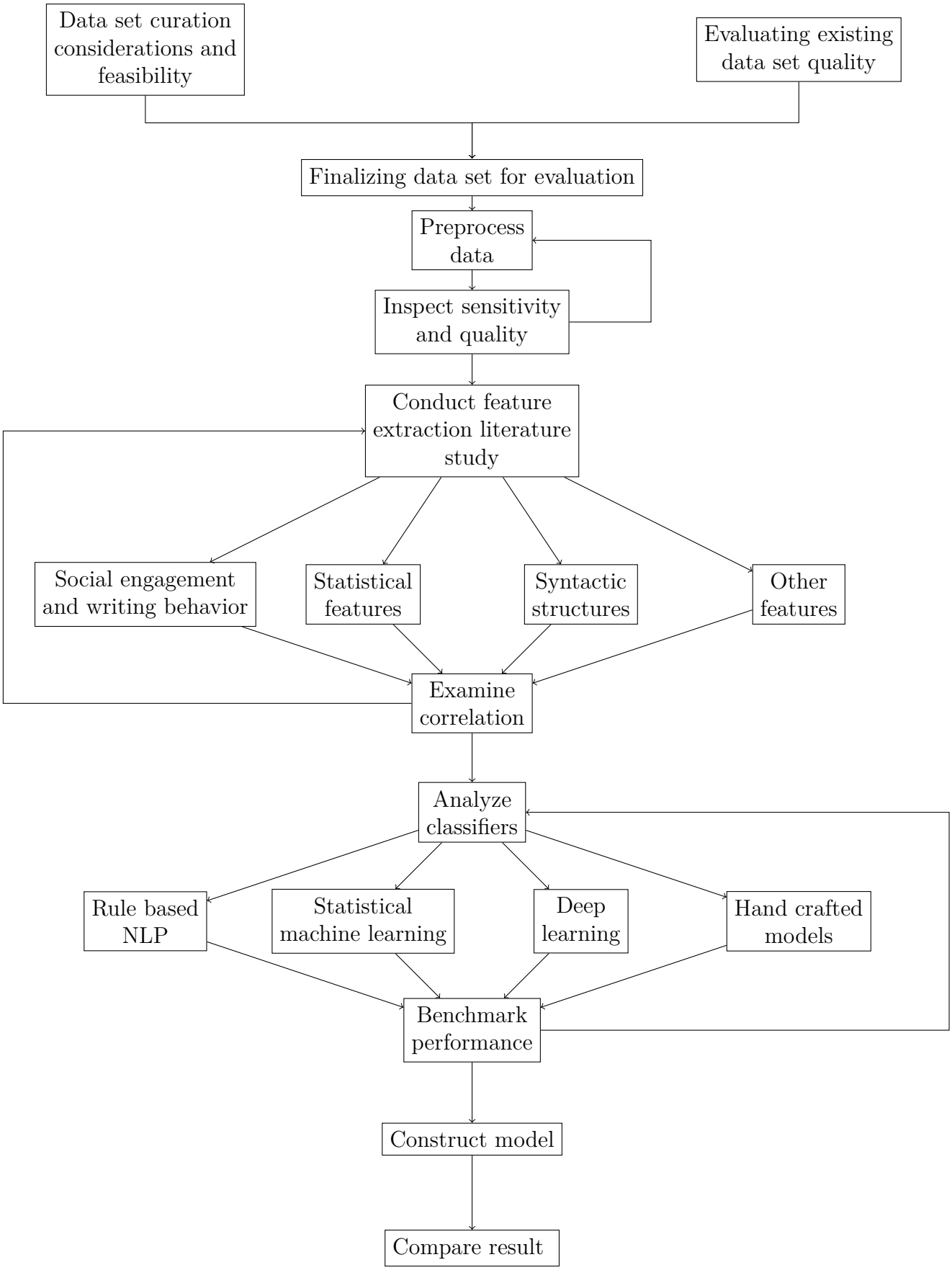
Semantic similarity matching is a way to find semantic similarity between two pieces of texts. This can be a powerful tool for tasks like sentiment analysis and depression classification. But the main problem of using this is the cost of training. Siamese network [4] is a type of architecture specially proposed for this which solves the problem. This network shares the weight between the two parts of the network and calculates similarity index for the opposing classes. Using triplet loss it minimizes the training cost thus making this a viable tool.

Chapter 5

Working Plan

The aim of our research is to construct and analyze different efficient models to detect depression among users automatically and early using social media text datasets. To achieve this we plan to follow the below procedure:
We first wanted to consider the feasibility of curating one dataset ourselves and also to benchmark other available ones.
After deciding on it, we experiment with several preprocessing techniques and the resulting data is inspected to see how much quality is retained.
Then we find out different feature extraction methods in text classification from existing literature and measure their correlation to the task to finalize on the ones to use. Social engagements and writing behavior came out as a common theme for feature extraction such as tracking post frequency, linguistic style and emotional expression in online interactions. Statistical features such as word frequency, sentiment analysis, and topic modeling were also used for text classification pattern identification. Representational features were also used to transform raw data into structured vector representations, using essential semantic characteristics in text. Other feature types were also explored for their valuable information such as LSA and LDA.
After that, the most important part, experimenting with different classifier models to decide upon which ones to use for better performance. After studying the literature to find out classifiers that perform well in this task, we will perform benchmarking to evaulate which classifiers have more potential to get better.
Rule-based Natural Language Processing (NLP) techniques were applied, using predefined rules to identify depression-related patterns or keywords. Statistical machine learning methods, such as Support Vector Machines, Logistic Regression, and Random Forest, demonstrated effectiveness in making predictions. Deep learning approaches like LSTM, GRU, and CNN outperformed statistical modeling approaches, and custom handcrafted architectures like Time Aware Affective Memories and t-SS3 were also developed, which showed great results.

Figure 5.1: Working plan



Chapter 6

Data

6.1 Data Collection Considerations

The data set we have used, curated by Losada and Crestani [36], have carefully selected sources to create a corpus centered on language usage and depression. They analyzed a number of variables, including the quality and quantity of the data. They also took into account the accessibility of lengthy connection histories. Other than that it was also taken into consideration the difficulty in distinguishing between depressed and non-depressed instances, as well as terms for data distribution.
Twitter is a very popular social media platform among the data sources that the creators have analyzed to collect data from. However,there are problems with the tweet's inability to adequately provide information in the context of depression. In addition to that the twitters constantly changing environment proved it difficult to get a complete record regarding the interactions of the user. Because permission is only given to attain a certain quantity of past tweets it becomes harder. Redistribution of the information on Twitter is also very severely constrained.
Another consideration was MTV's A Thin Line (ATL), which is a social media network created to address digital misuse among problematic youth. However, the limitations on the data redistribution and the absence of traceable identifiers prior to contact history made it difficult. The dataset that was acquired from ATL, included the personal tales, evaluations, comments etc. But, it was not appropriate for use in benchmark creation.
To get around these restrictions and limitations, Reddit was chosen to be the data source. Because, it is an open source website with sizable and diverse users where users participate in conversations and contribute materials to Reddit. Reddit provides detailed entry histories spanning many years in the past and it provides insightful information on a wide range of medical difficulties including depression among people. It is noteworthy to point out that Reddit's terms and conditions allow the usage of its information for research studies.
The focus was on spotting individuals with depression during the data collecting phase. A modified technique was suggested by Coppersmith [27] to use self-reporting to get the labels. Using that, specific searches were conducted on targeting self-expressions of depression diagnoses. It was examined with carefulness through matching posts, taking into considerations of elements such as length and explicitness and verifying the authenticity of the diagnoses that were done. The trust on the quality of the examinations was high enough. This trust was particularly high for

posts that have originated from depressive subreddits that acts as a support group for individuals who are struggling with their depression.
In addition to the depression group, a control group was formed to make the collection of the data classifiable. For this, random redditors and individuals who are interested in depression related subreddits but not suffering were chosen to make up the control group. They might become interested sometime because someone they knew suffered from depression. Although there might be some chances of having few depressed cases in the control group, it is estimated to have negligible effect on the result.
Overall, the data collecting strategy involved cautious data selection approaches utilizing Reddit to be the crucial platform for collecting meaningful data on depression as well as implementing manual review method for recognizing users with cases of depression. Consequently , this method resulted in collecting a large and diversified dataset which will also be very useful for future research and analysis.

6.2 Description of Data

The text-based dataset was obtained from Reddit with an interest in the language used by users who are depressed and are not depressed. This data was gathered using Reddit's API, which has a cap of 1000 comments and 1000 posts for each Reddit user. Consequently, the dataset contains a maximum of 2000 records from active Reddit users. The data contains comments and posts made on different subsidiary threads within Reddit.
In the "Train" dataset and the "Test" dataset, there are about 83 and 54 users in the depressed category, respectively. However, the non-depressed category consists of 352 users from the "Test" dataset and 403 users from the "Train" dataset section. The dataset consists of entries taken from an extensive number of subreddits, including information on videos,news, cuisine, and many more. The data is organized in a temporal arrangement where it is possible to observe the change in language over time. The average duration in days between the first and last submissions between depressed trained data is 572.22 days, whereas for depressed test data it is 586.31 days, covering a large amount of time for most Reddit users.
For the depressed user category, the dataset contains a total of 30,851 number of posts for test data and 18,729 number of posts for train data. However, for the non-depressed user category, the dataset contains a total of 2,64,172 number of posts for training data and 18,729 number of posts for testing data. Depressed category Reddit users have an average number of 371.70 and 346.83 posts for the training and testing sections, respectively. In addition, non-depressed category Reddit users have an average of 655.51 and 618.47 posts for the training and testing sections, respectively. The data displays an increased amount of participation in Reddit activities from non-depressed users.
Another insight looking into the language patterns can be the average count of words in each post. For the training sections, the average count of words in posts was 37.3 for the depressed category and 23.36 for the non-depressed category. Similarly, in testing sections, the average count of words in posts was 38.31 for depressed category and 27.04 for the non-depressed category. It is important to keep in mind that the increased word count is observed in depressed users in comparison to non-depressed users.

Table 6.1: Data description

Category	Depressed		Non-depressed	
	Train	Test	Train	Test
Total number of users	83	54	403	352
Total number of posts	30,851	18,729	2,64,172	2,17,701
Average number of posts	371.70	346.83	655.51	618.47
Average word count	37.73	38.31	23.36	27.04
Average days between first and last submissions	572.22 days	586.31 days	626.17 days	623.40 days

The dataset was curated by organizing in a series of XML files each of which is represented by a Reddit user. The entries of Reddit users are saved in chronological order, including each entry's title, content, and timestamp recorded separately. No further information or metadata are recorded in the entries.
Many scenarios were observed when it came to the timing of diagnosis. Some of the time, the recentness of the diagnosis is expressed by words such as "yesterday" or "this week". However, in such situations, the majority of the conversations collected occurred in the time preceding the diagnosis. Moreover, there are cases of diagnoses being made a long time ago, such as "in 2010" or "3 years before". As a result, the majority of Reddit user's content reflects their experiences after the diagnosis. In certain cases, the recovered information might include inputs from both before and after diagnosis. Despite containing some confusion regarding the accuracy of data diagnosis done, the rough estimation of the data is still noteworthy and may be utilized in many ways.

6.3 Data Pre-Processing, Cleaning

User submission data files were in XML format consisting of positive and negative chunks for the depressed and non -depressed users which indivicually were converted to Comma-separated values(CSV) files and then converted to panda dataframes. The two dataframes for the depressed and non-depressed is then finally merged into a single dataframe.

- Null value checking and removing:
- Contraction Removing:
- Lowercasing, Digit, Punctuation and White Space Removing:
- Lemmatizing and Stop Word Removal:
- Document Term Matrix Generation:

Chapter 7

Feature Extraction

For our experiments, we made use of some of the best performing feature extraction techniques for this task and natural language processing in general. More than a decade of work on this starting from Choudhury et al.'s [20] seminal work have made use of different assortments of feature extraction techniques as mentioned in the literature review. We made use of the following techniques for our work.

7.1 UMLS Metamap

The National Library of Medicine of the USA developed the Unified Medical Language System as a semantic mapping between medical-clinical texts and the relevant healthcare terminologies and ontologies. It uses a rich pool of data from medical records and different computational linguistic techniques to create the mapping. Sadeque et al. [40] first used this in depression classification tasks and has since been used in such. It has been popular since it has been close to top performance in CLEF eRisk Labs task. The metamap API [11] takes in texts and generates concept unique identifiers which are labels for biomedical concepts. We used only one source (SNOMED CT-US) and two semantic types (Mental or Behavioral Dysfunction, Clinical Drugs) as suggested by Sadeque et al.[40] From our experiments we got a set of 251 CUIs relevant for our task.

7.2 Doc2Vec embeddings

Embedding vectors are a common and successful method to extract features from textual data for classification and other tasks. Word2Vec embeddings have been very common in text classification tasks and have shown good results in CLEF eRisk labs. But for understanding the contextual meaning of the reddit posts, word2vec embeddings are not enough.
Mikolov and Lei [30] showed that paragraph vectors can better capture the contextual semantics in paragraphs. The vector presentations are learned to predict the surrounding words in a paragraph. In the paragraph-vector distributed-memory model, a softmax multiclass classifier is used to predict words in a paragraph. Every paragraph is a unique vector in matrix D, each word in that paragraph is a unique vector in matrix W. These two are averaged or concatenated to predict the next word. By maximizing the average log probability the predictions are calculated.

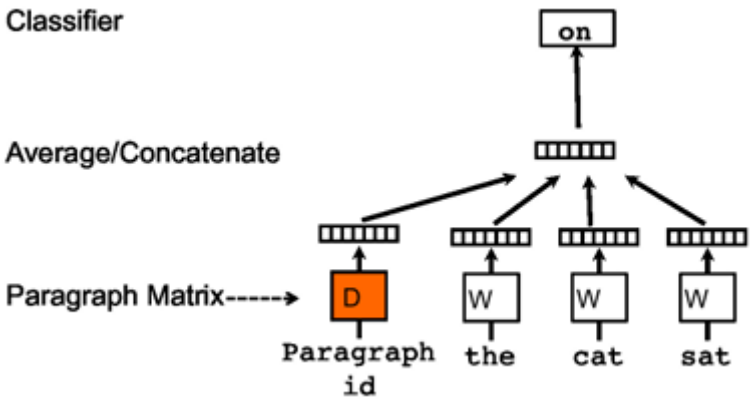


Figure 7.1: PV-DM in Doc2Vec[30]

Their results pointed out that such vector representations perform well on-par with other state-of-the-art text classification techniques. We used the gensim [12] implementation of doc2vec. For our experiments, we fine tuned the paragraph-vector distributed-memory algorithm with our training data to better fit our cases.

7.3 SBERT

BERT (Bidirectional Encoder Representation from Transformers) [46] is one of the most powerful techniques for generating representations of natural language. The core idea behind language modeling in BERT was to make it able to predict words in a sentence and generate representations for words as a result. Although many embedding techniques make use of this approach one main difference was in directionality. Rather than traversing a sentence from start or end to predict the missing word/words, BERT processes the whole sentence at once to understand the context around the missing one(s), which gives the model its bidirectional name. It uses positional embeddings of words to understand their positions in texts and uses a multihead attention layer to calculate the importance of words with respect to each other. This mechanism makes the model able to understand context and importance of different parts of the text.

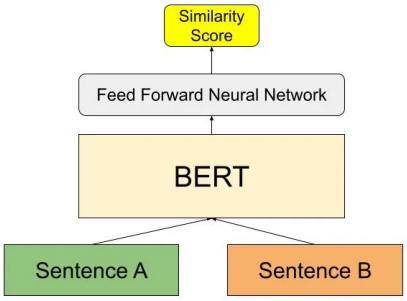


Figure 7.2: Semantic similarity using BERT

Now for our task at hand we need to get representations for post of users to classify it into depressed and non-depressed categories. To get sentence/post level embeddings from BERT is a complex task as it generates word level representations. Moreover,

if we model the task as semantic matching between posts then BERT becomes an infeasible model to work on. It is shown that BERT performs worse than GloVe which is an word2vec model for semantic matching. Semantic matching in BERT requires it to feed every sentence/post with every other to generate similarity scores. So for n posts, it would need n*(n-1) number of computations to generate that. This is where SBERT comes in.

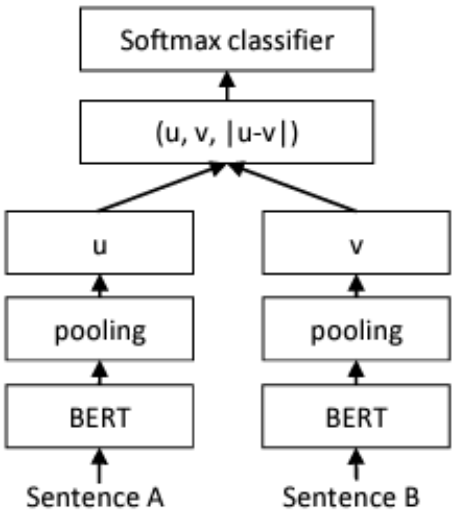


Figure 7.3: Architecture of SBERT [54]

SBERT makes use of both the bidirectional and attention mechanism of bert, and the Siamese network to share weight upgrading to make the task more efficient [54]. It uses two networks of BERT similar to before to find the similarity. The key difference here is that it uses triplet loss to update the weights. Which means it first gets the similarity score by comparing two sentences/posts from two classes with another sentence/post and then uses this score to calculate losses for every other sentence/post. In this way, it does not need to run computation on each of the post pairs. This model has shown to achieve greater performance in similarity matching. Making these models useful in different tasks requires training them over large sets of data. Both pretrained versions of BERT and SBERT are available which are trained on enormous datasets to achieve a very high level of language understanding. We have used all-mpnet-base-v2 pretrained SBERT model from huggingface [55].

Chapter 8

Classifiers

For our experiments, we used three models, one from statistical machine learning and two other from sequential neural networks. Here we present the description of how these models function to classify the texts.

8.1 Support Vector Machine

In SVM[5], the object is to be classified as a coordinate point in N-dimensional space and the coordinates are normally known as features. SVM executes classification tests by drawing hyperplanes in such a way that all points of both categories are separated completely. However, there can be multiples such hyperplanes and SVM tries to find the one that best separates the two categories in a way that maximizes the distance between points in either category . This distance is known as margin and the points that fall on the margin are called the supporting vectors. SVM optimizes in a way that maximizes the distance between the hyperplane and the classes thus generating a better classification.

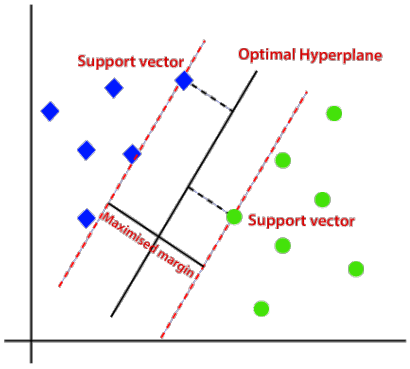


Figure 8.1: Linear Support Vector Machine

The base support vector machine is a linear classifier that takes into input the features and classifies them as belonging to one of the classes. This is done by calculating a hyperplane between the classes that maximize the distance between the dividing plane and the classes. For example , if the inputs belong to two classes and the value of the inputs are plotted on the graph we can see that a plane(line) can divide the two classes in a manner that separates the space. Now if a new input is fitted into this, the dividing line will ensure that the input falls in one of the two

divided spaces. This plane or dividing line is called the hyperplane which divides the classes. The points of each class which are closest to the hyperplane are called support vectors. The distance between the support vector and the hyperplane is called the margin. The idea is to maximize the margin so that the classifier can clearly identify which classes the new inputs belong to. The decision function is the function that decides which class an input belongs to.

$$f(x) = w^T \cdot x + b \tag{8.1}$$

Here x is the input, W contains the weights and b is the bias. While training, W and b are calculated to maximize the margin. The optimization function for this is defined as,

$$\begin{aligned} \min_{w,b,\xi} \quad & \frac{1}{2}||w||^2 + C\sum_i \xi_i \\ \text{subject to:} \quad & y_i(w^T x_i + b) \geq 1 - \xi_i \quad \forall \text{ training samples } (x_i, y_i) \\ & \xi_i \geq 0 \quad \forall \text{ training samples } (x_i, y_i) \end{aligned} \tag{8.2}$$

where C is the cost parameter which manages the balance between large margin and for some examples to violate the margin and ξ is the slack parameter for soft margin.
Now for complex tasks like sentiment analysis and text classification this linear separation of classes does not capture the complicated nature of the task, consequently producing poor results. This is where the kernel trick comes into play. It allows the input to be mapped to a higher dimensional space so that margin optimization can be applied to achieve a better decision boundary. Radial basis function (RBF) is one such kernel which uses a gaussian kernel function to get similarity between the new input and another belonging to a class to classify the input. The RBF kernel equation which calculates the similarity is,

$$K(x_i, x_j) = \exp(-\gamma \cdot ||x_i - x_j||^2) \tag{8.3}$$

where γ is a parameter that manages the width of the gaussian function. This kernel function is then used to generate the decision function,

$$f(x) = \sum \alpha_i \cdot y_i \cdot K(x_i, x) + b \tag{8.4}$$

where α is the lagrange multiplier. The changed optimization function for this kernel is given as follows,

$$\begin{aligned} \min_{\boldsymbol{\alpha}} \quad & \frac{1}{2}\sum \alpha_i \alpha_j y_i y_j K(x_i, x_j) - \sum \alpha_i \\ \text{subject to:} \quad & 0 \leq \alpha_i \leq C \quad \text{for all } i \\ & \sum \alpha_i y_i = 0 \end{aligned} \tag{8.5}$$

With the help of this kernel, complex relationships can be captured to map the inputs in higher dimensions to get improved classification results in cases such as ours.

8.2 Long Short Term Memory

Statistical machine learning algorithms and feed forward neural networks do not retain the information from different timesteps and do not process data as a sequence or stream. Tasks like ours which can perform better if the sequential nature is taken into account thus require a different type of modeling. Recurrent neural networks are such architectures which can treat the data set as a sequence to understand the development towards depression. But one main problem with basic recurrent neural networks is that it has the vanishing or exploding gradient problem. Like other learning models, recurrent neural networks calculate losses and update weights by back propagation. But while calculating the gradient by backpropagation through time, if the values are too small, the gradients can vanish or get to very small values, which means information from previous steps can be lost .

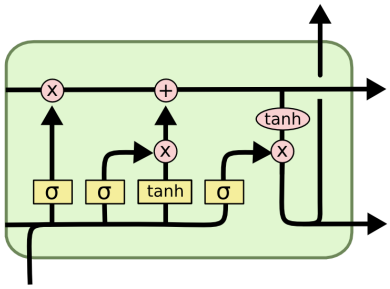


Figure 8.2: Architecture of Long Short Term Memory

Long short term memory proposes to solve this problem by using gates. Gates are information capture mechanisms which better retain the important information from previous steps and forget the information that is less important. LSTM uses three gates to process the information. It has a cell state which passes on the memory from before.

8.2.1 Forget gate

This gate decides which information from previous steps to get rid of.

8.2.2 Input gate

This gate decides the information with which to update the cell state

8.2.3 Output gate

This gate calculates the next hidden state.
The calculation of the mechanism are as follows,

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{8.6}$$

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{8.7}$$

$$\tilde{C}t = \tanh(W_C \cdot [ht - 1, x_t] + b_C) \tag{8.8}$$

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \tag{8.9}$$

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{8.10}$$

$$h_t = o_t \cdot \tanh(C_t) \tag{8.11}$$

In these equations, x_t represents the input at time step t, h_t represents the hidden state at time step t, C_t represents the cell state at time step t, σ represents the sigmoid activation function, and tanh represents the hyperbolic tangent activation function. W_f, W_i, W_C, W_o are weight matrices for the respective gates and candidate activation, while b_f, b_i, b_C, b_o are bias vectors. The dot symbol represents the dot product between vectors, and [a,b] represents the concatenation of vectors a and b.

8.3 Gated Recurrent Units

Gated recurrent unit [26] is another gated version of recurrent neural networks which can preserve information from previous sequences to better predict the class. It has two gates, update gate and reset gate to preserve some information from before and update new ones from the current time step. As it uses fewer gates, it requires fewer tensor operations and calculations making this a bit faster model than LSTM. Between the two gated versions of RNN, which one is better is up for debate, as both perform similarly or perform better than the other one in specific cases.

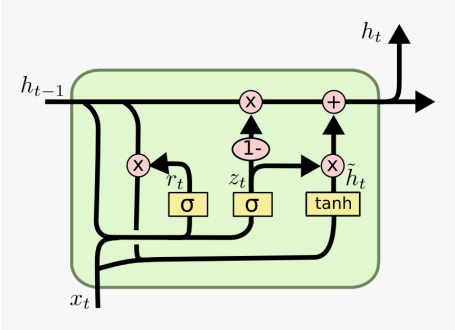


Figure 8.3: Architecture of Gated Recurrent Unit

8.3.1 Update Gate

This gate decides what information to get rid of from previous steps and what new information to add from the current step. The calculation for this gate is done by,

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{8.12}$$

where x_t is the current step input, h_{t-1}) is the information from previous steps, W_z is the weight matrix and z_t the output of this gate. Sigmoid function is used to calculate the output.

8.3.2 Reset Gate

The reset gate is another gate which picks out what information to forget from previous steps. The calculation for this gate is done by,

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{8.13}$$

where x_t and h_{t-1}) are similar, W_r is the weight matrix and r_t the output of this gate. Sigmoid function is used to calculate the output.
The outputs of these functions are then used to calculate the activation or hidden state which retains this information.

$$\tilde{h}_t = \tanh(W_h \cdot [r_t \cdot h_{t-1}, x_t]) \tag{8.14}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \tag{8.15}$$

where $\tilde{h}t$ and h_t are candidate activation and activation/hidden state respectively and W_h is the respective weight matrix.

Chapter 9

Experimental Setup

For the experiments of our research, we tested multiple feature extraction techniques as found in the literature and also multiple succeeding learning models. In our experiments, we found UMLS metamap, Doc2Vec embedding and SBERT embedding to be the most relevant feature extraction techniques. SVM with Metamap features were chosen as it was one of the best performing models in the 2017 version of CLEF eRisk labs. We put our belief in sequential modeling, consequently we tried the two memory retaining recurrent neural networks, LSTM and GRU.

9.1 Model 1 (Metamap-SVM)

From our experiments with the train set and Metamap, we got 251 concept unique identifiers which correlate to the classification task. We used the RBF kernel of SVM from scikit learn library to better model the complex relationship of non linearly separable features. To deal with the imbalanced dataset, we used weighting by class weights calculated from the previously mentioned library. We generated the probability at each post and experimented with different criteria to early predict the onset of depression.

9.2 Model 2 (Doc2Vec-BiLSTM)

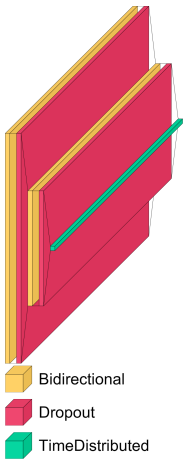


Figure 9.1: Architecture of Model 2

In this model, we fine tuned the Doc2Vec model(PV-DM) available on gensim [12] with our test data to better represent our case. Then we generated 100 length vector embeddings to feed into the model sequentially for each user. We took 411 user posts to create the sequence to feed into the system as memory retention of longer sequences is not that well for LSTM. The number 411 was chosen as it was the average number of posts in the train set. The classifier used was a Bidirectional LSTM to better integrate contexts from both ends. Two layers of BiLSTM with 128 and 64 units respectively were used. A dropout of 0.2 was used after both of the layers to prevent overfitting. The dense layer was time distributed to generate probabilities at each time step. To address the data imbalance issue we used a custom written loss function which is described in the next sub section.

9.3 Model 3 (SBERT-BiGRU)

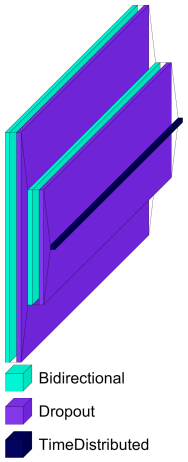


Figure 9.2: Architecture of Model 3

For the last model, we generated SBERT embeddings using the huggingface transformers library [56]. The embedding vectors were of length 768. Then the embeddings were fed into the model as sequences of length 411 for similar reasons stated above, for each user. The classification model was bidirectional GRU to retain contexts from both ends. It had similar configurations, 128 and 64 units in two layers. For the overfitting problem 0.2 dropout was used after both of the layers. Time distributed dense layer was used to get probabilities for each post. Similar to Model 2, custom written loss function was used to handle the imbalance issue in the dataset.

9.4 Handling Imbalanced Data:Custom Loss Function

The data set used for the experiments and benchmarking is an imbalanced one where the ratio of depressed posts to non-depressed ones is close to 1:5. This matches with the real word scenario as there are more people not affected by depression then there are the affected ones. But this presents a problem while training models as this lower portion leads to models tagging most or all of the ones as depressed to decrease loss.

In our experiments, we fitted the support vector machine with class weights calculated from the library. But for the neural network architectures as the input sequence are 3D tensors with shape (number of users, sequence size, embedding vector size), tensorflow does not support any kind of fitting with weights. So we found our way by creating a custom loss function to downplay the loss incurred from the non depressed class and to prioritize more on the loss from the depressed class.
We changed the categorical cross entropy to adjust it to the weights. The weights for the classes were calculated as before in the case for SVM. Now a standard loss function of categorical cross entropy is calculated by,

$$L(y_{\text{true}}, y_{\text{pred}}) = -\sum (y_{\text{true}} \cdot \log(y_{\text{pred}})) \tag{9.1}$$

where the variables y_{true} and y_{pred} represent the true binary label and the predicted probability of the positive class, respectively. Logarithm of the prediction probability is used for numerical stability and information compression. This calculated loss gives all the categories the same weight. We multiplied this loss with a weight vector, so when loss of each case is calculated, the loss from the depressed class gets more importance in the training. So the changed loss function is,

$$L(y_{\text{true}}, y_{\text{pred}}) = -\sum (w \cdot y_{\text{true}} \cdot \log(y_{\text{pred}})) \tag{9.2}$$

where w is the weight vector holding the weights for the classes.

Chapter 10

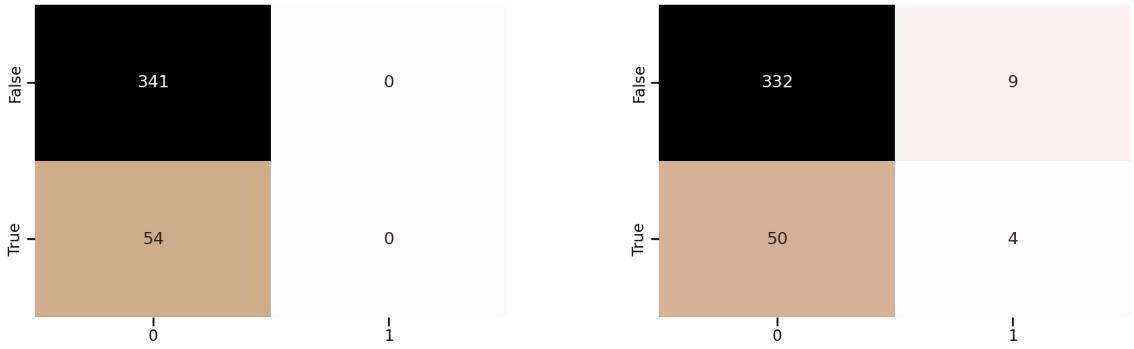
Results and Analysis

10.1 Experimental Reasoning for the Custom Loss Function

Now as our data set is largely imbalanced, we need a way to calculate the loss incorporating that factor. If not, the models will tend to classify all the data into the larger class which is the non-depressed one to decrease the loss.
In the first model, this was done by fitting the model with class-weights. But, the sequential models with 3D tensors do not have any built in method to do that. For these reasons we had to curate a custom loss function that caters to our needs.
We adapted the cross entropy loss function. Mathematically, it is expressed as the negative sum of the product of the true label and the logarithm of the predicted probability, serving to penalize larger deviations from the true label and encourage the model to improve its predictions over time. Now this loss function will penalize the deviation from the larger class as a result making a system that is not of much use to us.
So we multiplied the loss function with a weight matrix which weights the calculated loss according to the class's weight in the dataset. This allows us to better incorporate the loss of deviating from the depressed class.
We used the custom loss function in model 2 and 3. The performance difference achieved is reflected in the experimental results. Model 2 achieved significant differences, as the model with traditional loss function classified all the data as non-depressed.
Although model 3 with cross entropy loss function does not seem the achieve much differences in precision, recall and F1, it gets some gain in the latency part as it produces lower ERDE scores. The model 3 scores were achieved with higher weight values. We experimented both the models, with lower and higher weights than the actual weights in the class and these came out to be the optimum values for which the scores could be maximized.
Also, we can see the impact from the training and validation accuracies of the models. The model 2 basically stabilized to a version where all of the data was classified to the non-depressed class. But in model 3 we can see with training epochs the validation accuracy kept decreasing, but in the increased weight custom loss function version, the validation accuracy also increased with epochs.

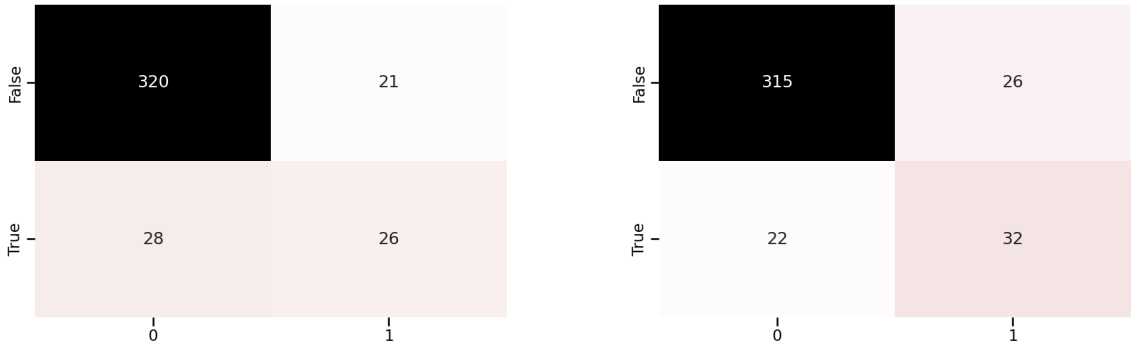
Table 10.1: Comparison of Loss Function

Model	Precision	Recall	F1	ERDE5	ERDE50	Latency Weighted F1
Model 2 Doc2Vec BiLSTM-Without Custom Loss	0.43	0.50	0.46	0.14	0.14	0
Model 2 Doc2Vec BiLSTM-With Custom Loss	0.62	0.69	0.64	0.08	0.08	0.64
Model 3 SBERT BiGRU-Without Custom Loss	0.73	0.72	0.72	0.08	0.08	0.73
Model 3 SBERT BiGRU-With Custom Loss (Increased Weight)	0.74	0.76	0.75	0.06	0.06	0.75



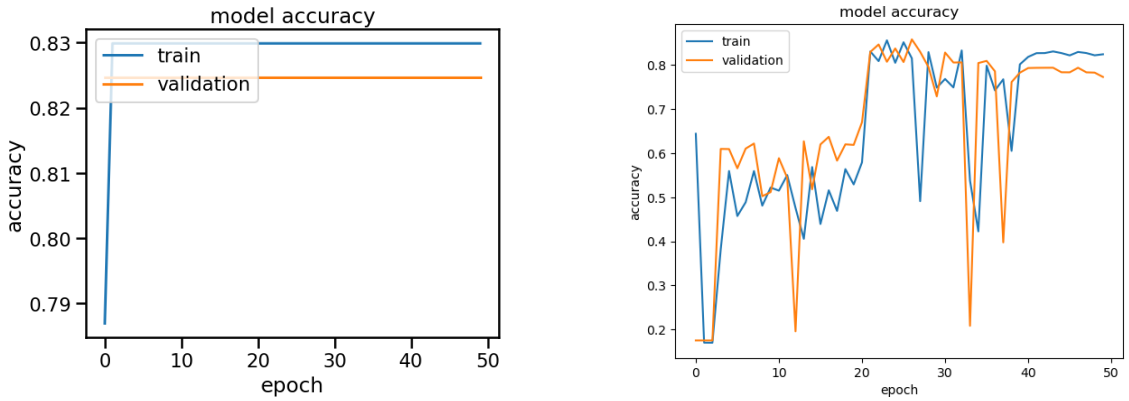
(a) Without Custom Loss Function (b) With Custom Loss Function

Figure 10.1: Confusion Matrix of Model 2



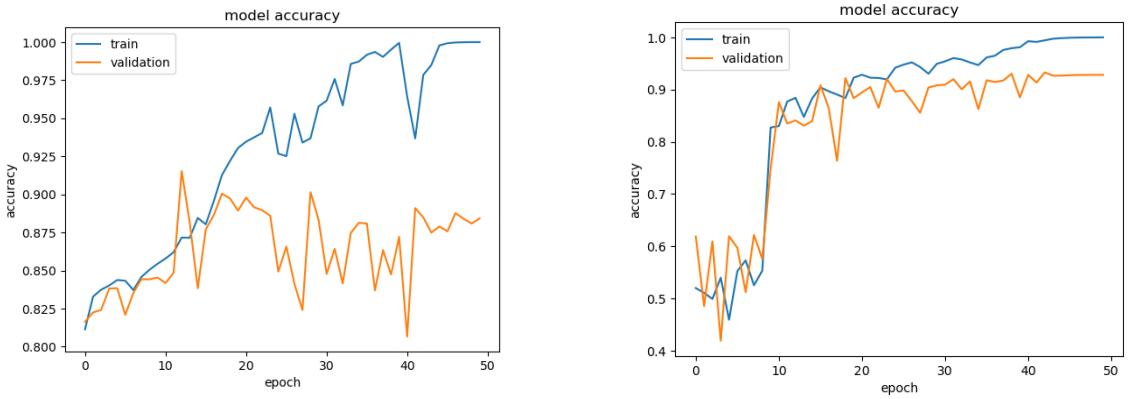
(a) Without Custom Loss Function (b) With Custom Loss Function

Figure 10.2: Confusion Matrix of Model 3



(a) Without Custom Loss Function (b) With Custom Loss Function

Figure 10.3: Training & Validation Accuracy of Model 2



(a) Without Custom Loss Function (b) With Custom Loss Function

Figure 10.4: Training & Validation Accuracy of Model 3

10.2 Comparison with SOTA

The results of the experiments along with the best achieved results till now are shown in Table 10.1. The best results are collected from CLEF eRisk labs, where early detection of depression was a shared task in 2017,2018 and 2022.
The names for the best models indicate the team name in the participating year, submission/run number and the year they participated. It also includes the features and models that were used to create the system.
Performance is measured in 6 metrics namely precision, recall, F1, $ERDE_5$, $ERDE_{50}$ and latency weighted F1. The higher the precision, recall and F1 scores the better the classification task was performed. To account for the early detection success $ERDE_5$, $ERDE_{50}$ and latency weighted F1 scores are used. $ERDE_5$ and $ERDE_{50}$ scores are better when lower and latency weighted F1 score is better when higher.
We used Metamap SVM as a baseline to model to compare with as it was one of best performing ones in the 2017 version of CLEF eRisk labs. As we can see both our models have performed better than the model in all the metrics.
Model 3 (SBERT-BiGRU) has a precision score better than all the previous ones which clearly shows how SBERT embeddings are great at preserving contextual meaning and semantic similarity matching. Although the recall was not the best, the harmonic mean of both, F1 score was really close to the best ones. Many BERT based models and variations of BERT did not achieve such score meaning using multi head attention encoder mechanisms with bidirectional capabilities are better for generating representations than classification for this task.
Now for the model to perform better early, that is to generate predictions earlier we used time distributed dense layers. The best of ours was Model 2 (Doc2Vec-BiLSTM). Although both Model 2 and Model 3 did not perform here on par with the best achieved models in $ERDE_5$ and $ERDE_{50}$, they got pretty close scores in latency weighted F1.
To sum up, both Model 2 and 3 are competitive with the best achieving ones till date on most of the metrics. This proves our point in using representational features and sequential modeling in handling the specific case of this problem.Although the NLPGroup IISERB submission 0 in the 2022 version of the CLEF eRisk labs shows that traditional statistical machine learning models with hand crafted and customized BoW and TF-IDF can achieve results on par with the state-of-the-art.

Table 10.2: Comparison of Results

Model	**Precision**	**Recall**	**F1**	$ERDE_5$	$ERDE_{50}$	**Latency Weighted F1**
Model 1 Metamap SVM	0.56	0.57	0.56	0.18	0.16	0.36
Model 2 Doc2Vec BiLSTM	0.62	0.69	0.64	0.08	0.08	0.64
Model 3 SBERT BiGRU	0.74	0.76	0.75	0.06	0.06	0.75
FHDOB 2017 Doc2Vec LR	**0.69**	0.46	0.55	0.13	0.10	N/A
Sunday-Rocker 2, 4 2022 MixUp BERT	0.11	**1.00**	0.19	0.08	0.04	0.191
NLPGroup IISERB 0 2022 BOW+TF-IDF RF	0.68	0.74	**0.71**	0.05	0.03	**0.69**
LauSAn 4 2022 Tuend DistilBERT concat2	0.20	0.72	0.32	**0.04**	0.03	0.32
UNSL 2 2022 BERT base uncased	0.40	0.75	0.52	0.05	**0.03**	0.52

Chapter 11

Conclusion & Future Work

A large number of the world population is suffering from depression among which the underprivileged and third world countries are victimized in greater numbers due lack of proper resources for both diagnosis and treatment. Therefore, there is a need for automated depression detection that provides early diagnosis to victims to make room for early interventions. We proposed that representational features and semantic similarity based representations might be the best available ways to extract information from textual data for classification. Moreover, sequential modeling was experimented with the belief that this can retain more data from previous posts to make better informed decisions. From our results, we saw that these surely do provide results that are on par with the state-of-the-art. But our modeling was limited to a certain number of posts because of the limitation of memory retention capability of the models. In future, we hope more ways can be applied to incorporate the whole sequences in the classification process which we believe will give even better results. But because of the memory retention limitations of LSTM & GRU, we could not fit the whole sequence of posts to create predictions which could improve the performance even more. We believe in the future, more advanced architectures with better memory retention capabilities can incorporate whole sequences to produce even better results.Although pre-trained encoder transformer models have produced great results in many tasks of natural language processing, the size of the dataset in this task limits its capabilities to perform better than the existing models. We expect larger collections will help to fine tune such models for even better performance. Also, the information gained from the metamap SVM model can be used to boost other models. The data labeling was done through self-declaration which is obviously not the gold standard, we hope to collect gold standard labeled data in the future so that better experimentations can be performed.

Bibliography

[1] A. Ellis, "Rational-emotive therapy: Research data that supports the clinical and personality hypotheses of ret and other modes of cognitive-behavior therapy," *The Counseling Psychologist*, vol. 7, no. 1, pp. 2–42, 1977. DOI: 10.1177/001100007700700102. eprint: https://doi.org/10.1177/001100007700700102. [Online]. Available: https://doi.org/10.1177/001100007700700102.

[2] L. S. Radloff, "The ces-d scale: A self-report depression scale for research in the general population," *Applied Psychological Measurement*, vol. 1, no. 3, pp. 385–401, 1977. DOI: 10.1177/014662167700100306. eprint: https://doi.org/10.1177/014662167700100306. [Online]. Available: https://doi.org/10.1177/014662167700100306.

[3] N. H. Frijda, "Comment on oatley and johnson-laird's "towards a cognitive theory of the emotions"," *Cognition and Emotion*, vol. 1, pp. 51–59, 1987.

[4] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a "siamese" time delay neural network," in *Advances in neural information processing systems*, 1994, pp. 737–744.

[5] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273–297, 1995. DOI: 10.1007/BF00994018.

[6] M. Pasquini and M. Biondi, "Depression in cancer patients: A critical review," *Clinical Practice and Epidemiology in Mental Health*, vol. 3, p. 2, Feb. 2007. DOI: 10.1186/1745-0179-3-2.

[7] G. L. Clore and A. Ortony, "Appraisal theories: How cognition shapes affect into emotion," in *Handbook of emotions*, J. M. Haviland-Jones, M. Lewis, and L. F. Barrett, Eds., vol. 3, New York: Guilford Press, 2008, pp. 628–642.

[8] J. G. Rabkin, "Hiv and depression: 2008 review and update," *Current HIV/AIDS Reports*, vol. 5, no. 4, pp. 163–171, 2008. DOI: 10.1007/s11904-008-0025-1.

[9] R. Detels and C. Tan, "The scope and concerns of public health," Jan. 2009. DOI: 10.1093/med/9780199218707.003.0001.

[10] A. Wierzbicka, "Language and metalanguage: Key issues in emotion research," *Emotion Review*, vol. 1, no. 1, pp. 3–14, 2009.

[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. DOI: 10.1136/jamia.2009.002733.

[12] R. Řehůřek and P. Sojka, "Software Framework for Topic Modelling with Large Corpora," English, in *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, http://is.muni.cz/publication/884893/en, Valletta, Malta: ELRA, May 2010, pp. 45–50.

[13] W. Jarrold, H. S. Javitz, R. Krasnow, B. Peintner, E. Yeh, G. E. Swan, and M. Mehl, "Depression and self-focused language in structured interviews with older men," *Psychological Reports*, vol. 109, no. 2, pp. 686–700, 2011, PMID: 22238866. DOI: 10.2466/02.09.21.28.PR0.109.5.686-700. eprint: https://doi.org/10.2466/02.09.21.28.PR0.109.5.686-700. [Online]. Available: https://doi.org/10.2466/02.09.21.28.PR0.109.5.686-700.

[14] C. D. Wilson-Mendenhall, L. F. Barrett, W. K. Simmons, and L. W. Barsalou, "Grounding emotion in situated conceptualization," *Neuropsychologia*, vol. 49, no. 5, pp. 1105–1127, 2011.

[15] L. F. Barrett, "Emotions are real," *Emotion*, vol. 12, pp. 413–429, 2012.

[16] R. Katikalapudi, S. Chellappan, F. Montgomery, D. Wunsch, and K. Lutzen, "Associating internet usage with depressive behavior among college students," *Technology and Society Magazine, IEEE*, vol. 31, pp. 73–80, Dec. 2012. DOI: 10.1109/MTS.2012.2225462.

[17] A. Majid, "Current emotion research in the language sciences," *Emotion Review*, vol. 4, pp. 432–443, 2012.

[18] American Psychiatric Association, *Diagnostic and Statistical Manual of Mental Disorders*, 5th. American Psychiatric Publishing, 2013. DOI: 10.1176/appi.books.9780890425596. [Online]. Available: https://doi.org/10.1176/appi.books.9780890425596.

[19] G. L. Clore and A. Ortony, "Psychological construction in the occ model of emotion," *Emotion Review*, vol. 5, pp. 335–343, 2013.

[20] M. Gamon, M. Choudhury, S. Counts, and E. Horvitz, "Predicting depression via social media," Jul. 2013.

[21] J. W. Pennebaker, D. Páez, and B. Rimé, *Collective memory of political events: Social psychological perspectives*. Mahwah, NJ: Erlbaum, 2013.

[22] R. Rettner, "Depression higher in rich countries, study suggests," *Live Science*, May 2013. [Online]. Available: https://www.livescience.com/35792-global-depression-rates.html.

[23] S. E. Rivers, M. A. Brackett, M. R. Reyes, N. A. Elbertson, and P. Salovey, "Improving the social and emotional climate of classrooms: A clustered randomized controlled trial testing the ruler approach," *Prevention Science*, vol. 14, pp. 77–87, 2013.

[24] S. C. Widen, "Children's interpretation of facial expressions: The long path from valence-based to specific discrete categories," *Emotion Review*, vol. 5, pp. 72–77, 2013.

[25] C. D. Wilson-Mendenhall, L. F. Barrett, and L. W. Barsalou, "Situating emotional experience," *Frontiers in Human Neuroscience*, vol. 7, p. 764, 2013.

[26] K. Cho, B. van Merrienboer, Ç. Gülçehre, F. Bougares, H. Schwenk, and Y. Bengio, "Learning phrase representations using RNN encoder-decoder for statistical machine translation," *CoRR*, vol. abs/1406.1078, 2014. arXiv: 1406.1078. [Online]. Available: http://arxiv.org/abs/1406.1078.

[27] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in Twitter," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Baltimore, Maryland, USA: Association for Computational Linguistics, Jun. 2014, pp. 51–60. DOI: 10.3115/v1/W14-3207. [Online]. Available: https://aclanthology.org/W14-3207.

[28] M. Gendron, D. Roberson, J. M. van der Vyver, and L. F. Barrett, "Perceptions of emotion from facial expressions are not culturally universal: Evidence from a remote culture," *Emotion*, vol. 14, pp. 251–262, 2014.

[29] A. D. I. Kramer, J. E. Guillory, and J. T. Hancock, "Experimental evidence of massive-scale emotional contagion through social networks," *Proceedings of the National Academy of Sciences*, vol. 111, pp. 8788–8790, 2014.

[30] Q. V. Le and T. Mikolov, "Distributed representations of sentences and documents," *CoRR*, vol. abs/1405.4053, 2014. arXiv: 1405.4053. [Online]. Available: http://arxiv.org/abs/1405.4053.

[31] G. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "CLPsych 2015 shared task: Depression and PTSD on Twitter," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, Denver, Colorado: Association for Computational Linguistics, Jun. 2015, pp. 31–39. DOI: 10.3115/v1/W15-1204. [Online]. Available: https://aclanthology.org/W15-1204.

[32] K. A. Lindquist, A. B. Satpute, and M. Gendron, "Does language do more than communicate emotion?" *Current Directions in Psychological Science*, vol. 24, pp. 99–108, 2015.

[33] D. Preoţiuc-Pietro, M. Sap, H. A. Schwartz, and L. Ungar, "Mental illness detection at the world well-being project for the clpsych 2015 shared task," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015.

[34] J. D. Bernard, J. L. Baddeley, B. F. Rodriguez, and P. A. Burke, "Depression, language, and affect: An examination of the influence of baseline depression and affect induction on language," *Journal of Language and Social Psychology*, vol. 35, no. 3, pp. 317–326, 2016. DOI: 10.1177/0261927X15589186. eprint: https://doi.org/10.1177/0261927X15589186. [Online]. Available: https://doi.org/10.1177/0261927X15589186.

[35] G. Gkotsis, A. Oellrich, T. J. P. Hubbard, R. J. B. Dobson, M. Liakata, S. Velupillai, and R. Dutta, "The language of mental health problems in social media," in *Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2016.

[36] D. Losada and F. Crestani, "A test collection for research on depression and language use," vol. 9822, Sep. 2016, pp. 28–39, ISBN: 978-3-319-44563-2. DOI: 10.1007/978-3-319-44564-9_3.

[37] H. Almeida, A. Briand, and M.-J. Meurs, "Detecting early risk of depression from social media user-generated content," in *Working Notes of CLEF 2017–Conference and Labs of the Evaluation Forum*, 2017.

[38] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," in *Proceedings of the 4th Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality (CLPsych@ACL'17)*, 2017, pp. 32–40.

[39] A. Mousa and B. Schuller, "Contextual bidirectional long short-term memory recurrent neural network language models: A generative approach to sentiment analysis," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 1023–1032. [Online]. Available: https://aclanthology.org/E17-1096.

[40] F. Sadeque, D. Xu, and S. Bethard, "Uarizona at the clef erisk 2017 pilot task: Linear and recurrent models for early depression detection," *CEUR workshop proceedings*, vol. 1866, Sep. 2017.

[41] R. Trifu, B. Nemes, C. Bodea-Haţegan, and D. Cozman, "Linguistic indicators of language in major depressive disorder (mdd). an evidence based research," *Journal of Evidence-Based Psychotherapies*, vol. 17, pp. 105–128, Mar. 2017. DOI: 10.24193/jebp.2017.1.7.

[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," Jun. 2017.

[43] E. Villatoro-Tello, G. Ramírez-de-la-Rosa, and H. Jiménez-Salazar, "Uam's participation at clef erisk 2017 task: Towards modelling depressed bloggers," in *Working Notes of CLEF 2017—Conference and Labs of the Evaluation Forum*, 2017.

[44] WHO TEAM, *Depression and Other Common Mental Disorders: Global Health Estimates*. World Health Organization, Jan. 2017. [Online]. Available: https://www.who.int/publications/i/item/depression-global-health-estimates.

[45] F. Cacheda, D. Fernández Iglesias, F. J. Nóvoa, and V. Carneiro, "Analysis and experiments on early detection of depression," in *Working Notes of CLEF 2018–Conference and Labs of the Evaluation Forum*, 2018.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. arXiv: 1810.04805. [Online]. Available: https://dblp.org/rec/journals/corr/abs-1810-04805.

[47] D. G. Funez, M. J. Garciarena Ucelay, M. P. Villegas, S. Burdisso, L. C. Cagnina, M. Montes-y-Gómez, and M. Errecalde, "UNSL's participation at erisk 2018 lab," in *Working Notes of CLEF 2018–Conference and Labs of the Evaluation Forum*, 2018.

[48] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," Jan. 2018, pp. 88–97. DOI: 10.18653/v1/W18-0609.

[49] N. Liu, Z. Zhou, K. Xin, and F. Ren, "Tua1 at erisk 2018," in *Working Notes of CLEF 2018—Conference and Labs of the Evaluation Forum*, 2018.

[50] F. Sadeque, D. Xu, and S. Bethard, "Measuring the latency of depression detection in social media," in *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, ser. WSDM '18, Marina Del Rey, CA, USA: Association for Computing Machinery, 2018, pp. 495–503, ISBN: 9781450355810. DOI: 10.1145/3159652.3159725. [Online]. Available: https://doi.org/10.1145/3159652.3159725.

[51] M. Stankevich, V. Isakov, D. Deviatkin, and I. Smirnov, "Feature engineering for depression detection in social media," Jan. 2018, pp. 426–431. DOI: 10.5220/0006598604260431.

[52] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *CoRR*, vol. abs/1804.07000, 2018. arXiv: 1804.07000. [Online]. Available: http://arxiv.org/abs/1804.07000.

[53] S. G. Burdisso, M. Errecalde, and M. Montes-y-Gómez, "T-ss3: A text classifier with dynamic n-grams for early risk detection over text streams," *CoRR*, vol. abs/1911.06147, 2019. arXiv: 1911.06147. [Online]. Available: http://arxiv.org/abs/1911.06147.

[54] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *CoRR*, vol. abs/1908.10084, 2019. arXiv: 1908.10084. [Online]. Available: http://arxiv.org/abs/1908.10084.

[55] ——, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[56] ——, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Nov. 2019. [Online]. Available: https://arxiv.org/abs/1908.10084.

[57] "Global health data exchange," May 2021. [Online]. Available: http://ghdx.healthdata.org/gbd-results-tool?params=gbd-api-2019-permalink/%20%5C%5C%20d780dffbe8a381b25e1416884959e88b.

[58] E. A. Ríssola, D. E. Losada, and F. Crestani, "A survey of computational methods for online mental state assessment on social media," *ACM Trans. Comput. Healthcare*, vol. 2, no. 2, Mar. 2021, ISSN: 2691-1957. DOI: 10.1145/3437259. [Online]. Available: https://doi.org/10.1145/3437259.

[59] R.-A. Gînga, A.-A. Manea, and B.-M. Dobre, "Sunday rockers at erisk 2022: Early detection of depression," in *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, Bologna, Italy, Sep. 2022.

[60] N. Liu, Z. Zhou, K. Xin, and F. Ren, "Tua1 at erisk 2022: Exploring affective memories for early detection of depression," in *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum*, Bologna, Italy, Sep. 2022, pp. 1026–1037.

[61] J. Parapar, P. Martín-Rodilla, D. E. Losada, and F. Crestani, "Overview of erisk 2022: Early risk prediction on the internet," in *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 13th International Conference of the CLEF Association, CLEF 2022, Bologna, Italy, September 5–8, 2022, Proceedings*, Bologna, Italy: Springer-Verlag, 2022, pp. 233–256, ISBN: 978-3-031-13642-9. DOI: 10.1007/978-3-031-13643-6_18. [Online]. Available: https://doi.org/10.1007/978-3-031-13643-6_18.

[62] R. Poświata and M. Perełkiewicz, "OPI@LT-EDI-ACL2022: Detecting signs of depression from social media text using RoBERTa pre-trained language models," in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, Dublin, Ireland: Association for Computational Linguistics, May 2022, pp. 276–282. DOI: 10.18653/v1/2022.ltedi-1.40. [Online]. Available: https://aclanthology.org/2022.ltedi-1.40.

[63] K. S and D. Thenmozhi, "Data set creation and empirical analysis for detecting signs of depression from social media postings," Feb. 2022.

[64] I. Tavchioski, B. Škrlj, S. Pollak, and B. Koloski, "Early detection of depression with linear models using hand-crafted and contextual features," in *Proceedings of the Working Notes of CLEF 2022–Conference and Labs of the Evaluation Forum*, Bologna, Italy, Sep. 2022, pp. 1005–1013.