DEPRESSION DETECTION IN SOCIAL MEDIA VIA DIFFERENTIAL TEXT EMBEDDING

By

Norah Alfadhli

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

Computer Science - Master of Science

2022

ABSTRACT

Deep learning models have shown promising results for depression detection using social media data (i.e., Twitter), but the difficulties of maintaining explainability and few-shot adaptation of models for new problems remain an open challenge. The standard solution for depression detection modeling is to represent the natural language text of the tweet as a numerical vector via embedding first then training a classification model that uses the vectors to predict the depression status. In this study, we propose a few-shot learning technique to improve the performance of depression detection classification models. More specifically, we represent tweets as *differential embeddings*: a set of embedding vectors that measure the tweet's (Sentence BERT) embedding location with respect to a set of depression tweet templates (anchor points) derived from clinically backed depression symptoms described in the literature. Intuitively, the *differential embeddings* describe the similarities between different tweets and the set of depression templates. We have assessed the capability of our approach on random samples we drew from a source of tweets to create multiple datasets as follows: (1)20 random balanced datasets and (2)20 random unbalanced dataset. We trained a supervised model using different approaches derived from Sentence-BERT and the anchor points. The results show that the proposed solution improved SBERT in a supervised task by 0.035 and .023 relative improvements in terms of Partial AUROc @FPR: 0.10 in balanced and unbalanced datasets, respectively.

ACKNOWLEDGMENTS

First and foremost, I must acknowledge my limitless thanks to Allah, the Ever-magnificent, the Ever-Thankful, for His help and bless by giving me the opportunity, courage and enough energy to carry out and complete the entire thesis

I would like to express my deepest appreciation to my advisor Dr. Mohammad Ghassemi for his support, encouragement, and guidance during this journey. He did not only teach me how to be a better researcher, but also how to talk about science. Without his assistance and the many thought-provoking discussions, I would never be able to complete this thesis. Dr. Ghassemi was always available and willing to give even if it will cost his valuable time. I am thankful to Dr. Ross and Dr. Johnson for accepting to be in my master thesis committee. I am also grateful to all my colleagues for their continuous support and help, including Sari Sadiya, Niloufar Eghbali and Najla'a Alsaedi.

I must express my very profound gratitude to my parents for their unfailing support, continuous encouragement and prayers throughout my years of study. I would like to express my special thanks to my beloved mother who, through her thorough discussions and patience throughout my life, made me the person who I am today. I am deeply thankful to my husband for his confidence, patience and positive energy when I have been stressed out. I appreciate my brothers, sisters, and friends for their great motivation.

TABLE OF CONTENTS

Chapter 1: Introduction and Literature Review

1.1 Introduction

1.1.1 Background to the study

Globally, more than 264 million people of all ages suffer from depression [1]. The scale of the problem is so immense that it is both logistically and financially impossible to monitor the early signs of depression using human agents alone. Additionally, studies have shown that those suffering from depression (especially young people) find it more challenging to discuss their feelings in face-to-face settings [2] but are less hesitant to share their thoughts on social media platforms. This provides an incentive for developing automated systems that can detect the presence of depression symptoms in social media for a possible downstream intervention.

1.1.2 Research Objective

The present research aims to apply techniques in natural language processing (NLP) and ML to build a system that can identify social media posts (i.e., tweets) that contain the elements of the symptoms of depression. More specifically, the current research aims to design a few-shot approach that synthesizes NLP algorithms (i.e., Sentence-Bert) with clinical intuitions about depression (i.e., clinical depression symptoms). We demonstrate that this process can help detect depression more effectively than algorithms or insights alone.

Previous works have shown that data mining of social media activity can aid in detecting cases of depression automatically [3], [4], [5], [6]. Especially in the case of using text data from social media, ML approaches have unique advantages in the detection of depression; With people's participation in online platforms and their public sharing via the internet, text data from social media records are a treasure trove of psychological data that may assist in

screening for depressive symptoms among the users of social media. In fact, the number of daily social media users is increasing and a significant percentage of the users have reported being depressed. ML techniques also offer opportunities for identifying hidden patterns in online communication and interaction on social media that may reveal users' mental states. The automatic detection of depressive symptoms through ML algorithms applied to social media data has the potential to identify those at risk of depression through large-scale monitoring of online social networks, potentially complementing traditional screening procedures.

1.1.3 Research Contributions

As described in the literature review section below (see section 1.4), the existing depression detection systems have two primary limitations:

1. *Limited clinical relevance:* Many state-of-the-art (SOTA) techniques (i.e., deep learning) achieve excellent performance but do not explicity account for factors that are known to be clinically relevant for depression diagnosis [5, 7].
2. *High false positive rates*: Many existing approaches are ineffective at discriminating between depressed tweets from those tweets that carry a negative sentiment but are not necessarily depressed. Indeed, many depression detection systems deliberately exclude negative sentiments from the nondepressed class during training and evaluation [8].

The present research addresses these challenges through a simple but effective few-shot solution that utilizes a small amount of labeled data to enhance model generalization to unseen data. More specifically, our work has two advantages over existing alternatives:

1. Augments the power of contemporary representation learning techniques with a clinically grounded indication of depression symptoms in texts.
2. Is trained on a dataset that includes a portion of *negative tweets* in the non-depressed class, making the learning task more challenging but also more relevant for real-world

deployment.

1.2 Outline

The remainder of the current thesis is organized as follows: Chapter 1 focuses on a review of the key concepts of NLP and ML that are relevant to the thesis; we also present a literature review, in which the research related to depression detection in social media is discussed. Chapter 2 describes the proposed methodological approach and baseline of the research. It also includes a discussion of the data, pre-processing approach, and feature extraction pipeline. Chapter 3 provides a detailed description of the experiments performed to assess the research method and their results; It also provides a discussion of the research results, the context of the literature, and the propensity for downstream clinical use of the approach in this research. Finally, Chapter 4 is a discussion of the key takeaways, limitations, and future directions of work.

1.3 Theoretical Background

The present thesis utilizes NLP and ML to build an automated depression detection system. In this chapter, we review the key concepts and theories that our work builds on. NLP is the analysis of human (i.e., "natural") language using computational and statistical techniques. The objective of NLP is to allow computers to represent, understand, analyze, and derive meaning from human language [9]. In recent years, there have been significant advances in NLP techniques. Recent advances have utilized deep neural architectures with multi-headed attention (i.e., BERT approaches) for language modeling. These models are trained using the wealth of natural language human text available online (and offline). The SOTA language models provide contextual representations of natural language that are useful for a variety of downstream tasks, including text summarization, question answering, and topic classification with little-to-no costly human annotation required.

Generally speaking, NLP systems are composed of two types: rule-based approaches and statistical NLP. The rule-based approaches are based on a set of rules that guide the system. One example of this is the sentence parsing systems that use a nominally complete set of rules that define allowable words, parts of speech, and allowable sequences of the parts of speech. Despite its strengths, its shortcomings includes the difficulty in modeling natural language using a set of rules based on a predefined vocabulary. On the other hand, statistical NLP consists of all the quantitative approaches (often probabilistic) to automated language processing and modeling language implicitly instead of using explicit rules. One criticism of the statistical approach is that the statistical assumptions may not match the intuition of current research on how languages work. To this effect, text corpora–based approaches could be subject to criticism because they have insufficient data. In the current study,

instead of hand-coding rules, we used statistical NLP to learn these rules by analyzing a set of examples and making statistical inferences. This approach relies on methods based on ML algorithms [10].

Arthur Samuel described ML as the computer's ability to learn without being explicitly programmed. Hence, ML can make decisions about new data without being instructed on how to do so. Machines can learn different tasks and are able to do so in one of two forms: supervised learning and unsupervised learning. In supervised learning, an algorithm is given training data and the desired solution (labels). The two main types of supervised learning problems include classification, which involves predicting a class label, and regression, which involves a numerical value. The supervised learning uses historical data (data from the past) to learn patterns and uncover relationships between features and the target [11]. In unsupervised learning, an algorithm models the underlying data structure or distribution to learn the pattern within the structure. It is used to solve ML problems when there are no ground truths (known as the target for training or validating the model with a labeled dataset). Clustering is an example of unsupervised learning [11].

ML applications typically involve the following steps: data collection and preprocessing, features engineering, training a predictive model, and testing the performance.

1.3.1 Feature Engineering

This is referred to as the process of using knowledge of the domain to select, manipulate, and transform raw data into features that can be used in supervised and unsupervised learning. A feature is a form of information that is useful for prediction. In computer vision, an image is an observation, but a feature could be a line in the image. In natural language processing, a document or tweet could be an observation, and a phrase or word count could be a feature. In speech recognition, an utterance could be an observation, but a feature might

be a single word or phoneme [12]. Feature engineering in deep learning is direct and can be performed by an algorithm. More specifically, deep learning attempts to mimic the activity in the layers of neurons in the human neocortex to enable it to transform the input raw data into features, a process that occurs in an early stage of the training process.On the contrary, in conventional ML (shallow learning), features engineering is carried out outside the algorithmic stage. Experts and nonmachines are in charge of analyzing raw data to transform it into valuable features [13]. A dataset that is given to an ML algorithm contains dependent and independent variables. The outcome of a prediction is the dependent variable. The expert can provide features as part of the data set or may be derived from the data in the case of textual data. As a machine learns, it finds patterns in data (associated or not with given classes). In all ML applications, the data are first converted into a representation (a set of features) that can be interpreted. To apply ML algorithms to NLP applications, the text has to be transformed into a numeric (or discrete) representation.

Features Representation in NLP: Finding useful features is an integral part of conventional Machine Learning research. In the case of NLP tasks, these features are extracted from text. Some kinds of features rely on word frequencies such as Bags of Words and n-grams. Other features are more problem-specific, such as the sentiment value of a document, its readability level, tone, etc. Generating these features involves extracting information from the text and converting it into a form that machine learning algorithms can understand. As an example, NLP uses deep learning to represent information from the text in the form of embedding representations. Deep Learning methods out-compete other statistical and linguistic models for NLP tasks [14]. Deep learning can learn the features from the natural language required by the model, rather than requiring that the features be specified and extracted by an expert. This learned representation is called embedding. The way

words and documents are represented is a key breakthrough in deep learning when it is applied to challenging NLP problems. A word or document embedding is similar for words or documents that have the same meaning. A SOTA approach to learning document representation embedding is by utilizing Bidirectional Encoder Representations from Transformers (BERT) networks. BERT is SOTA language model for NLP. The key technical innovation of BERT is its application of Transformer, a popular attention model, in bidirectional training to language modeling. Prior research looked at text sequences left to right or combined left-to-right and right-to-left training, but BERT examines a text sequence bidirectionally. BERT shows that bi-directionally trained language models provide a deeper understanding of language context and flow than their single-direction counterparts. To learn the context of words (or sub-words) within a text, BERT uses a Transformer. A transformer consists of two mechanisms such as an encoder that reads text input and a decoder that produces a prediction. Given that BERT aims to produce a language model, only the encoder mechanism is necessary. The Transformer encoder reads the entire sequence of words at once, in contrast to directional models, which read the text input sequentially (from right to left or left to right). As a result, it is regarded as bidirectional. This characteristic enables the model to learn a word's context depending on all of its surroundings (left and right of the word). A series of tokens are used as the input and are first embedded into vectors before being processed by a neural network. The result is a series of vectors of size H, each of which corresponds to a token from the input with the same index [15]. A modification of BERT was introduced to better handle sentence embedding. Sentence-BERT (SBERT) uses siamese and triplet network structures to derive semantically meaningful embedded sentences that can be compared using cosine similarity. This modification of BERT enables BERT to be used for certain tasks such as large-scale semantic similarity comparison, clustering, and

information retrieval through semantic search. However, BERT uses a cross-encoder which requires two sentences to be fed into the network and the target value is predicted. Due to a large number of possible combinations, this setup is not suitable for various pair regression tasks. An inference computation via cross-encoding would need to be completed 100K times if we wanted to search for similarity in a small 100K sentence dataset. To cluster sentences, we would have to compare all 100K sentences, resulting in just under 500M comparisons. Therefore, to address the issue of the expensive computation, we need to pre-compute sentence vectors that can be stored and then used whenever required. SBERT was developed to handle the limitation of BERT by processing one sentence at a time. By using siamese network architecture in SBERT, it ensures that fixed-sized vectors for input sentences can be derived. Using a similarity measure like cosine similarity or Manhatten / Euclidean distance, semantically similar sentences could be found. Clustering and semantic search are commonly accomplished by mapping each sentence into a vector space where semantically similar sentences are grouped together. BERT has been used to identify fixed-size embedding of individual sentences and the common approaches include averaging the BERT output layer (known as BERT embedding) or by using the first to- ken ([CLS]). This common practice shows bad sentence embedding [16]. This research uses textual data from Twitter. Though a fine-tuned version of BERT was proposed to handle twitter data (BERTweet), the model uses the same architecture as BERT which results in the same shortcoming of BERT that were mentioned above

1.3.2 Feature Selection

Features selection is one of the core principles in ML that extremely affects the overall performance of the predictive model. The process of automatically or manually selecting the features that make the most contribution to the desired prediction variable or output is

known as feature selection. Having irrelevant or partially relevant features in the data can decrease the accuracy of the models and make the model learn based on irrelevant features. The model can benefit from features selection in three ways: 1) improve accuracy, 2)reduce overfitting by reducing number of features, and 3) reduce the training time.

One of the features selection methods is sequential features selection (SFS). Farwrd-SFS is a greedy process that iteratively finds the best new feature to add to the set of selected features. Concretely, at first, we start with a zero feature and find the one feature that increases a cross-validated score when an estimator is trained on this single feature. We repeat the process by adding a new feature to the set of selected features after selecting the first one. When the desired number of selected features has been reached or there is no improvement, the procedure ends [17].

1.3.3 Classifier Selection

Some of the most widely used algorithms in NLP include logistic regression (LR), support vector machine (SVM), naive Bayes (NB), K-nearest neighbors (KNN), and ensemble methods. Below we provide a brief overview of the method we used in this research.

Logistic Regression: LR is a statistical model that is often used for classification and predictive analysis. LR estimates the probability of an event occurring given a set of indicators (i.e., features). Because the outcome is a probability, it lies between 0 and 1. This model uses a logistic function (Sigmoid function) to map the predicted values to probabilities [12]. Logistic regression is commonly used to predict binary target variables, but it can be expanded and further divided into three different types: binomial, where a target variable can only have two types, for example, predicting whether an email is spam; polynomial, which is when the target variable has more than two types that may not have quantitative meaning, for example, predicting illness; and ordinal, where the categories of the target variable are

ordered, for example, a web series rating from 1 to 5.

Cost function: it is a mathematical formula used to quantify the error between the predicted values and expected values. More specifically, a cost function is a measure of how wrong the model is in terms of its ability to estimate the relationship between x and y. The value returned by the cost function is referred to as the cost or loss or, simply, the error [18].

1.3.4 Performance Measures

To evaluate the ML models, performance evaluation metrics are used. Based on their outputs, the performance measurements are different for supervised and unsupervised algorithms. In supervised algorithms, performance measurements rely on the correctly classified labels. Examples of evaluation metrics that can be used with supervised algorithms are accuracy, F1 score, precision, recall, and the receiver operator characteristics (ROC) curve.

- Precision:

  The precision is calculated as the ratio between the number of positive samples correctly classified to the total number of samples classified as positive (either correctly or incorrectly). The precision measures the model's accuracy in classifying a sample as positive.

$$P = \frac{TP}{TP + FP} \tag{1.1}$$

  where,

  TP : True positive,

  FP : False positive

  A high precision indicates that 1) the model makes many correct positive classifications

(maximize true positive). 2) the model makes fewer incorrect positive classifications (minimize false positive) [19].

- Recall:

  The recall is determined as the proportion of positive samples that were correctly identified as positive to all positive samples. The recall measures how well the model can identify positive samples. The more positive samples that are identified, the higher the recall will be.

$$P = \frac{TP}{TP + FN} \tag{1.2}$$

  where,

  TP : True positive,

  FN : False negative

  The recall cares only about how the positive samples are classified. This occurs independently of how the negative samples are classified, for example, for the precision. When the model classifies all the positive samples as positive, then the recall will be 100%, even if all the negative samples were incorrectly classified as positive [19].

- Precision-Recall Curve:

  The precision-recall curve shows the trade-off between precision and recall for different threshold. High precision indicates a low false positive rate, while high recall indicates a low false negative rate. A high area under the curve reflects both high recall and high precision. High scores for both indicate that the classifier is yielding results that are accurate (high precision) and that are mostly positive (high recall). The precision-recall curve is recommended when the classes are imbalanced. [19].

- ROC Curve:

  A ROC (receiver operating characteristic) curve is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: the true positive rate (TPR) and false positive rate (FPR). The true positive rate (TPR) is a synonym for recall. False positive rate (FPR) is defined as follows:

$$P = \frac{TP}{TP + TN} \tag{1.3}$$

  where,

  TP : True positive,

  FN : True negative

- AUC: Area under the ROC Curve:

  AUC stands for "Area under the ROC curve." That is, the AUC measures the entire two-dimensional area underneath the entire ROC curve. The AUC provides an aggregate measure of performance across all possible classification thresholds. The values range from 0 to 1. A model that predicts 100% incorrectly has an AUC of 0.0, while a model whose prediction is 100% correct has an AUC of 1.0.

  The AUC is recommended when we are looking for a metric that 1) measures the ranking of predictions, not absolute values and that 2) measures the quality of the model's predictions, regardless of which classification threshold is chosen [20].

Another factor that affects the choice of the evaluation metric is the nature of the dataset. It would be misleading to evaluate an imbalanced dataset using an accuracy score; in a test set that contains majority and minority examples, a model that predicts the majority class for all examples will have a classification accuracy as high as 99%, reflecting the distribution

of the major and minor examples expected on average in the test set [21]. For this reason, we use the ROC and precision-recall curves in the current research.

1.4 Literature Review

1.4.1 Depression Detection in Social Media

Worldwide, depression affects more than 264 million individuals of all ages [1]. Because of the problem's enormous scope, it is both logistically and monetarily unfeasible to detect early indications of depression solely with human agents. Additionally, research has indicated that people with depression, especially young people, find it more difficult to talk about their feelings in person than they do online [2]; however, they are less afraid to share their challenges on social media. This has also sparked the creation of automated algorithms that look for signs of depression in social media to potentially offer an intervention. To build systems that can detect depression, a features-based approach could be used. This approach requires some knowledge of the problems' domain to extract meaningful features from texts. Another approach that is often used involves the use of SOTA deep-learning techniques. This research has shown that deep learning has promising results in a wide range of problems, but it has its shortcomings when it come to clinical context as it compromises the clinical relevance. Few-shot learning approach is the most recent approach for detecting depression. The purpose of few-shot learning models is to improve model generalizabilily for cases where training data is scarce.

1.4.2 Approaches for Depression Detection in Social Media

1.4.2.1 Features-Based Approaches

Classical depression-detection systems rely heavily on expertly crafted features, including linguistic [22], psycholinguistic [23], textual [24], [3], [25], semantic [25], and sentiment

features [26], [27]. These features are often used in conjunction with shallow modeling frameworks (e.g., SVM) for the task of depression detection. However, not all features are equally valuable. For example, Cacheda et al. reported that textual features tend to outperform their semantic counterparts when used for depression detection [25], and Alsagri and Yakhlef found that combining synonyms with LIWC (linguistic inquiry and word count), sentiment analysis, and social activity increases the accuracy of detection models [3]. De Choudhry et al. performed crowd-sourcing to identify Twitter users who were reported as being depressed based on psychometric measures; to identify the symptoms of major depressive disorder, the authors used behavioral characteristics identified under engagement, egocentric social graph, depressive language, emotion, and linguistic style; they reported an accuracy of 70% and a higher precision of 74% for the depression class [28]. Taking a similar route, Tsugawa et al. considered a user's activity history to collect ground truth data for predicting depression among Twitter users [29]. In this work, the authors used bag of words and word frequencies, in addition to the features used in [28], to identify the ratio of tweet topics. Although subtle differences in behavioral features were found between Tsugawa et al.'s research and that of De Choudry et al., the analysis performed by Tsugawa et al. found similar patterns for the use of negative words, frequency of posting, retweet rate, and URLs contained in tweets. They found that support vector machine (SVM) classifiers using features generated from Twitter user activity resulted in an accuracy of 69%, with a precision of 0.64 and recall of 0.43 [29].

1.4.2.2 Deep Learning

Contemporary systems increasingly leverage representation-learning approaches for depression detection. The recent work by Hussein et al. explored the performance of several word-level embedding techniques (random trainable, skip-gram, and CBOW) with convolu-

tional neural network (CNN) and recurrent neural network (RNN) models. The researchers demonstrated the superiority of CNNs and RNNs over simple feature-based approaches (SVM classifier using TF-IDF) [7]. More recent work by Zogan et al. used a concatenated CNN and bidirectional gated recurrent unit (BiGRU) to identify social media users at risk of depression using temporal, semantic, and behavioral data [30]. Indeed, the general direction of research in this domain focuses on the use of deep-learning techniques. Although this has (in many cases) improved model performance compared with feature-based approaches, it has also (in some cases) compromised the clinical relevance and interpretability of the models. However, clinical relevance is important in the automated depression-detection task to inform an appropriate potential intervention [31]. SOTA deep-learning approaches do not, for instance, provide a way to understand the underlying distorted thinking patterns that may be responsible for depression classification. Being able to associate these patterns with the depression classification task can help in managing depression at early stages, with treatment being as simple as cognitive reconstruction therapy [32] [31].

1.4.2.3 Few-Shots Learning

An attempt to detect and confirm symptoms of depression has been created [5]. The researchers proposed a zero-shot learning model that predicts a possible relationship of a sample to an unseen label, that is, to a label that the model did not see during the training. They have proposed a set of depression symptoms and descriptors representing the symptoms (words). For depression detection in a post, they assigned a membership score for each of the symptoms that appears in the post. As a result, they used a set of membership scores as the representation of a post sent to an SVM classifier. The authors reported a significant capability of few-shot learning models compared with the baselines. Additionally, one study constrained the behavior of depression detection methods by the presence of symptoms

known to be related to depression (i.e., clinically backed symptoms) while producing a model that is easy to inspect [8]. These researchers have proposed a questionnaire model and depression model; the questionnaire model used a pattern-based method that matched every post against symptom patterns; this model comprises nine symptom classifiers, such as anhedonia, concentration, eating, fatigue, mood, psychomotor, self-esteem, self-harm, and sleep. The model takes BERT embeddings and weakly labeled symptoms data as the input and generates the final question scores (i.e., symptom scores) or the hidden layers (i.e., symptom vectors) of the nine submodels. The depression model then uses the outputs of the questionnaire model to predict depression in posts. The researchers in [8] have shown that their approach performs well compared with strong baselines (unconstrained BERT) while generalizing better results. Constraining the behavior of depression detection models by the presence of depression symptoms has different advantages. This type of model has the advantage of being inherently more reliable than a black-box model because it determines classifications based on the presence of specific symptoms in specific posts so that it can be inspected to assess the quality of the evidence for a diagnosis. Apart from this argument, the model can generalize more effectively by limiting the use of spurious shortcuts. [8].

1.5 Summary

This chapter has introduced the key concepts of NLP and ML that are relevant to the depression detection system designed in the current research. This chapter focused on the commonly used techniques, features, classifiers, and performance measures used for an NLP task. It also presented the existing social media systems that have been developed to identify depression symptoms while providing a literature review that has uncovered the need for a constrained, automated mental illness detection system.

Chapter 2: Methodology

In this chapter, we first describe the approach we used to detect depression in social media and provided an overview of our final proposed system. Figure 2.1 illustrates the methodological pipeline.

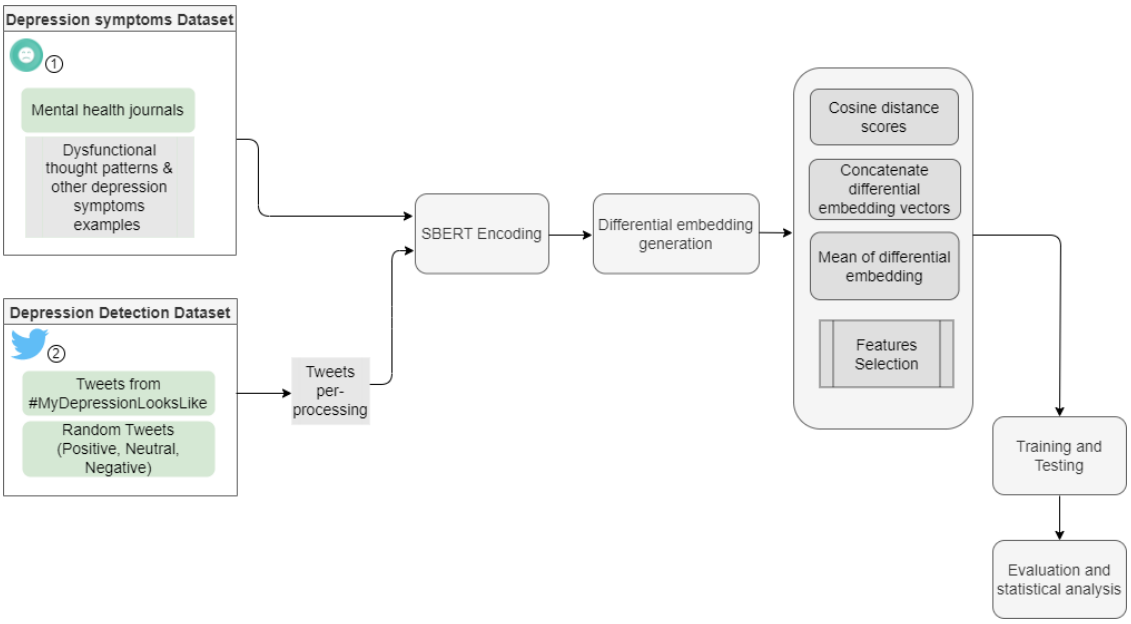


Figure 2.1: The methodological pipeline. More details about (1) can be found in 3.1.1, a detailed description of (2) is available in 2.1.1. The differential embedding generation explained in 2.1.4.1

The problem of detecting depression symptoms from posts on social media has been formulated as a binary classification problem. The target classes are depressed and non-depressed, where the class indicates the presence of depression symptoms in a post. That is, our depression detection system is required to simply predict whether a text belongs to the depressed class or not.

The research questions we are trying to answer in this study are: (1) can we leverage the strength of SOTA deep learning techniques without any training or fine-tuning, while maintaining clinical relevance in a semi-supervised depression detection system? (2) can

dysfunctional thought patterns be effectively used in a depression detection system?

2.1 Data Collection and Processing

For this study, we used two datasets which we described in greater detail below.

2.1.1 Depression Symptoms Dataset

In this research, we considered two sets of depression symptoms. More specifically, we combined a set of (1) cognitive depression symptoms that have been used in several studies and clinically confirmed to be depression symptoms and (2) a set of cognitive distortions (i.e dysfunctional thought patters) that we propose and investigate as depression symptoms.

2.1.1.1 Dysfunctional Thought Pattern Definitions:

According to cognitive therapy, even a seemingly insignificant event such as forgetting an appointment can cause individuals to feel anxious or depressed if unwarranted negative interpretations are made, such as, "That's just like me; I forget everything," or "I blew it; they'll never want to talk to me again." A negative interpretation usually makes an event appear to be worse than it truly is [33]. Feeling hopeless, guilty, angry, or discouraged can be triggered by thoughts; a core principle of cognitive therapy is that negatively based thoughts contribute to mood (and other) disorders. These thoughts are referred to as "dysfunctional thoughts," or "cognitive distortions." Dysfunctional thoughts are negative perceptions of oneself, others, and the world [34]. Although everyone generates an occasional inaccurate interpretation; depressed individuals have an overall, systematic bias towards dysfunctional thoughts [35]. A significant relationship has been found between the frequency of dysfunctional thoughts and the severity of clinical symptoms [36]. In an attempt to look closely at how users talk about their depression on social media, Lachmar et al. performed a study on the popular hashtag #Mydepressionlookslike [2]. They found that one of the most common themes is

using language that shows dysfunctional thought patterns such as fortune-telling, emotional reasoning, labeling, mind-reading, overgeneralizing, personalizing, and "should" thinking. A feasibility test was done in an attempt to classify dysfunctional thought automatically by Cromer et al. [37]; their system reliably detects the seven types of dysfunctional thought categories that were mentioned earlier. This automatic identification is based on language markers that make it feasible for the system to distinguish between different types of dysfunctional thoughts. However, no prior work has been done to show the impact of dysfunctional thoughts analysis on the depression-detection task in social media.

2.1.1.2 Dysfunctional Thought Patterns Dataset

We collected a set of sentences that represent clinically grounded dysfunctional thought categories. The dataset consists of 63 exemplary statements that reflect the presence of the dysfunctional thought categories. These statements were collected from a review of several mental health journal articles (see Table 1 in Appendix for complete details). These sentences serve as anchor points for features matrix generation. Table 2.1 shows examples of the representative sentences of the seven dysfunctional thought categories (A full set is available in Appendix Table2).

Table 2.1: Dysfunctional thought categories and representative sentences

Category	**Example**
Mind reading	He thinks I am a loser.
Labeling	I must be a worthless person.
Fortune telling	I will get rejected.
Overgeneralizing	I am going to fail at everything.
Emotional reasoning	My boyfriend is upset; therefore, I must have done something wrong.
Personalizing	The world has got it in for me.
Should & Must	I should always give everything I do 100%

This dataset was used to construct the features matrix in subsequent steps.

2.1.1.3 Introducing Other Depression Symptoms

Depression symptoms could be self-reported or observable symptoms [38]. Dysfunctional thought patterns are an example of signs that are detected by either an observer or by self-observation. On the other hand, [38] provided a set of self-reported and clinically-observed signs. In this work, we used a subset of the self-reported symptoms, such as loss of interest, pleasure loss, inability to feel, etc. We included eight symptoms that we suspected could be found in writing rather than diagnosed clinically.

Using the same approach we followed to represent dysfunctional thought categories, we collected a set of exemplary statements as anchor points that reflect eight self-reported symptoms. Unlike dysfunctional thoughts that appear in one's writing, this set of symptoms were self-reported. Finding a set of representative statements was not feasible due to the nature of these symptoms. To cope with this, we used keywords that could be used to describe each of the symptoms and used it in a short sentence. For these symptoms, we used four to five sentences for each symptom as, unlike dysfunctional thought categories, we were not looking for varying language markers. Table 2.2 shows the self-reported symptoms included in this work, and representative statements.

2.1.2 Depression Detection Dataset

In all experiments, a depression detection dataset included instances from two sources of tweets, representing the two classes:

1. Depressive tweets: contained a #MyDepressionLooksLike hashtag; we made the reasonable assumption that tweets with this hashtag were (1) generated by individuals that self-identify as depressed and (2) would contain self-reported symptoms of the

Table 2.2: Self-reported symptoms and representative sentences

Category	**Example**
Loss of insight	Lack of understanding
Pleasure loss	I feel miserable
Interest loss	I am finished with it
Feeling bothered	I am not happy with this
Energy loss	I feel mentally drained
Inability to feel	I feel unmoved
Feeling needed	Be valued at something
Feeling happy	I am over the moon

individual's depression. To address why we particularly used the tweets from this hashtag, from a review of multiple publicly available depression datasets, we found that (in most cases) these datasets contain (1) information about the depressed users (with no tweets available) [39], [**?**], (2) sentiment tweets; where the researchers who bases their work on this kind of datasets consider the negative class as the depressed class [41]. However, this research argues that negative tweets should be a part of the non-depressed class as not all "sad" or "negative" users are depressed.

2. Non-Depressive tweets: Tweets were selected from [42]: a dataset of 1,600,000 tweets and contains tweets annotated with neutral, positive, and **negative scores.**.

The non-depressive tweets has over a million tweets. If we use all of them against the depressive tweets (≈ 2000), severe class imbalance may cause the predictive performance of the machine learning algorithm to be biased toward the majority class during model training. For this reason, we randomly drew a number of samples from the non-depressive tweets to test against the depressive tweets. We evaluated two scenarios: (1) datasets with balanced classes. (2) datasets with imbalanced classes. To get the random samples, we made sure they contain equal portions of positive, neutral and negative tweets. To assess the statis-

tical integrity of our results, we evaluated our model using *multiple* random samples from the source data and create multiple datasets. Figure 2.3 illustrates the dataset generation procedure.

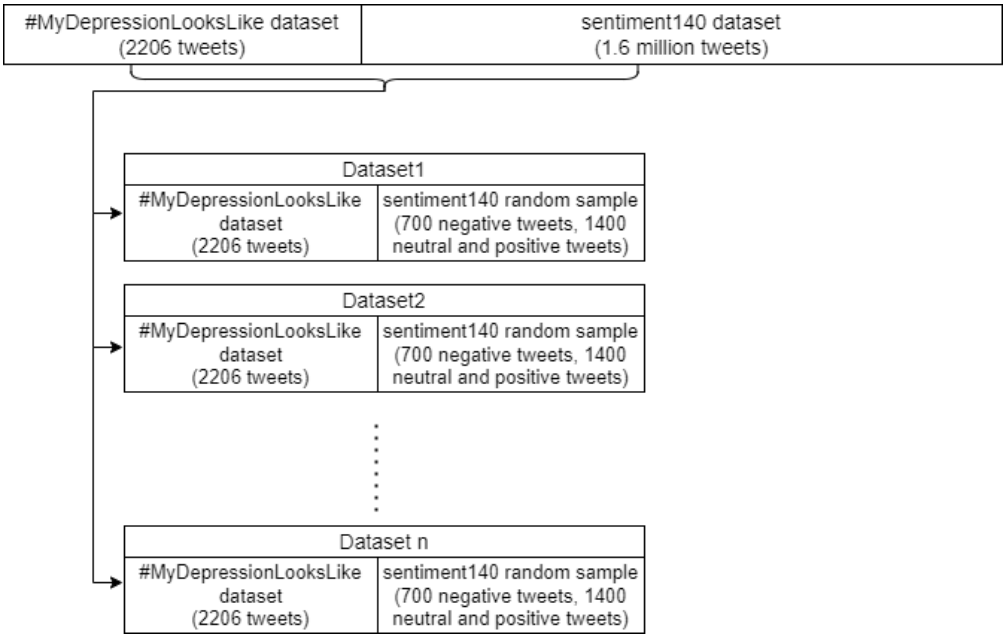


Figure 2.2: Random Balanced Datasets Construction

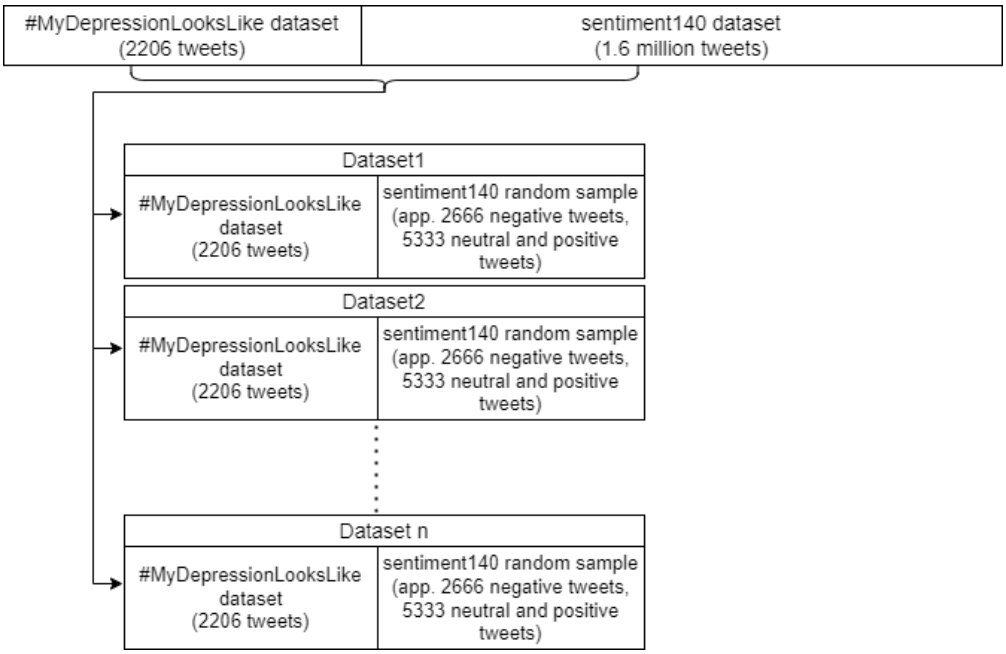


Figure 2.3: Random Imbalanced Datasets Construction

2.1.3 Pre-Processing

We expecting linguistic markers to identify dysfunctional thought categories [37]. However, tweets include links, mentions, emojis, and hashtags. Hence, we removed all links, mentions, emojis, and hashtags using *regex*. Tweets may also contain bad Unicode text such as mojibake (encoding mix-ups). We used *ftfy* to correct unicode errors, keeping only the main thought [43].

2.1.4 Feature Matrix Construction

The feature matrix consists of n rows and m columns, where n is the number of tweets and m is the number of features. Each row is a feature vector that is presented as an input to the classifiers. The details of each dataset, n and m is described in 3.2 and 2.1.2.
For this research, we have constructed two different features sets as follows:

1. Tweets were embdded using SBERT.
2. Differential embeddings that denote depression symptoms. This differential embedding represents the distance between two vectors in the embedding space of the depression symptoms exemplary sentences and tweets.

2.1.4.1 Depression Symptoms Features Extraction

In [2], ten dysfunctional categories are included to study if these categories can be identified by linguistic markers. These categories are All-or-nothing thinking, Negative predictions, Disqualifying the positive, Emotional reasoning, Labeling, Magnification, Mind reading, Overgeneralization, "Should" thinking, and Personalization. Based on the feasibility test that was performed by Lachmaret al., seven of the dysfunctional thought categories have the highest matches with the linguistic markers that were developed as a part of their system. In this study, we included the seven dysfunctional thought types that were shown

to be identified correctly. The dysfunctional thought concept was incorporated in the depression detection task by the means of differential embedding. Examples of the selected dysfunctional thought categories are listed in Table 2.1.

The other depression symptoms that we considered were provided in [38]. In this article, a set of depression symptoms were discussed. As we are looking for symptoms that could possibly appear in one's writing, only a subset of these symptoms were included in this research. More specifically, we included eight depression symptoms which are Loss of insight, Pleasure loss, Interest loss, Feeling bothered, Energy loss, Inability to feel, Feeling needed, and Feeling happy. Part of these symptoms show a positive correlation with depression, while the presence of others would indicate a non-depressed individual. For instance, pleasure loss is a symptom that would possibly indicate depression. On the contrary, feeling needed and feeling happy are not signs of a depressed individual. This difference may impact further expectations and analysis. Examples of the selected dysfunctional thought categories are listed in Table 2.2.

The differential embedding that is the difference between a tweet and dysfunctional thought anchors was generated as follows:

1. A set of representative sentences of dysfunctional thought categories was collected from psychological journals (A list of citations is available in Appendix Table 1). The limitation of taking the sentences only from such journals was to ensure that they were identified by experts. For each category, nine sentences were collected. In addition to dysfunctional thought patterns, a set of representative sentences was generated using keywords to represent other depression symptoms. For each depression symptom, four to five sentences were generated.

2. The set of depression symptoms representative sentences was represented in an embedding a pace using SentenceBERT. SentenceBERT derives semantically meaningful sentence embedding that can be averaged or compared using cosine-similarity for further uses. The pre-trained model we used in this research is bert-base-nli-mean-tokens which encodes sentences/texts in 768-dimension vectors.

Figure 2.4 and 2.5 show plots of the first and second dimensions of Multi-Class Linear

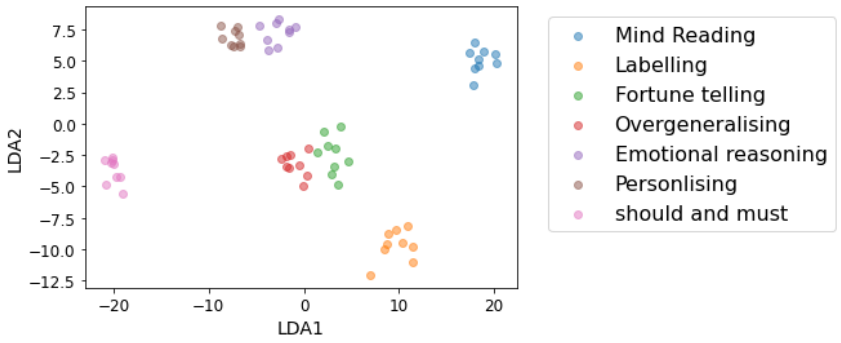


Figure 2.4: First and second dimensions of MDA on SBERT embedding space that represent dysfunctional thought anchors

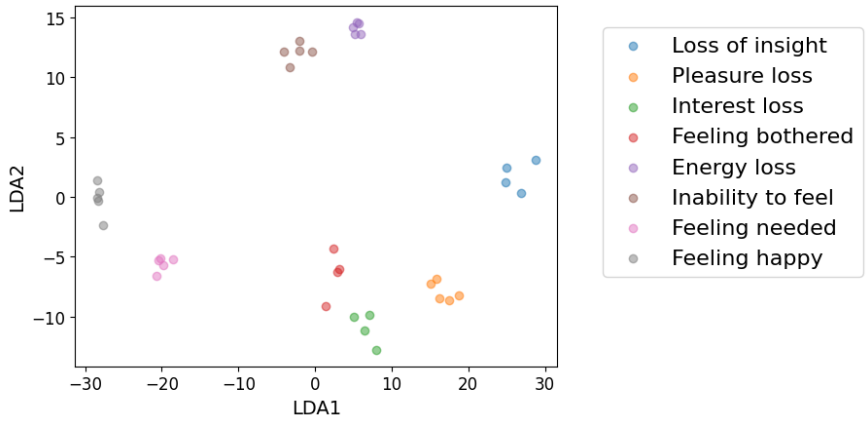


Figure 2.5: First and second dimensions of MDA on SBERT embedding space that represent other depression symptoms anchors

Discriminant on SBERT embedding space. They show the separation between dysfunctional thought categories and between other depression symptoms, respectively.

Depression symptoms were incorporated into the depression detection task by the means of

differential embedding. To get the differential embedding, an encoded tweet in a dataset is considered to be one vector and the other vector is the embedding of an anchor point in the depression symptoms dataset. Therefore, for each of the depression symptom in the dataset, we encoded and selected the depression symptom's exemplary sentence that is the closest sentence to a tweet to serve as a feature in the features vector. We tested four approaches to incorporate the dysfunctional thought categories as described below:

- We used the cosine distance between two embedding's vectors (tweets and selected anchor points) to generate the features vectors.
- Instead of cosine distance scores, we concatenated the differential embeddings of the depression symptoms that we selected earlier, based on the distance, and used them as one feature vector.
- We calculated the mean of the differential embeddings of the depression symptoms.
- We made use of one of features selection algorithms to get the most contributed depression symptoms.

Further details and explanation can be found in 3.1 and 3.2.

2.2 Features Exploration and Visualization

To better understand the similarity between depression symptoms anchors and tweets in our dataset, we measured the cosine distance between a tweet and the closest anchor point to all depression symptoms in a balanced dataset. The result is a feature vector that contains the distances to 15 anchor points. Each represents one depression symptom.

In Figure 2.6, we show the distribution of the values that capture the cosine distance scores between the tweets and their closet anchor points to the "Emotional Reasoning" symptom. The skewed histogram suggests that emotional reasoning language tends to appear more in

tweets that show depression symptoms. This can be inferred by looking at the distances that are reserved in each bin and the number of depression-indicative tweets that each bin represents. This provides some evidence that this category could be discriminative and be used for the prediction task.

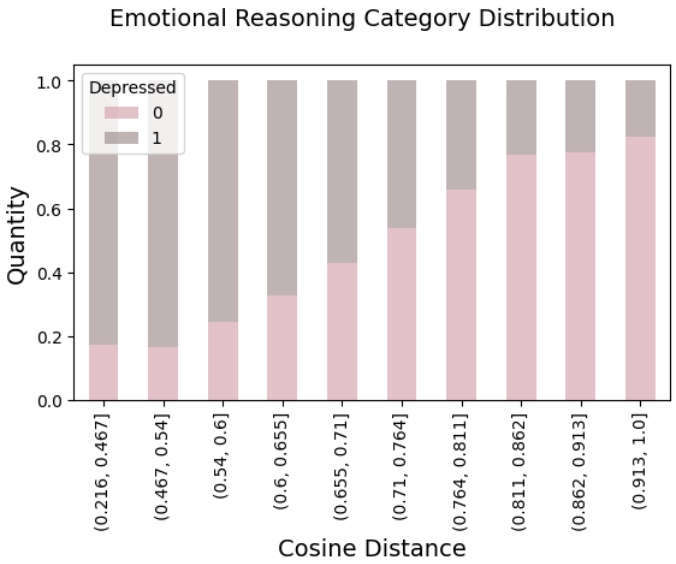


Figure 2.6: Cosine distance distribution of the Emotional Reasoning category - 0: "No depression symptoms", 1: "Shows Depression symptoms"

On the other hand, Figure 2.7 shows the distribution of the cosine distance scores between the tweets and their closet anchor points to the “Feeling happy” symptom. Looking at the histogram, unlike the histogram in 2.6, we see that we have higher number of non-depressed instances when the cosine distance score gets the closest to 0. This observation means we rarely have tweets that show happy feelings within the depressed class. Though the observation is not surprising, it suggests that this depression symptom is also can be used for the prediction detection task for its negative correlation.

Figure 4.1 in appendix shows the distribution of the 15 depression symptoms.

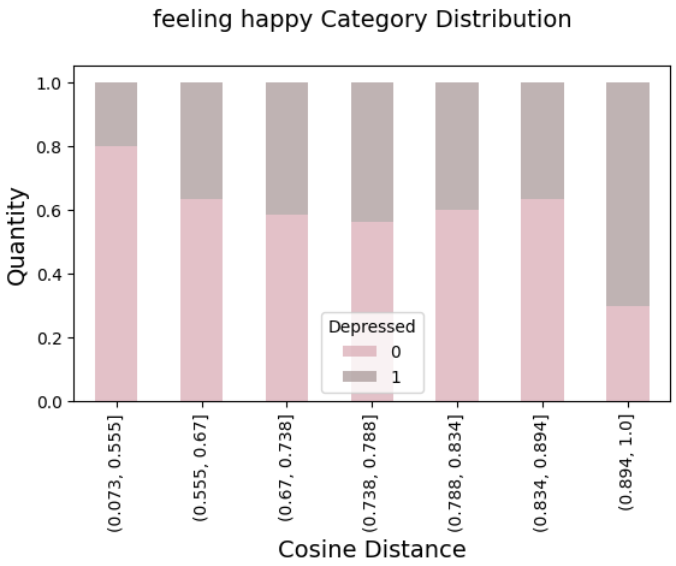


Figure 2.7: Cosine distance distribution of "Feeling happy" depression symptom - 0: "No depression symptoms", 1: "Shows Depression symptoms"

2.3 Supervised Classification Using SBERT Embedding

This machine learning task is a standard binary classification problem i.e., to predict if a tweet contains any depression symptoms, or not.

2.3.1 Logistic Regression

We tested a simple model (logistic regression). We evaluated how different independent variables effect the outcomes by training different features sets as described in 2.1.4.1. One of the main disadvantages of logistic regression is the overfitting that may occur as the number of observations approaches the number of features; to avoid overfitting, we utilized L1/L2 regularization in addition to dimensionality reduction via PCA.

2.3.2 Performance Measurements

Any dataset that exhibits an unequal distribution between its target classes can be considered imbalanced. Commonly, imbalanced data refers to datasets that exhibit significantly or even extremely unequal class distribution. In such cases, we require a classifier that will provide high accuracy for the minority class, without severely jeopardizing the accu-

racy in the majority class. This also suggests that the conventional evaluation practice, such as the overall accuracy or error rate, does not provide adequate information in the case of imbalanced learning. Therefore, more informative assessment metrics such as the receiver operating characteristics (ROC) curves, precision-recall curves, and cost curves, are necessary for conclusive evaluations of performance in the presence of imbalanced data. In this research we used AUROC and AUC-PR to evaluate the performance of the supervised learning models.

Chapter 3: Experiments and Results

For this research, all development was done in Python. Classifiers were trained and evaluated using 10*10-fold cross validation to help assess over-fitting. The results presented in this chapter are for the test set only.

3.1 Experimental Settings

To answer (1) if we can leverage the strength of SOTA deep learning techniques without any training or fine-tuning while maintaining clinical relevance in a semi-supervised depression detection system and (2) if dysfunctional thought patterns can be effectively used in a depression detection system, we used SBERT embedding vectors and dysfunctional thought patterns (as depression symptoms). More specifically, we used SBERT embedding vectors to measure the similarities between tweets and clinically backed anchor points as distance in the vector space and fed them directly into the machine learning model.

3.1.1 Datasets

To assess the statistical integrity of our results, we re-performed our experiments on 40 random samples of the data. We then compared the distributions of the results obtain using our approach with those using the baseline method. In this way, we were able to test for the statistical significance of any differences in performance that were observed. The random samples were drawn from the two datasets described in Section 2.1.2 as follows: (1) datasets with balanced classes: each random sample contained a total of ≈4,216 tweets; 2,108 tweets were taken from the #MyDepressionLooksLike hashtag, and the remaining were drawn randomly from random tweets dataset (30% negative tweets). (2) datasets with imbalanced classes: each random sample contained a total of ≈10,108 tweets; 2,108 tweets were taken from the #MyDepressionLooksLike hashtag, and the remaining were drawn randomly from

random tweets dataset (30% negative tweets).

3.1.1.1 Generalization

One of the useful properties of machine learning is the ability to create a model that can generate accurate predictions for a certain task. An effective machine learning model has the ability to make predictions on not only the data that it has seen but also on data that it has not seen. In a binary classification problem, we can assume there is a perfect model or function that can discriminate between two classes. In the context of a given problem, the perfect discriminant function is likely to have profound relevance to the domain experts. When we build a predictive model, we want to understand that relevance and try to best approximate this perfect discriminant function.

To approximate the perfect discriminant function, we use a sample, or a subset, of all possible data collected from the domain. This data contains the structure that is appropriate for the ideal discriminant function. When we prepare the data, we do so in a way that best exposes this structure to the predictive model. However, the data also includes information that is not related to the discriminant function, such as biases caused by the selection of the data and random fluctuations that disguise the underlying structure. For this reason, we aim to create a predictive model that does not model all the noise in the sample but generalizes beyond the seen data.

To evaluate a model's ability to generalize from the sample of data, we use data that the model has not seen before or during training. The problem with evaluating using a sample of data that the model was trained on is that doing so prohibits awareness of how well the model will perform on new, unseen data. If a selected model is chosen for its perfect accuracy on the training dataset rather than on unseen test data, it is very likely that the model will perform poorly on unseen data. This phenomenon, called overfitting, occurs because the

model was trained to recognize a specific structure in the training dataset [44].

3.1.1.2 The Overfitting Problem

To deal with overfitting, the dataset is divided into training and test datasets. The predictive model is created using a portion of the training dataset while the model's perfomance is tested using the unseen test dataset. Another way to deal with overfitting is through cross-validation. A common example of the use of cross-validation is 10-fold cross-validation. In 10-fold cross-validation, the dataset is split into 10 portions and the algorithm is run 10 times. In each run, the model is trained on 90% of the data and tested on the remaining 10%. The 10% testing portions are different in each run.

In this study, all experiments were evaluated using 10*10-fold cross-validation, and the results reported on the test set.

3.1.2 The Baseline

The baseline for our experiments was the SBERT embedding vectors of tweets fed directly into a machine learning model. This is because the main contribution of this research is the leveraging of SOTA representation learning techniques without model training where 1) the interpretability of the model is important, i.e, in clinical contexts, and 2) there is not enough labelled data.

We used SBERT embedding vectors of the representative sentences in the depression symptoms dataset and computed new vectors using subtract, average, and concatenate vectors operations.

3.1.3 Distance Scores

Cosine similarity is often used in text analysis to measure how similar two documents are, regardless of their size. This score determines whether two vectors are pointing roughly

Table 3.1: Sorted cosine distances to depression symptoms' representative sentences generated by applying the method on the sentence "I always fail"

Depression symptom	Cosine Distances
Overgeneralizing	[0.302, 0.307, 0.363, 0.376]
Labeling	[0.308, 0.406, 0.536, 0.538]
Fortune telling	[0.4, 0.454, 0.531, 0.623]
Personalizing	[0.467, 0.533, 0.565, 0.57]
Inability to feel	[0.472, 0.491, 0.524, 0.605]
Emotional reasoning	[0.481, 0.493, 0.54, 0.689]
Pleasure loss	[0.488, 0.627, 0.673]
Loss of insight	[0.506, 0.508, 0.543]
Interest loss	[0.528, 0.544, 0.799, 0.82]
Mind reading	[0.554, 0.605, 0.624, 0.641]
Feeling bothered	[0.582, 0.641, 0.657, 0.68]
Energy loss	[0.584, 0.734]
Loss of insight	[0.607]
Shoulds and musts	[0.648, 0.686, 0.745, 0.782]
Feeling needed	[0.761, 0.873, 0.951]
Feeling happy	[0.879, 0.903, 0.907]

in the same direction. In this experiment, we first measured the similarity (i.e., cosine distance) between a tweet's SBERT embedding and the SBERT embedding of all representative sentences expressing depression symptoms. Table 3.1 presents an example measuring the cosine distance between the encoded "I always fail" sentence and the embedding of 15 depression symptoms (dysfunctional thought patterns included). In the context of dysfunctional thought patterns, the main category of this type of sentence is identified by an expert as 'overgeneralizing thinking pattern'. The table shows the depression symptom and the cosine distance scores from the given sentence to the representative sentences of all depression symptoms (only the top four distances are represented in the table, if available). If a depression symptom shows fewer than four scores, it means there were not enough representative sentences of that symptom similar to the given sentence (cosine distance < 1).

Once we had the cosine distance scores sorted for each depression symptom, we generated features vectors by concatenating the cosine distance scores to the closest sentences in each depression symptom. To illustrate using 3.1, if we wanted to generate a features vector for "I always fail," then the values [0.302, 0.308, 0.4, 0.467, 0.472, 0.481, 0.488, 0.506, 0.528 , 0.554, 0.582, 0.584, 0.607, 0.648, 0.761, 0.879] would be the features vector input to the predictive model.

3.1.4 Mean and Concatenation of Differential Embedding

We also carried out experiments using models fed embedding vectors. This approach was inspired by the mathematical operations that can be applied to word embedding vectors. Various operations can be used to obtain new embedding vectors, such as sum, average, and concatenation. Figure 3.1 illustrates the latter two approaches.

Following the techniques that are used on word embedding to get document-level embeddings, we used the averaging and concatenation approaches in this research to get the final features vectors for each tweet. The concatenation approach is expected to be more informative than the mean as it maintains the original representation of all the depression symptoms. Table 3.2 provides a summary of the approaches and respective features

Figure 3.1 illustrates the latter two approaches.

3.1.5 Concatenated Differential Embedding of Selected Features

We implemented a features selection method (i.e., Sequential Features Selection (SFS) algorithm) to 1) determine what features we should exclude/include to improve performance and 2) get a general idea of which depression symptoms can be used effectively in a depression detection system.

A traditional SFS method works by considering every column in the dataset as a feature.

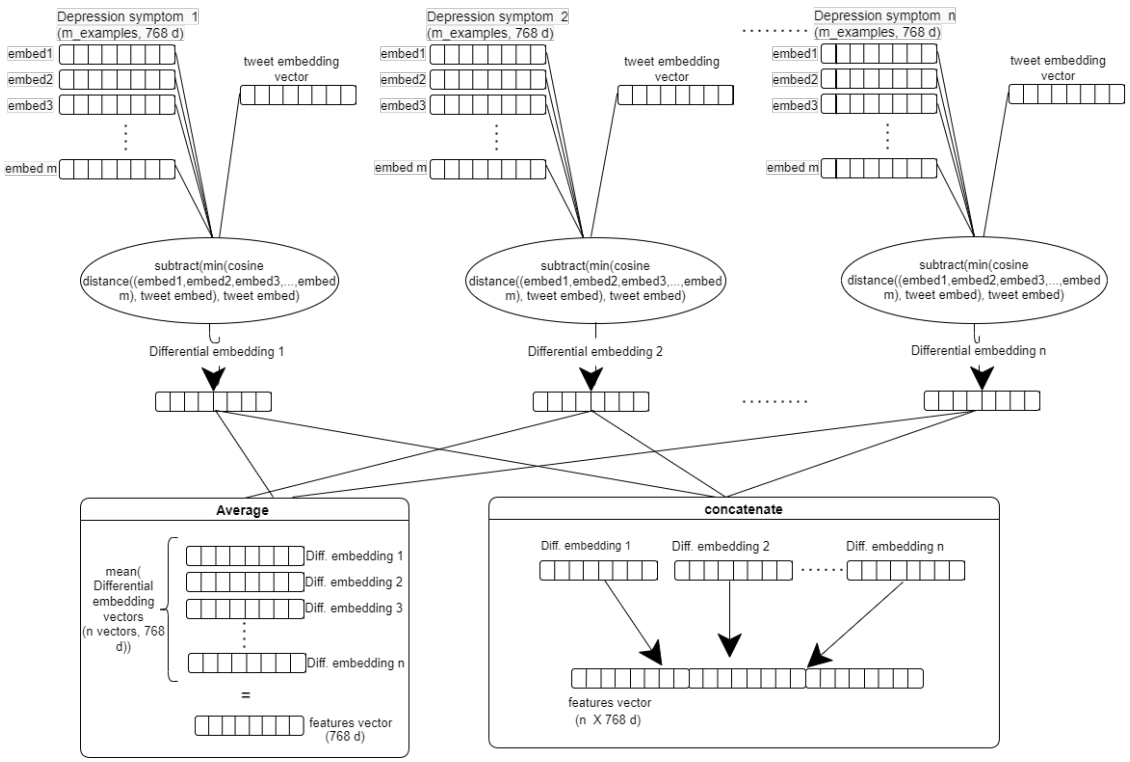


Figure 3.1: Overview of the averaging and concatenating approaches to obtaining new embedding vectors (features vector).(1) Measure the distance scores between a tweet and all representative sentences of depression symptoms.(2) Choose the sentence that has the smallest score.(3) Subtract operation to get the differential embedding between the tweet's embedding vector and the embedding vector of the chosen sentence.(4) Generate the features vector using average and concatenation

However, in our dataset, a feature is a depression symptom and every feature/depression symptom is a set of 768 dimensions/columns in the dataset. Therefore, we modified the original algorithm to account for this special case.

We applied SFS to 50 different random datasets that we created in 3.1.1. In all experiments, SFS consistently selected two depression symptoms: "mind reading" and "feeling happy". Figure 3.2 shows the set of features and the respective performance in each iteration.

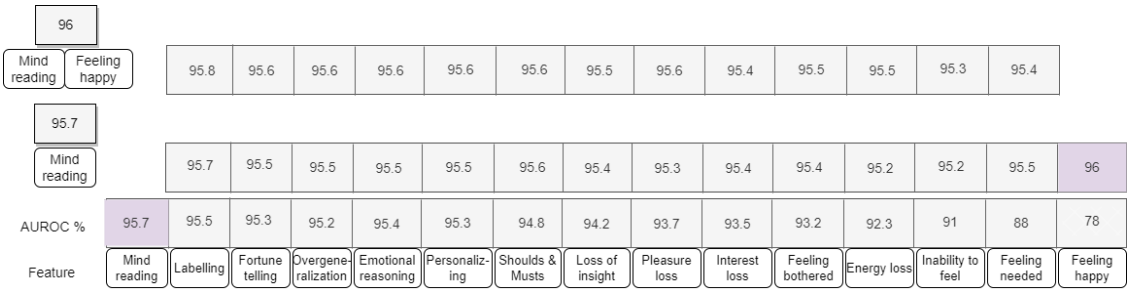


Figure 3.2: Features selection using SFS. Starting from the bottom: depression symptoms are in white, gray squares represent the AUROC of the included depression symptoms starting from an empty features set, and the AUROC of the best feature in each iteration is identified by the color purple

Table 3.2: Summary of the approaches. Encoder refers to the transformer used to generate the original embedding, and #Features refers to the original number of features before applying dimension reduction methods. Diff. emb: Differential embedding, Dep. symptoms: Depression symptoms

Approach	Type of Features	#Features
SBERT - baseline	BERT-generated emb	768
Distance scores	Cosine distance scores	#Dep. symptoms = 15
Average	Mean of diff. emb	768
Concatenation	Concatenation of diff. emb	768 X #Dep. symptoms = 11,520
Features selection	Concatenation of diff. emb of selected features	768 X #Selected features = 1,536

3.2 Supervised Learning

The classification task began by applying dimension reduction (i.e, principal components analysis) before using both concatenation approaches. To concatenate differential embeddings, we increased the features space while maintaining the number of training samples. This may lead to overfitting, which is why we implemented dimension reduction. In approach 3 and 4, we used 350 and 280 principal components, respectively.

We implemented a logistic regression model on each of the 50 random datasets that we created (described in 3.1.1).

We used L2 regularization and 10*10-fold cross-validation in all LR models.

3.2.0.1 Comparing two models

Comparing machine learning approaches and choosing a final model are frequent operations in applied machine learning. Common resampling techniques for model evaluation include k-fold cross-validation, from which the mean of a model's performance can be derived and directly compared to the means of other models'. Although straightforward, this method may be deceptive because it is difficult to determine if the difference in a mean model's performance is real or a statistical fluke. A difference in a model's performance is considered statistically significant if the null hypothesis, or the assumption, is rejected.

A statistical hypothesis test measures the probability of observing two data samples assuming that they have the same distribution. The null hypothesis is a presumption that underlies a statistical test, and we can compute and analyze statistical measures to determine whether or not to accept or reject the null hypothesis. In the case of selecting models based on their estimated performance, we are interested in knowing whether there is a real or statistically significant difference between the two models. There are two possible outcomes in the comparison of models: 1) if the test's outcome indicates that there is not enough data to rule out the null hypothesis, then any observed difference in model performance is probably the consequence of statistical chance, and 2) if the test's outcome indicates that there is enough evidence to reject the null hypothesis, then any observed difference in model performance is probably caused by a difference in the models [11].

The products of a statistical test are a test statistic and p-value, which can both be interpreted and used to quantify the degree of confidence or significance in the difference between models.

In addition to reporting performance metrics, we also report the p value that indicates if the

difference between two models is statistically significant. In this research, the threshold of the p value was set to .05.

Figure 3.3 summarizes the LR results. We measured model performance using the AUROC scores for each approach. Figure 3.3 illustrates that, across datasets, using the baseline and concatenating the differential embedding of depression symptoms show relatively similar performances. However, using the concatenated differential embedding of the selected features.

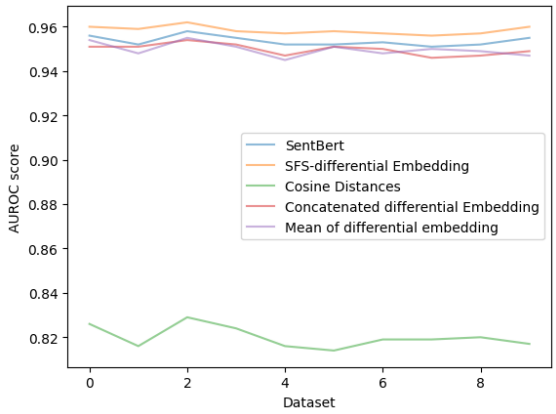
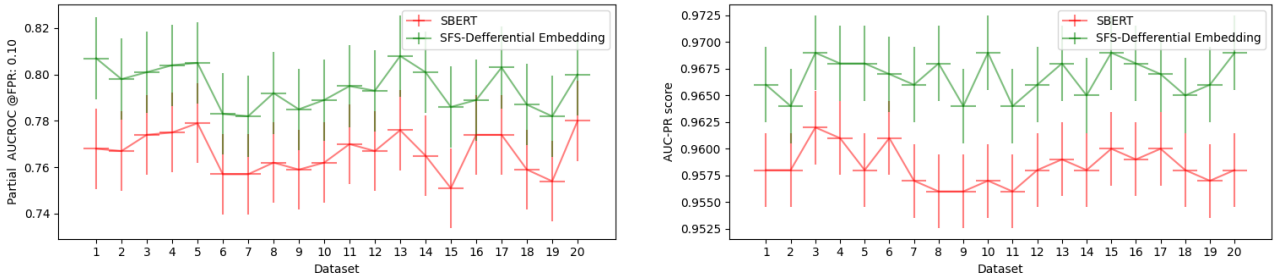


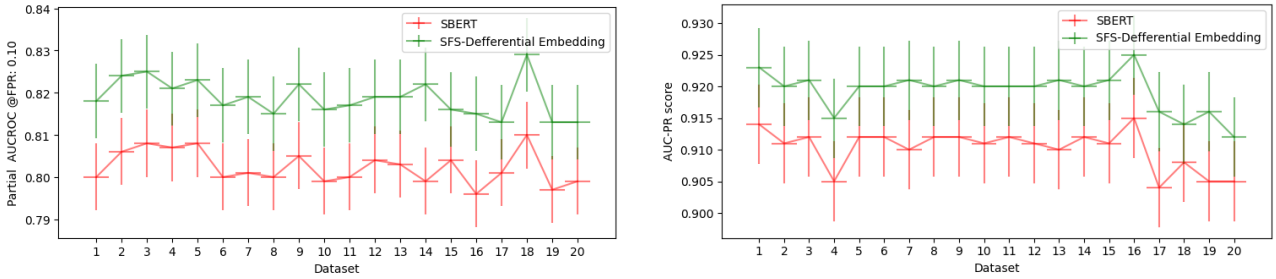
Figure 3.3: LR results using different approaches on 10 random samples

Using the baseline and the best performing model from the previous test, we reported a precision-recall (PR) curve and Partial ROC curve at low positive rate (FPR: 0.10) for the LR model trained on 20 different random datasets (Figure 3.4). The differences in the means of the PR and ROC scores were statistically significant ($p < 0.05$) for all datasts.



a) b)

Figure 3.4: a) Partial AUROC scores @FPR: 0.10 and b) AUC-PR scores across different random balanced datasets



a) b)

Figure 3.5: a) Partial AUROC scores @FPR: 0.10 and b) AUC-PR scores across different random imbalanced datasets

3.3 Discussion

In this study, we proposed a constrained few-shot learning model that makes use of SOTA representation learning techniques and clinically relevant depression symptoms. We used this model in supervised settings on 50 random samples.

3.3.1 Supervised Classification Using SBERT Embedding and Differential Embedding

For standard binary classification, 3.3 shows that the proposed model performs best at low false positive rate when using the concatenation of the selected-feature differential embedding. Although it is a relatively small improvement, this performance is based on a constrained depression detection model. More specifically, the depression detection task is constrained to the presence of depression symptoms rather than a black box detection task (i.e, deep learning models). This distinction makes the proposed model more reliable, with a performance comparable to that of SOTA models (i.e., SBERT).

To assess the integrity of the results, we reperformed the experiments using 20 random balanced datasets and 20 random imbalanced datasets. 3.4 and 3.5 show that the improvement is consistent and statically significant across multiple random balanced/imbalanced datasets ($p < 0.05$) for all reported results.

3.3.2 Interpretation of the results

The best-performing model is one that uses two depression symptoms that are the result of Forward-SFS (i.e., "mind reading" and "feeling happy"). One can intuitively recognize that "mind reading" is positively correlated with depression, while the absence of "feeling happy" is a depression indicator. 3.2 presents the AUROC scores for individual depression symptoms. We observed that "feeling happy" does not perform well by itself which is likely due to having negative tweets in the non-depressed class. Therefore, the "feeling happy" symptom is not informative if used alone. This observation confirms the necessity of including negative tweets in the non-depressed class, as this changes the way we extract features. Also, it poses a challenge to get more informative features.

Although "feeling happy" is not an informative depression symptom that we considered

alone, it can significantly improve the model's performance when combined with "mind reading." Combining two correlated features that measure different characteristics provides complementary information to the predictive model.

Self-reported depression symptoms that can be identified by asking questions, which implies that the user is aware of their mental health distress, have been the focus of recent works on constrained depression detection systems [5], [8]. However, we based this study on depression symptoms that can be inferred from one's writing (i.e., dysfunctional thought patterns) rather than self-reported symptoms, as we hope this approach will help detect depression symptoms at early stages, even if a user is not aware of their symptoms. Although one of the best features comes from dysfunctional thought patterns, and the other comes from a set of self-reported depression symptoms, the distribution of the distance scores of these symptoms shows less discrimination between the depressed and non-depressed classes 4.1. In addition to the difficulty of identifying self-reported symptoms in one's writing, this research was limited by using representative sentences to express these symptoms, unlike dysfunctional thought patterns, where there are confirmed representative sentences.

3.3.3 Comparing to literature review

Although we propose a few-shot learning method, similar to other recent studies, we included a different set of depression symptoms that do not rely on the presence or absence of specific words and do not require fine-tuning and training [5]. A similar study was undertaken to constrain the behavior of depression detection methods by the presence of symptoms known to be related to depression (i.e., clinically backed symptoms), while producing a model that is easy to inspect. However, we observed certain shortcomings in this work because the authors manually selected only neutral and positive examples for the non-depressed class [8]. To properly address this issue, we proposed a constrained depression detection system and

carried out various experiments that included negative tweets. We showed that including negative tweets imposed a challenge and changed how well some features performed.

Chapter 4: Conclusion

4.1 Summary and Conclusion

In this research, we attempted to strike a balance between expert-defined features and machine-learned features; we represent 15 depression symptoms as Sentence-BERT embeddings which are the results of encoding a set of representative sentences of each symptom. To train a depression-detection system, we then used the cosine distance to measure the similarity between the tweets and the depression symptoms. Additionally, we used the differential embedding that is the difference between tweets embedding vectors and the depression symptoms representative sentences' embedding vectors. The results support the theory that depressed individuals on social media use dysfunctional thought patterns more than individuals with no depression symptoms. This research shows that we can perform a classification task based on clinically relevant depression symptoms.

Additionally, this research presents a methodology with which we express and incorporate depression symptoms in a depression detection system by the means of differential embedding. We showed that the proposed methodology outperformed SOTA embedding generation techniques (i.e, SBERT).

4.2 Future Work

The current research bases the experiments on a set of depression symptoms and their representative sentences that were collected and generated manually. Although the current set yields good results, a method to automatically generate representative sentences would serve in providing a variety of sentences and not be limited to what we find online.

Contrastive learning is a technique that has been used recently to train a model to learn representations of sentences such that similar samples are closer in the vector space. Investi-

gating this approach for the purpose of getting the representations of the anchor points would be an important next step and expected to achieve better performance. This technique is also expected to perform well in the unsupervised learning aspect of this research.

BIBLIOGRAPHY

[1] "Depression," Jan 2020.

[2] E. M. Lachmar, A. K. Wittenborn, K. W. Bogen, and H. L. McCauley, "#mydepressionlookslike: Examining public discourse about depression on twitter," *JMIR Ment Health*, vol. 4, Oct 2017.

[3] H. S. AlSagri and M. Ykhlef, "Machine learning-based approach for depression detection in twitter using content and activity features," *IEICE Transactions on Information and Systems*, vol. 103, no. 8, pp. 1825–1832, 2020.

[4] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, "A textual-based featuring approach for depression detection using machine learning classifiers and social media texts," *Computers in Biology and Medicine*, vol. 135, p. 104499, 2021.

[5] N. Farruque, R. Goebel, O. R. Zaiane, and S. Sivapalan, "Explainable zero-shot modelling of clinical depression symptoms from text," *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2021.

[6] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. 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*, (New Orleans, LA), pp. 88–97, Association for Computational Linguistics, June 2018.

[7] A. H. Orabi, P. Buddhitha, M. H. Orabi, and D. 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*, pp. 88–97, 2018.

[8] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan, "Improving the generalizability of depression detection by leveraging clinical questionnaires," *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2022.

[9] B. I. C. Education, "What is natural language processing?."

[10] "Supervised learning vs deep learning: Learn top 5 amazing differences," Mar 2021.

[11] J. Brownlee, "14 different types of learning in machine learning," Nov 2019.

[12] J. Brownlee, "Discover feature engineering, how to engineer features and how to get good at it," Aug 2020.

[13] "What is the difference between deep learning and machine learning? quantdare," Dec 2019.

[14] I. C. B. Owner, L. B. Developer, N. J. I. Architect, J.-L. M. Professor, and J. H. S. D. Scientist, "Deep learning for natural language processing," Oct 2021.

[15] R. Horev, "Bert explained: State of the art language model for nlp," Nov 2018.

[16] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," 2019.

[17] "1.13. feature selection."

[18] "Perfect recipe for classification using logistic regression."

[19] "Precision-recall."

[20] "Classification: Roc curve and auc nbsp;|nbsp; machine learning nbsp;|nbsp; google developers."

[21] J. Brownlee, "Failure of classification accuracy for imbalanced class distributions," Jan 2021.

[22] 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*, pp. 51–60, 2014.

[23] M. R. Islam, M. A. Kabir, A. Ahmed, A. R. M. Kamal, H. Wang, and A. Ulhaq, "Depression detection from social network data using machine learning techniques," *Health information science and systems*, vol. 6, no. 1, pp. 1–12, 2018.

[24] T. Shen, J. Jia, G. Shen, F. Feng, X. He, H. Luan, J. Tang, T. Tiropanis, T.-S. Chua, and W. Hall, "Cross-domain depression detection via harvesting social media," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pp. 1611–1617, International Joint Conferences on Artificial Intelligence Organization, 7 2018.

[25] F. Cacheda, D. Fernandez, F. J. Novoa, and V. Carneiro, "Early detection of depression: Social network analysis and random forest techniques," *J Med Internet Res*, vol. 21, p. e12554, Jun 2019.

[26] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, 2013.

[27] X. Tao, X. Zhou, J. Zhang, and J. Yong, "Sentiment analysis for depression detection on social networks," in *International Conference on Advanced Data Mining and Applications*, pp. 807–810, Springer, 2016.

[28] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media," *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, pp. 128–137, Aug. 2021.

[29] S. Tsugawa, Y. Kikuchi, F. Kishino, K. Nakajima, Y. Itoh, and H. Ohsaki, "Recognizing depression from twitter activity," pp. 3187–3196, 04 2015.

[30] H. Zogan, X. Wang, S. Jameel, and G. Xu, "Depression detection with multi-modalities using a hybrid deep learning model on social media," *arXiv preprint arXiv:2007.02847*, 2020.

[31] Melinda, "Depression treatment," Oct 2021.

[32] N. Schimelpfening, "How to positively conquer common cognitive distortions," Mar 2020.

[33] W. Irwin and G. Bassham, "Depression, informal fallacies, and cognitive therapy," *Inquiry: Critical Thinking Across the Disciplines*, vol. 21, no. 3, p. 15–21, 2003.

[34] I. M. Blackburn and K. M. Eunson, "A content analysis of thoughts and emotions elicited from depressed patients during cognitive therapy*," *British Journal of Medical Psychology*, vol. 62, no. 1, p. 23–33, 1989.

[35] A. T. Beck, *Cognitive therapy and the emotional disorders*. Penguin Books, 1991.

[36] M. J. Fennell and E. A. Campbell, "The cognitions questionnaire: Specific thinking errors in depression," *British Journal of Clinical Psychology*, vol. 23, no. 2, p. 81–92, 1984.

[37] K. Wiemer-Hastings, A. S. Janit, P. M. Wiemer-Hastings, S. Cromer, and J. Kinser, "Automatic classification of dysfunctional thoughts: a feasibility test," *Behavior Research Methods, Instruments, & Computers*, vol. 36, no. 2, pp. 203–212, 2004.

[38] E. I. Fried, J. K. Flake, and D. J. Robinaugh, "Revisiting the theoretical and methodological foundations of depression measurement," *Nature Reviews Psychology*, vol. 1, no. 6, p. 358–368, 2022.

[39] "There are 6 depression datasets available on data.world.."

[40] Möbius, "The depression dataset," Feb 2021.

[41] Isanbel, "Depression on twitter," Jan 2020.

[42] V. Romero, "Detecting-Depression-in-Tweets," 2 2019.

[43] R. Speer, "ftfy." Zenodo, 2019. Version 5.5.

[44] J. Brownlee, "A simple intuition for overfitting, or why testing on training data is a bad idea," Aug 2016.

[45] K. Cherry, "Cognitive psychology is the science of how we think," Feb 2022.

[46] N. Dinovitz, "Can you read minds? if not, stop trying.... - dinovitz counseling llc: Philadelphia and bala cynwyd therapist," Sep 2019.

[47] D. Bryan, *Panic Attacks Think Yourself Free: The Self-Help Book to Overcome Panic Attacks*. Xlibris, 2011.

[48] "Cognitive distortions and thinking errors: Mindreading."

[49] "Thinking traps: 12 cognitive distortions that are hijacking your brain."

[50] B. Elizabeth Hartney, "10 cognitive distortions you'll learn about in therapy."

[51] A. Bonfil, "Cognitive distortions: Labeling," May 2017.

[52] K. Cherry, "How cognitive behavior therapy works."

[53] "Thinking traps," Sep 2019.

[54] S. S. Casabianca, "15 cognitive distortions to blame for your negative thinking," Jan 2022.

[55] E. McAdam, "Skill 18 cognitive distortions part 1 cognitive behavioral therapy techniques," Aug 2021.

[56] "Midtown fellowship."

[57] M. Cortese, S. J. Hemmeter, and N. Schrandt. Practising Law Institute, 2008.

[58] D. D. Burns, *The feeling good handbook*. Penguin, 1999.

[59] R. J. Stanborough, "Cognitive distortions: 10 examples of distorted thinking," Dec 2019.

[60] "Cognitive distortions and thinking errors - how can cbt help?," Apr 2022.

[61] "Succeed socially.com: Free social skills guide for adults."

[62] K. Kost, "What are cognitive distortions and why should i care?," Aug 2021.

APPENDIX A: DYSFUNCTIONAL THOUGHT REPRESENTATIVE SENTENCES REFERENCES

Table 4.1: Dysfunctional thoughts representative sentences references

Representative Sentence	Reference
1) I just know that my therapist thinks I am a waste of his time 2) I am a total loser 3) I must have failed that test because I feel so bad about my performance 4) I feel anxious, so I know something dangerous is going to happen	[45]
5) he thinks I am a loser	[46]
6) John's in a terrible mood It must have been something I did 7) I could tell he thought I was stupid in the interview 8) I can tell they hate my shirt 9) It is obvious she does not like me, otherwise she would have said hello 10) This relationship is sure to fail 11) I feel hopeless, therefore my situation must be hopeless	[47]
12) He is ignoring me so he must not like me anymore	[48]
13) I knew they hated me 14) They are all making fun of me behind my back 15) She is bored of hanging out with me 16) I am an awkward person 17) I am a failure 18) I should not eat any junk food	[49]
19) I must be a worthless person	[50]
20) He is a jerk 21) She is irresponsible 22) He is an idiot 23) I am useless	[51]
24) I will never find love or have a committed and happy relationship 25) I will get rejected 26) I will make a fool of myself 27) I have got nothing done 28) I am going to fail everything	[52]
29) If I do not get out of here, I am going to faint 30) I am going to make a fool of myself and people will laugh at me 31) I always screw up 32) I must not fail 33) I must get over this fear 34) I should not have made so many mistakes	[53]
35) What if I have not turned the iron off and the house burns down 36) If I do not perform well, I will get the sack 37) My neighbor did not speak to me this morning, therefore I must have done something to upset them	

To be continued

Table 4.1 (cont'd)

Sentence	Reference
38) The world has got it in for me	[54]
37) I have always been like this; I will never be able to change 38) He did not want to go out with me, so I will always be lonely 39) I will never be asked on a second date 40) I have the worst luck in the entire world 41) My daughter failed her exam because I have not helped her	[55]
42) I never can speak publicly without messing up 43) We were late to the dinner party and caused everyone to have a terrible time	[56]
44) I am a terrible speaker and always screw up 45) I feel guilty, therefore I must have done something bad 46) I feel so depressed, this must be the worst place to work in 47) This shows what a bad mother I am 48) If only I were better in bed, he would not beat me 49) He should not be so stubborn and argumentative	[57]
50) I must be a complete loser and failure 51) I am not in the mood to do anything, therefore I might as well just lie in bed 52) I am furious with you, this proves that you have been acting badly and trying to take advantage of me	[58]
53) I am a horrible student and should quit school	[59]
55) My boss is irritable today so I must have annoyed her 56) It is my fault that my son is not studying 57) My husband hit me because I am a bad wife 58) It is all my fault that the meeting ran on so long	[60]
60) I should always give everything I do 100% 61) I must not be rude so others should not be either	[61]
62) I really should exercise 63) I should not be so lazy 64) I should pick up after myself more	[62]

APPENDIX B: DEPRESSION SYMPTOMS REPRESENTATIVE SENTENCES

Table 4.2: Depression Symptoms representative sentences

Depression Symptom	Representative Sentence
Mind reading	I just know that my therapist thinks I am a waste of his time
Mind reading	he thinks I am a loser.
Mind reading	John's in a terrible mood It must have been something I did
Mind reading	when you like that I know you were not telling the truth
Mind reading	he is ignoring me so he must not like me anymore
Mind reading	I knew they hated me
Mind reading	they are all making fun of me behind my back
Mind reading	she is bored of hanging out with me
Mind reading	It i obvious she does not like me, otherwise she would have said hello
Labelling	I am an awkward person
Labelling	I am a worthless person
Labelling	he is a jerk
Labelling	she is irresponsible
Labelling	I am a born loser
Labelling	I am a phony
Labelling	I am a failure
Labelling	He is an idiot
Labelling	I am useless
Fortune telling	i will never find love or have a committed and happy relationship
Fortune telling	I will get rejected
Fortune telling	I will make a fool of myself
Fortune telling	If I don not get out of here, I am going to faint
Fortune telling	I am going to make a fool of myself and people will laugh at me
Fortune telling	what if I haven not turned the iron off and the house burns down
Fortune telling	If I do not perform well, I will get the sack
Fortune telling	I have always been like this; I will never be able to change
Fortune telling	This relationship is sure to fail
Overgeneralising	I never can speak publicly without messing up
Overgeneralising	I have got nothing done
Overgeneralising	People are all mean and superficial
Overgeneralising	shopping will always be a stressful experience
Overgeneralising	I am going to fail everything
Overgeneralising	all sales clerks are rude
Overgeneralising	I always screw up

To be continued

Table 4.2 (cont'd)

Sentence	Reference
Overgeneralising	I must be a complete loser and failure
Overgeneralising	I am a horrible student and should quit school
Emotional Reasoning	I must have failed that test because I feel so bad about my performance.
Emotional Reasoning	I feel hopeless, therefore my situation must be hopeless
Emotional Reasoning	I feel guilty, therefore I must have done something bad
Emotional Reasoning	I am not in the mood to do anything, therefore I might as well just lie in bed
Emotional Reasoning	I am furious with you, this proves that you have been acting badly and trying to take advantage of me
Emotional Reasoning	I feel anxious, so I know something dangerous is going to happen
Emotional Reasoning	I feel so depressed, this must be the worst place to work in
Emotional Reasoning	my neighbour did not speak to me this morning, therefore I must have done something to upset them
Emotional Reasoning	my boss is irritable today so I must have annoyed her
Personalising	It is my fault that my son is not studying
Personalising	we were late to the dinner party and caused everyone to have a terrible time
Personalising	this shows what a bad mother I am
Personalising	I have the worst luck in the entire world
Personalising	My daughter failed her exam because I have not helped her
Personalising	My husband hit me because I am a bad wife
Personalising	if only I were better in bed he would not beat me
Personalising	It is all my fault that the meeting ran on so long
Personalising	the world has got it in for me
Shoulds and Musts	I should always give everything I do 100%
Shoulds and Musts	I must not fail
Shoulds and Musts	I must not be rude so other should not be either
Shoulds and Musts	I should not be so lazy
Shoulds and Musts	I should pick up after myself more
Shoulds and Musts	I should not eat any junk food
Shoulds and Musts	He should not be so stubborn and argumentative
Shoulds and Musts	I must get over this fear
Shoulds and Musts	I should not have made so many mistakes
Loss of insight	lack of understanding
Loss of insight	insufficient understanding
Loss of insight	lack of awareness
Loss of insight	false interpretation
Pleasure loss	I feel miserable
Pleasure loss	I feel unhappy

To be continued

Table 4.2 (cont'd)

Sentence	Reference
Pleasure loss	I feel sorrow
Pleasure loss	life is joyless
Pleasure loss	I feel distressed
Interest loss	I am finished with it
Interest loss	I am sick of it
Interest loss	everything is boring
Interest loss	I am done with it
Feeling bothered	it is disturbing
Feeling bothered	I feel irritated
Feeling bothered	I am pissed off
Feeling bothered	feeling upset
Energy loss	mentally drained
Energy loss	I can not leave my bed
Energy loss	I stay in bed all day
Energy loss	power draining
Energy loss	I feel energyless
Inability to feel	I am unemotional
Inability to feel	not being able to feel
Inability to feel	I feel heartless
Inability to feel	I feel unmoved
Inability to feel	I feel apathetic towards everything
Feeling needed	be valued at something
Feeling needed	feeling needed
Feeling needed	be wanted
Feeling needed	My family needs me
Feeling needed	I help my friends
Feeling happy	life is joyful
Feeling happy	I am happy
Feeling happy	I am in high spirits on the last day of school
Feeling happy	I am over the moon about being accepted to the university

APPENDIX C: DEPRESSION SYMPTOMS VALUES DISTRIBUTION

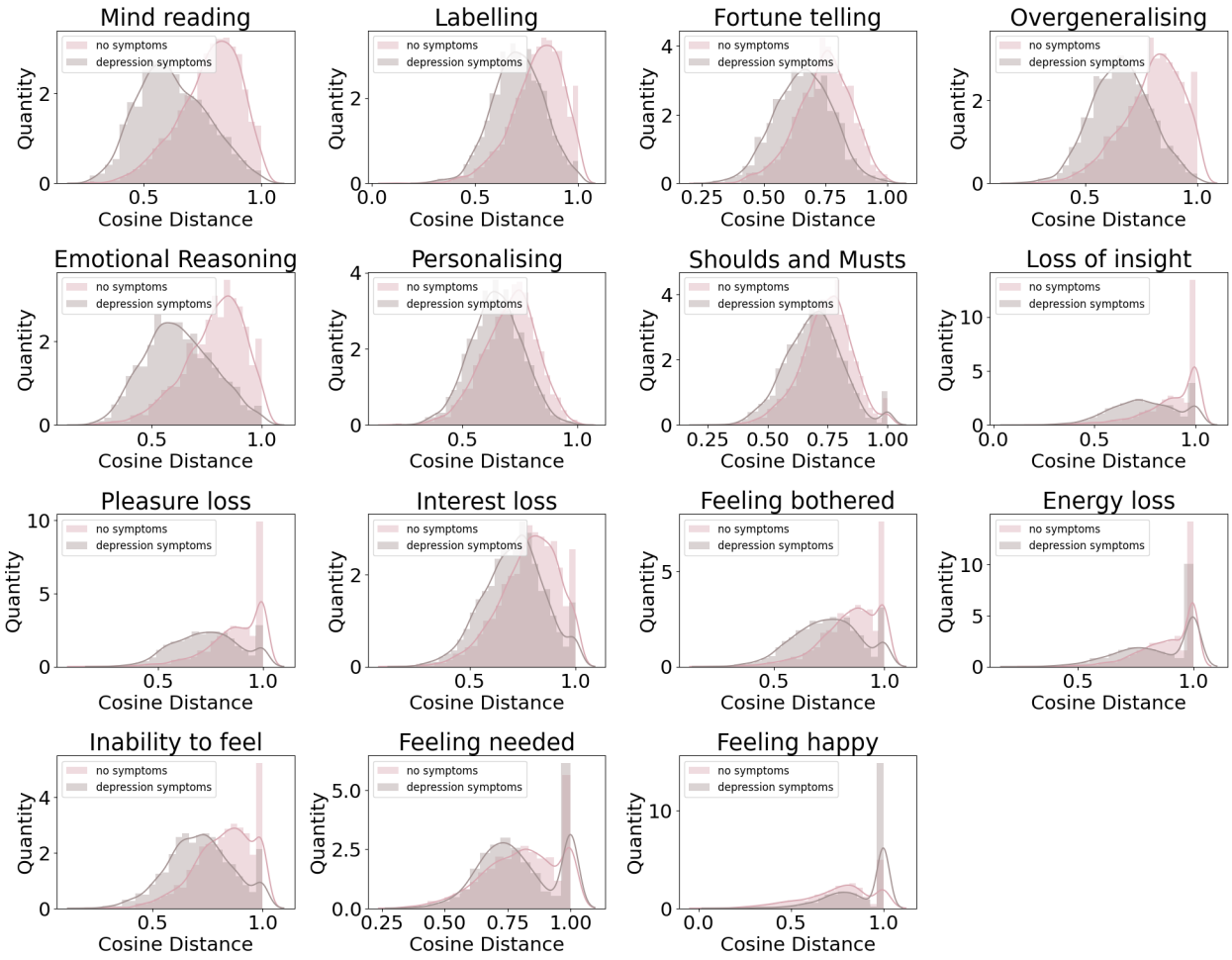


Figure 4.1: Cosine distances distribution of dysfunctional thought categories on a balanced dataset