Towards Natural Language Modelling of Clinical Depression

by

Nawshad Farruque

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Computing Science

University of Alberta

Abstract

Traditional survey based methods for clinical depression detection are not always effective; the patient may not reflect their actual mental health condition because of the cognitive bias exhibited while filling out questionnaires about depression. Established through ample earlier work, social media language has been found to be a reflection of a user's real-time mental health status. Being influenced by this potential of social media posts, in this dissertation, we describe a framework for natural language modelling of clinical depression from public social media posts, e.g., tweets from a Twitter user's timeline. Such modelling requires extraction of depression symptoms from the social media posts, then following clinical psychiatry guidelines to calculate depression scores for all two-weeks episodes; then, based on these scores, we infer whether a user is depressed or not. In this process, the first important challenge is the data scarcity for developing a Depression Symptoms Detection (DSD) model.

To address data scarcity, we follow two steps. First, we curate a Clinical Expert Annotated Depression Symptoms tweets (CEADS) dataset. We bring important innovations for curating a better quality of CEADS dataset that reflects both clinicians' insights and depression symptoms distribution of self-disclosing depressed Twitter users. Second, we train our DSD model using CEADS dataset and further make the model robust with the help of our proposed Semi-supervised Learning (SSL) framework. In this framework, we iteratively harvest depression symptoms tweets and re-train our DSD model. Moreover, we propose a Zero-Shot Learning

(ZSL) model to make our iterative data harvesting process more effective.

Further, with the help of the DSD model, we develop our Temporal User-level Clinical Depression Detection (TUD) model that can extract clinical depression scores through a user's Twitter timeline; much like what a depression rating scale, e.g., Patient Health Questionnaire - 9 (PHQ-9) would do.

Finally, we draw insightful conclusions on user-level clinical depression modelling by using the following: (1) depression score based features, (2) pure semantic representation based features, along with (3) their temporal representations and (4) experimentations with various clinical depression detection settings in several data distributions. To the best of our knowledge, our experimentations and analyses are unique in the literature.

Preface

Nawshad Farruque is the lead author responsible for problem formulation, design and development of the implementations, experimental evaluation and analysis and writing of this dissertation and the relevant published peer reviewed and under review papers listed below. These papers are further linked with the chapters of this dissertation. Most of the co-authors of these papers participated through providing feedback and participating in discussions. Otherwise, explicit acknowledgement of the co-author contribution is provided.

1. Refereed Conference and Workshop papers

   (a) Farruque, N., Zaïane, O., & Goebel, R. (2019, September). Augmenting semantic representation of depressive language: from forums to microblogs. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (ECML-PKDD) (pp. 359-375). Springer, Cham. (Chapter 3 contains the updated experiments from the paper, Chapter 4 uses the models proposed in the paper).

   (b) Farruque, N., Goebel, R., Zaïane, O. R., & Sivapalan, S. (2021, December). Explainable Zero-Shot Modelling of Clinical Depression Symptoms from Text. In 2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA) (pp. 1472-1477), IEEE. (Chapters 4, 5). Sudhakar Sivapalan helped in creating depression label descriptors.

   (c) Farruque, N., Zaïane, O. R., Goebel, R., & Sivapalan, S. (2022, May). DeepBlues@ LT-EDI-ACL2022: Depression level detection modelling through domain specific BERT and short text Depression classifiers.

In Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion (LT-EDI@ACL) (pp. 167-171). (Chapter: 7). **Stood 3rd out of 30 teams in the competition for depression level detection through Reddit posts**.

(d) Farruque, N., Huang, C., Zaïane, O., & Goebel, R. (2023, February). Basic and Depression Specific Emotions Identification in Tweets: Multi-label Classification Experiments. In Computational Linguistics and Intelligent Text Processing: 20th International Conference, CICLing 2019, La Rochelle, France, April 7–13, 2019, Revised Selected Papers, Part II (pp. 293-306). Cham: Springer Nature Switzerland. (Chapter 7). Nawshad Farruque was responsible for designing and developing the experiments of traditional machine learning models and analysis. Chenyang Huang was responsible for deep learning based experiments design and development.

2. Under review papers

   (a) Farruque, N., Goebel, R., Sivapalan, S., & Zaïane, O. (2022). Depression Symptoms Modelling from Social Media Text: A Semi-supervised Learning Approach. arXiv preprint arXiv:2209.02765. (Chapter 5 contains edited version of the paper uploaded at arXiv). Sudhakar Sivapalan helped in creating annotation guideline and annotating samples.

   (b) Farruque, N., Goebel, R., Sivapalan, S., & Zaïane, O. R. (2022). Deep Temporal Modelling of Clinical Depression through Social Media Text. arXiv preprint arXiv:2211.07717. (Chapter 6 contains edited version of the paper uploaded at arXiv).

For conducting social media based depression detection, data collection, and annotation, three ethics approval have been received from University of Alberta, Research Ethics Office (REO):

1. Depression Detection from Social Media Language Usage (Pro00099074).

2. Depression Dataset Collection (Pro00082738).
3. Social Media Data Annotation by Human (Pro00091801).

To my parents, Farruque and Tahera.

To my wife Sumaiya, and daughter Zoeya.

To the people who are suffering from mental health disorders.

It is hard to find a needle in a hay stack, it is much harder if you haven't seen a needle before.

– Judea Pearl

Acknowledgements

First and foremost, I would like to express my gratitude towards the Almighty Allah for giving me enough mental strength to endure this tough journey. This is the most challenging yet worth-while and incredibly eye-opening pursuit I have ever taken in my life. The path of my PhD was pretty rough and in so many occasions I was overwhelmed with self-doubts; however, I had to survive till the end to appreciate the virtue of this journey. Most importantly, I learned how to become an independent researcher and a better human being through this process. During my tenure, I have crossed paths with so many beautiful souls and I am thankful for it by all means.

I would like to thank my supervisors for their endless support and trust they put on me. Thank you for teaching me how to gradually become an independent researcher and a better writer. Thank you Randy for being there always when I needed you. Thank you for reminding me of the ultimate goal I needed to achieve in my PhD program and helping me make an important collaboration with a group of multi-disciplinary researchers to achieve that goal. I am also grateful to you for providing me with a very generous funding package in the final years of my PhD program, especially it was much needed at that time to stay focused in my research. I always admire your wisdom and how you always put student's interest at the forefront. Thank you Osmar for being like a friend always during my PhD tenure. I have learned how to be always positive no matter what and see the good side of anything bad. Thank you very much for giving me access to the hardware resources, it made my life a lot easier. I would like to thank my supervisory committee members, Sudhakar Sivapalan from Psychiatry, and Lili Mou from Computing Science, for arranging time from their busy schedule to examine my thesis and providing feedback during committee meetings. Thank you Sudhakar for participating in the discussions on clinical aspects of this research, annotating samples and reviewing the papers. With you, I had an opportunity to just scratch the surface of language analysis of patients by the clinician; it's no doubt an enormously interesting area! Thanks to my candidacy committee member, Chris Westbury from Psychology and examination committee member, Geoffrey Rockwell from Philosophy for their invaluable suggestions which helped me improve the writing of my thesis.

I would like to thank my wife, Sumaiya, for being my confidante. Thank you very much for being a good listener all the time. Thank you for sharing my frustrations and helping me keep my chin-up at the times when it was really hard. I would like to thank my daughter, Zoeya for lighting up my dark days. Thanks to the Almighty for gifting me with such a bundle of joy! I would like to thank my parents; my father, Farruque Ahmed, who instilled the value of education in me; my mother, Tahera Khatun, for always being by my side and encouraging me to achieve my goals no matter what and my brother Rubaiyat, who influenced me in different ways to pursue higher studies. I thank my in-laws for their encouragement time to time.

I would like to thank my colleagues at CSA lab, Edmonton Police Service (EPS). During my final years, EPS office was a nice relief to me that helped me keep going despite I was totally mentally exhausted. I would always remember the interesting chats with Mehdi ranging from Philosophy to Astronomy and everything in between! Thank you Brian for providing the reminder that I should finish sooner than later! Thanks to my friends, Shihab Bhai, Shazan the "Jabbar Saab", "Ever-smiling" Housam, "Spirited" Chenyang and the xAI and Conversational Agents lab members for introducing me with different perspectives of life and research.

Finally, I am thankful to the University of Alberta, Alberta Machine Intelligence Institute (AMII), Natural Sciences and Engineering Research Council of Canada (NSERC) and Mathematics of Information Technology and Complex Systems (MITACS) for their grants and support. I apologize if I have missed anyone to acknowledge here, but I can assure that I haven't forgotten anyone I met. All of you have contributed to make my journey meaningful and I am thankful for that.

Contents

List of Tables

List of Figures

Acronyms

AR Absence Ratio. 120, 138

ATE Augmented Twitter Word Embedding. xvi, 34, 43, 44, 46, 49, 50, 51, 52, 54, 157

ATEA All Tweets Embedding Average. 130

BART Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. 62, 74, 192

BDI Beck Depression Inventory. 59, 82, 196

BERT Bidirectional Encoder Representations from Transformers. xiii, xiv, 21, 62, 65, 68, 86, 89, 90, 192, 193, 202

BoW Bag of Words. 20, 34, 40, 44, 45, 51, 52

CES-D Center for Epidemiologic Studies Depression Scale. 59, 82, 196

CS Clinical Scoring. 122

DLD Depression Level Detection. 19, 20, 21

DPD Depressive Post Detection. xii, xiii, xiv, xv, 4, 5, 7, 9, 11, 19, 20, 21, 33, 34, 35, 45, 52, 53, 54, 57, 65, 66, 68, 69, 70, 77, 78, 79, 81, 86, 95, 109, 146, 151, 153, 156, 157, 159, 202, 203

DRFS Depression Recurrence Frequency Score. 131

DS Depression Score. 120, 137, 138, 139, 140, 141

DSD Depression Symptoms Detection. xii, xiii, xiv, xv, xvi, 4, 5, 6, 7, 10, 11, 21, 22, 30, 31, 56, 57, 58, 65, 66, 67, 68, 69, 70, 74, 75, 76, 77, 78, 79, 81, 83, 85, 86, 88, 92, 93, 95, 99, 101, 109, 119, 130, 139, 147, 151, 152, 153, 158, 159, 202

DSE Depression Specific Word Embedding. xvi, 33, 34, 42, 43, 44, 46, 50, 51, 52, 53, 61, 69, 146, 156, 157

DSM-5 Diagnostic and Statistical Manual of Mental Disorders - 5. 7, 8, 19, 30, 56, 59, 60, 82, 105, 106, 107, 146, 149, 151, 188, 196

DTR Depressive Tweets Repository. xvi, 10, 79, 81, 82, 83, 89, 90, 99

ELMo Embeddings from Language Model. 193

ELSEA Episode Level Sentence Embedding Average. 130

ES Episodic Semantic Representation. 138, 139

GPT Generative Pre-trained Transformer. 62, 192

HAM-D Hamilton Depression Rating Scale. 59, 82

HTS All Historic Tweets Semantic Representation. 138

LIWC Linguistic Inquiry and Word Count. xvi, 20, 25, 34, 41, 44, 45, 49, 50, 51, 57, 69, 107, 111, 113

LR Logistic Regression. 44, 45

LSTM Long Short Term Memory. 142, 191, 193

LSVM Linear Support Vector Machine. 44

MADRS Montgomery–Åsberg Depression Rating Scale. 59, 60, 67, 70, 82, 195, 196

MDE Minimal Depression Expression. 139, 140, 141, 142

MVC Majority Voting Classifier. 52

MVCP Majority Voting with Clinician Preference. xiv, 83, 84, 88

NB Naïve Bayes. 44

NHEL NRC Hashtag Emotion Lexicon. 41, 44, 45

NLI Natural Language Inference. 10, 56, 57, 62, 63, 64, 67, 69, 70, 74

PCA Principal Component Analysis. xvi, 34, 49, 50

PHQ-9 Patient Health Questionnaire - 9. 59, 82, 146, 196, 201

RoBERTa A Robustly Optimized BERT Pretraining Approach. 192

RSVM Support Vector Machine with Radial Basis Function Kernel. 44, 45

SBERT Sentence Bidirectional Encoder Representations from Transformers. 44, 53, 54, 193

SDD Signs of Depression Detection. xiv, 14, 20, 22, 28, 29, 30

SE Sentence Embedding. 67, 69, 70, 74

SSL Semi-supervised Learning. xvi, 4, 7, 10, 11, 76, 78, 79, 85, 86, 87, 88, 89, 93, 94, 95, 99, 102, 147, 151, 152, 159

STEP Syntax Tree Guided Semantic Explanation. xiv, 57, 71, 72, 73, 74, 158

SVM Support Vector Machine. 45

TE General Twitter Word Embedding. xvi, 33, 34, 42, 43, 44, 49, 50, 51, 52, 53, 54, 69, 146, 157

TP Temporal Patterns. 131, 138, 139

TUD Temporal User-level Clinical Depression Detection. xiii, xv, xvii, 4, 5, 6, 7, 11, 119, 120, 122, 128, 129, 130, 139, 146, 147, 148, 149, 150, 151, 159, 161, 202, 203

UDD User-level Depression Detection. xiv, 9, 14, 25, 26, 28, 29, 30

USE Universal Sentence Embedding. xiv, 44, 46, 53, 54, 61, 62, 69, 73, 79, 122, 193

WE Word Embedding. 67, 69, 70, 74

ZSL Zero-Shot Learning. xii, xiv, xvi, 4, 5, 7, 10, 56, 57, 58, 59, 60, 62, 63, 65, 66, 67, 68, 69, 70, 72, 73, 74, 75, 77, 91, 92, 94, 99, 101, 102, 147, 158

Chapter 1

Introduction

1.1 Overview

Major Depressive Disorder (MDD), otherwise called "Clinical Depression," is one of the most common mood disorders estimated to affect around 300 million people, worldwide which is around 4.4% of the global population [70]. MDD is different from temporary sadness and, if long-lasting with moderate to severe intensity, can cripple a person from functioning properly in their personal lives [28]. Such a condition can even lead to other debilitating consequences in one's life, including self-harm, substance abuse, and suicide [139]. Unfortunately, detecting depression itself is a very challenging task. According to statistics, half of the depressed people worldwide do not seek treatment for depression because of societal stigma, ignorance, or failure to acknowledge this as a disorder that requires treatment [4], [127], [150]. Because of this, there is a considerable need for an effective, inexpensive, and real time intervention of depression for this high-risk population.

According to the research by Gowen et al. [43] and Naslund et al., [92], [93], it has been found that depressed people show increased use of social media platforms to share their daily struggles, connect with others who might have experienced the same, and seek help. Interestingly, among young adults, social media is very popular where they share their day-to-day activities, and the availability of social media services is growing exponentially year by year [100]. Moreover, it has been established by an ample early research that it is possible to detect signs of depression from social media language used in social media posts [20], [25]–[27], [117], [124],

[130], [133], [152], [155]. According to these studies, linguistic features, such as n-grams, psycholinguistic and sentiment lexicons, word and sentence embeddings extracted from social media posts can be very useful for detecting depression compared to other social media-related non-linguistic features, such as social network structure of depressed users and their posting behavior. A majority of these studies use public social media data, i.e., Twitter and Reddit mental health forums for user-level depression detection because the relative ease of accessing such datasets unlike Facebook and other social media which have strict privacy policies. Despite of this plethora of work, there are some fundamental problems which are still to be addressed and are required for creating robust models for depression detection following clinical guidelines. For example, it has been found in earlier research that traditional survey-based methods of depression screening through telephone or online questionnaires can be ineffective due to the lack of truthfulness of the patients, typically caused by some cognitive bias [48]. This bias can be further aggravated because of filling out the surveys at a later time as opposed to real-time. On the other hand, through social media posts people share their day-to-day affairs, ups and downs, emotional states and overall mental health status. Therefore, it is sensible to look for signs of clinical depression in a social media user's timeline for real-time mental health status monitoring and intervention if required.

The clinical process of depression detection involves the analysis of temporal patterns of depression symptoms for at-least a two-weeks period [29]. This temporal component is very important for clinicians so that they can monitor a patient's depression over time and take necessary steps. Previous research focused on detecting depression through a digest of social media posts in a user's timeline based on different lexical, vocabulary, or topic-based clues of depression [18], [27], [59], [68], [96], [102], [117], [155]. These studies did not put effort in extracting clinically meaningful temporal patterns over all the social media posts in a user's timeline to infer the presence of their depression. Clinical depression modelling requires underlying models for depression symptoms detection. Very few studies attempted to do so by training their models with the training data gathered primarily through lexicon keyword based crawling of random social media users [88],

[96], [152], [155]. These studies did not attempt to (1) create a large enough clinician annotated samples for depression symptoms through the active participation of practicing clinical experts, (2) build a system that can help expand this annotated set of samples so that the expanded set maintains a similar quality to the clinician annotated samples and, (3) reflect the distribution of depression symptoms found in samples from users' timeline with genuine disclosure of their depression diagnosis.

It should be noted that we only focus and improve upon the research on depression modelling in social media posts compared to doing the same in other depression-related language resources, such as depression forums [67], [69] and/or interviews [122], [123]. Depression or related forums and/or interviews usually contain stronger language-specific signals of depression because of the nature of these sources. Most people participating in depression or depression-related forums are usually depressed. These people are comfortable sharing their daily struggles with like-minded and sympathetic participants, including psychologists, to seek help. While online forums allow for a less censored expression of emotions, clinical interviews, on the other hand, limit the discussion to focus on only a few dimensions of symptoms. Furthermore, such a setting focuses only on the past two weeks of a patient's subjective experience in those few clinical dimensions of depression. To overcome these challenges, we use social media posts where users share their daily affairs over a longer period of time. They are not bound to share only their depression related struggles here. Therefore, depression signals found in these samples are subtler than the ones found in depression forums or interviews and also more challenging to detect.

The datasets we use in our research are curated by external research groups [20], [133]. These datasets contain tweet samples from the users who disclosed their depression diagnosis through a statement, often called "self-disclosure". These research groups ensure the genuineness of these self-disclosures based on human annotation or other strict curation strategies. These datasets have been used as benchmark datasets in most of the earlier work in the area of social-media user-level depression detection [18], [20], [102], [133]. Tweet samples from depression rating scale rated depressed users [27] is treated as gold-standard in this area. However, the language-usage and social media posting behavior characteristic of both self-disclosure based and depression rating scale rated depressed users are found to be similar [27], [133].

In addition to the overall goals and motivations described above, this dissertation also provides specialized motivation for each chapter. In Chapter 3, we discuss our motivation behind creating Depressive Post Detection (DPD) models when data is scarce. In Chapter 4, we discuss our motivation behind using state-of-the-art pre-trained language models to develop a Zero-Shot Learning (ZSL) framework when we have no data for training a Depression Symptoms Detection (DSD) Model. In Chapter 5, we provide the motivation behind creating robust DSD model with the help of Semi-supervised Learning (SSL) strategy. Finally, in Chapter 6, we provide the motivation behind Temporal User-level Clinical Depression Detection (TUD) modelling.

1.2 Goals and Research Questions

Based on the existing research gaps, we have the following overarching research question: **"How can we model clinical depression to detect signs of depression in samples of social media posts from a user's social media timeline ?"**. Answering this question focuses on developing a clinical depression model while tackling the biggest obstacle in developing the same, i.e., data scarcity. Moreover, clinical modelling also requires integrating clinical insights into the overall modelling process. The modelling process starts with creating a model for detecting signs of depression from social media posts. Later, with the help of this model, we develop a model for detecting signs of depression symptoms. Finally, we develop a clinical depression score extractor based on the earlier created models and use it for modelling user-level clinical depression. So we define the following goals that help address the main research question:

1. Consider ways of detecting signs of depression through social media posts such as tweets, provided there is a very small number of training samples.

    - How can we develop a Depressive Post Detection (DPD) model from

   - tweets, given there is not enough clinical expert annotated tweets for the task? We answer this question in Chapter 3.

2. Find alternative ways of detecting depression symptoms when we have very small training data.

   - Can we develop a Depression Symptoms Detection (DSD) model even in the absence of training data, i.e., with the help of Zero-Shot Learning (ZSL)? We discuss this in Chapter 4.

3. Develop a system to gather more clinically relevant depression symptoms samples, which reflect symptoms distribution of the self-disclosing depressed users.

   - Starting with an initial model trained only on clinician-annotated depression symptoms samples from a subset of an existing dataset of users who self-disclosed their depression diagnosis, how can we harvest more clinically relevant depression symptoms samples and improve upon the initial model? How do aforementioned DPD and ZSL models contribute in this process? We discuss this in Chapter 5.

4. Develop a Temporal User-level Clinical Depression Detection (TUD) model with the help of the Depression Symptoms Detection (DSD) model created on clinically relevant harvested data.

   - How can we develop a temporal depression detection model through the DSD model which conforms to the clinical criteria of depression detection? We discuss this in Chapter 6.

The four main goals stated above focus on developing a clinical depression model when there are almost no datasets available to develop the underlying models, which are essential to finally develop TUD model. For developing such a model, we take advantage of state-of-the-art pre-trained embeddings, language models, natural language inference models and, temporal deep-learning models. Then we integrate clinical insights in this modelling through utilizing well-known clinical psychiatry resources, clinician annotated samples and clinician's advice. While much earlier research attempted to detect user level depression through temporal social media posts [27], [117], [155], they hardly put effort on clinical modelling, which involves careful creation of an underlying DSD model that encodes clinical insights. Lack of innovation in tackling challenges of data scarcity is also evident in the earlier work at the phase of developing the DSD model. Finally, effort for using this DSD model to develop a user-level *clinical depression detection* model, that adheres to the criteria followed in Psychiatry, is also missing. Therefore, our work is unique compared to the earlier work in a way that, we put effort in developing TUD model, which conforms to clinical criteria of depression detection as much as possible. Further, we provide extensive analyses of different clinical features and their representations and evaluate them in different data distributions and clinical depression detection settings.

Most of the earlier work that curated datasets for depression detection from social media posts are based on Twitter because of its public nature. Unfortunately, due to the privacy issues related to users identity, most of these Twitter datasets are not shared by the researchers. Of the few research groups that do share data, most still require institutional ethics approval and/or signed Data Use Agreement (DUA). Also, it is customary to apply for ethics approval while collecting these datasets on our own or even conducting such research. Collecting one's own data is problematic because it requires advanced access to expensive APIs; without that access, it is impossible to collect data on a large scale. So in this work, we primarily use Twitter datasets, which we have access to. This includes two benchmark datasets collected by Coppersmith et al. [20] and Shen et al. [133], which are used to evaluate much research in this area [20], [53], [87], [152] and closely resembles other gold label but private datasets for detecting depression through social media [27], [117].

1.3 Contributions

The main contributions of this research are in the following areas:

1. Development of clinical sub-modules for Temporal User-level Clinical De-

pression Detection (TUD) model:

(a) First, we develop a Depressive Post Detection (DPD) model that detects signs of depression by leveraging small but rigorously clinical expert annotated training data and relevant embedding representations. We bring intriguing innovation in developing powerful embedding representations, which help us achieve very good accuracy in this task compared to other relevant baseline models.

(b) Next, we emphasize the development of a Zero-Shot Learning (ZSL) model for Depression Symptoms Detection (DSD) with the help of state-of-the-art text representation techniques and clinical resources, such as clinical descriptions of depression symptoms from Diagnostic and Statistical Manual of Mental Disorders (DSM-5), depression rating scales and a practicing clinician's advice. ZSL works reasonably well compared to a strong supervised baseline and random baselines. Another advantage of ZSL is that it does not require any supervised training, which means that we do not require any labelled samples to create a ZSL model. Furthermore, we can also use ZSL model to label tweets with candidate depression symptoms.

(c) Finally, we develop an Semi-supervised Learning (SSL) system, which helps us gather more relevant data starting from relatively small clinician-annotated data. By relevant data, we mean the data that reflects the natural distribution of depression symptoms from self-disclosing depressed users and linguistic clues of depression learned from clinician-annotated samples. Our clinician-annotated dataset is the largest of its kind. Further, our SSL system results in gathering the largest number of depression symptoms samples of its kind. Also, using this dataset, we learn a DSD model which has better accuracy than the one which is only trained on clinician annotated training data.

2. Design and development of Temporal User-level Clinical Depression Detection (TUD) model and experiments to evaluate it.

(a) We identify important clinical features from DSM-5, standard clinical practice, and early research. We then integrate all those into a model that can support temporal modelling based on state-of-the-art deep temporal modelling constructs, i.e., Bidirectional Long Short Term Memory (BiLSTM) followed by Attention.

(b) We design experiments to shed light on the contributions of the extracted features through feature attribution tests in different data distributions and clinical depression detection settings.

1.4 Dissertation Organization

The thesis manuscript is organized as follows:

- **Chapter 2:** In this chapter, we lay the foundational background of this thesis and then discuss the related work. We are mainly interested in solving the problem of depression detection through the language used in social media posts by social media users (or simply "users"). We start with a description of what depression is and the associated challenges surrounding depression detection in general. Later, we discuss why and how language can be used to identify and monitor mental health conditions, especially depression. We then discuss how social media language can provide us with ample language data for analyzing people's mental health status and thereby alleviate many challenges that are part of depression detection methods, such as real-time depression detection and intervention. Finally, we discuss the related work in the area of user-level depression detection in social media. For the sake of better organization, we categorize this discussion into two main and interconnected themes:

  1. Signs of Depression Detection (SDD) and
  2. User-level Depression Detection (UDD)

  Within those categories we discuss three fundamental areas related to any machine learning based text classification task:

1. Dataset Curation
2. Feature Extraction and
3. Modelling

We provide a summary of all the publicly available UDD Twitter datasets that we could access in our research. This summary includes the details of these datasets in the light of their size, efforts on curation, availability, and timeline to acquire. Finally, we highlight the research gaps where we can make progress in several areas of SDD and UDD.

- **Chapter 3:** In this chapter, we discuss the development of a Depressive Post Detection (DPD) model for detecting signs of depression in social media posts. Unfortunately, DPD is a very low resource task, which means it is very hard to find enough human annotated depressive post samples to train a DPD model. In such a scenario, utilizing pre-trained resources is a good option because these resources already encode useful knowledge that could be valuable. In this chapter, we discuss the efficacy of several pre-trained word embedding representations and their enhancement through the accuracy achieved in the DPD task while those are used as the feature representation for the tweets. We learn and evaluate these models in two kinds of datasets, (1) an extensively annotated depression dataset of tweets, and (2) a large but noisy dataset. Through quantitative and qualitative analyses in both datasets, we show that domain specific word embedding, i.e., word embedding learned on depression texts, is more effective in detecting signs of depression. Furthermore, we show that the domain specific embedding has a potential to enhance an existing pre-trained embedding confirming the efficacy and applicability of depression specific semantic representation for the DPD task. We also compare these word embedding representations with more advanced and state-of-the-art sentence embedding representations and found sentence embedding based DPD models are, in general, better. Finally, we show that we can construct a majority voting model based on the best embedding models and can achieve significantly better accuracy than any of those models individually.

- **Chapter 4:** In this chapter, we describe a Zero-Shot Learning (ZSL) framework for Depression Symptoms Detection (DSD) from tweets. ZSL is a machine learning paradigm allowing between-class attribute transfer at test time to predict samples from classes that were not observed during training. ZSL models have promising potential to alleviate the data scarcity problem by helping to create an initial training dataset for a supervised learning task. In this chapter, we use existing state-of-the-art pre-trained embedding representations and large language model based Natural Language Inference (Natural Language Inference (NLI)) systems to represent tweets and clinical descriptions of depression symptoms to formulate a ZSL based DSD model. This model mainly leverages semantic similarity between the tweet and the symptoms descriptions to assign labels to a tweet. We experiment with various combinations of these representation techniques, clinical descriptions as well as few relevant parameters for the ZSL modelling. We establish the fact that these models are in general better in the DSD task compared to naïve baselines and supervised models fine-tuned on very small training data. We also outline experiments on how to make an explainability friendly DSD system later in the chapter through the proposed ZSL framework.

- **Chapter 5:** In this chapter, we describe a Semi-supervised Learning (SSL) framework, where we use an initial supervised learning model that leverages state-of-the-art large mental health forum text pre-trained language model further fine-tuned on a clinician annotated DSD dataset, a ZSL-based DSD model and use them together to harvest depression symptoms related samples from a large Depressive Tweets Repository (DTR) curated by us. DTR is created from the samples of tweets in self-disclosing depressed users' Twitter timeline, which helps preserve the depression symptoms distribution of self-disclosing Twitter users' tweets samples. Next, we annotate a portion of DTR with the help of clinical experts. Our clinician annotated dataset is the largest of its kind. Later, we retrain our initial DSD model with the harvested data iteratively. Finally, we discuss the stopping criteria and limits of this SSL process. We also elaborately discuss all the underlying constructs which play a vital role in the overall SSL process. Through our SSL framework, we create a final dataset that is the largest of its kind. Furthermore, a DSD and a DPD model trained on it achieve a significantly higher accuracy than their initial versions.

- **Chapter 6:** In this chapter, we use learned models for depression symptoms as described in Chapter 5 and design a deep learning based depression detection model to detect user-level clinical depression through their temporal social media posts. This chapter provides insight into the strengths and weaknesses of the underlying depression symptoms detection model to extract clinically relevant features. These features include, depression scores and the temporal patterns based on those scores, and user posting activity patterns of an user. To evaluate the efficacy of these extracted features, we create three kinds of datasets and a test set from the two existing well-known benchmark datasets for the user-level depression detection task. Later, we provide accuracy measures through single features, baseline features and feature ablation tests; in several temporal granularity, data distributions, and clinical depression detection related settings. Based on this, we draw a complete picture on the impact of different features across our created datasets. We show that, in general only semantic representation based models perform the best. However, clinical features may enhance overall performance very slightly provided the training and testing distribution is same and there is more data in a user's timeline. Predictive capability of depression score increases significantly while used in a more sensitive settings that we also discuss in this work.

- **Chapter 7:** In this chapter, we provide the summary findings of this research in the main areas of contribution mentioned in Chapter 1, i.e., (1) creating building blocks of natural language oriented Temporal User-level Clinical Depression Detection (TUD) modelling, and (2) analyzing the TUD model through various feature analysis in several data distributions and clinical depression detection settings. Later, we describe the limitations of our research in terms of validity and reliability of the dataset curation and modelling approaches. We also discuss how we conform to the existing best practices of ethics in social media based mental health monitoring research. Finally, we discuss future directions of our research including an outline to an actual depression monitoring system which might be useful for practicing clinicians to monitor their patients in clinical white space, i.e., the time between the visits to clinician's office.

Chapter 2

Background and Related Work

In this chapter, we lay the foundational background of this research and discuss related work. We are mainly interested in solving the problem of depression detection through language used in social media posts by social media users (or simply "users"). So we start with a description of our problem, i.e. what depression is and what the primary challenges are in the area of depression detection. We later discuss why and how language can be used to identify and monitor mental health conditions, especially depression. We then discuss how social media language can provide us with ample language sample for analyzing people's mood fluctuations and, eventually, depression in real-time. The hope is to understand and alleviate the challenges in depression detection.

Although we are interested in detecting signs of depression from the posts of a social media user's Twitter timeline, this task requires underlying sub-components which look into several linguistic clues relevant to clinical depression in the user's social media posts, to finally predict depression. So in this chapter, we also discuss related work in the area of detecting signs of depression from social media posts. We categorize the plethora of earlier work into two main and interconnected themes:

1. Signs of Depression Detection (SDD) and

2. User Level Depression Detection (UDD)

We separately discuss SDD and UDD in the dimensions of three fundamental areas related to any machine learning based text classification task (Figure 2.1):

1. Dataset Curation

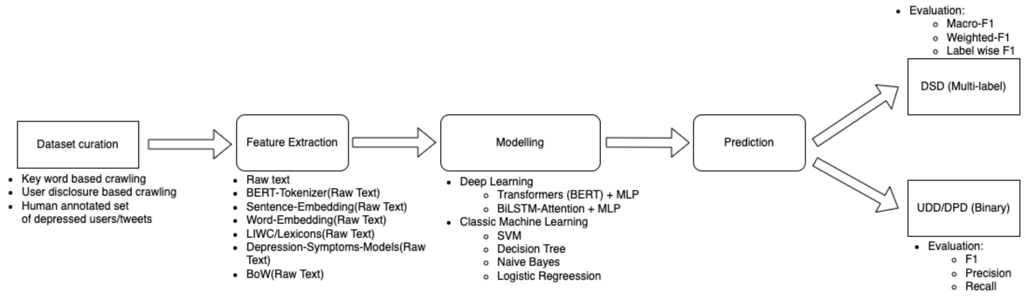


Figure 2.1: High level modelling pipeline followed in our research. In feature extraction phase, first parenthesis means a function, so BERT-Tokenizer(Raw Text) means it takes a text and tokenizes using BERT-Tokenizer. In Deep Learning model, "+" indicates followed by. MLP means Multi-layer Perceptron.

2. Feature Extraction and

3. Modelling

We provide a summary of all the available datasets that we could access in our research. This summary includes the details of these datasets in the light of their size, curation effort, availability and timeline to acquire. Finally based on all this, we highlight the research gaps where we can make progress for both SDD and UDD. We emphasize on discussing the most recent, relevant and substantial early contributions. In absence of novelty in research insights, we discuss the work, which has the most novelty.

2.1 Depression Detection and Its Challenges

"Major Depressive Disorder (MDD)" otherwise "clinical depression" or just "depression" is one of the most common mood disorders. One year and lifetime prevalence of depression are 12.9% and 10.8% respectively [65]. Prevalence for schizophrenia is 0.4% [126] and bipolar is around 4% [56]. Depression is also significantly higher in women (14.4%) and countries with a medium human development index (29.2%) [65]. Depression has the ability to disrupt one's life if left untreated. Lack of diagnosis eventually results in suicide, drug abuse, crime and many other societal problems and related costs. It has been found to be a major cause behind more than 700,000 deaths committed through suicide each year worldwide [139]. The economic burden created by depression is estimated to have been 210 billion USD in 2010 in the USA alone [44].

According to the Diagnostic and Statistical Manual of Mental Disorders - Fifth Edition (DSM-5) [34], a manual widely used by clinicians world-wide for mental disorder diagnosis, depression is defined as experiencing five or more symptoms of depression along with low mood or loss of interest (anhedonia) for atleast two weeks [29].

The challenges of depression detection can be easily understood if we carefully analyze this definition. First of all, a patient's mood pattern needs to be monitored over time to detect depression. Secondly, these mood patterns are exhibited through the symptoms of depression that needs to be identified. The basic problem is to find a real-time source for observing the depression symptoms, which should be as objective as possible, means, the patient has less cognitive bias for being untruthful to their ongoing depressed state. At the same time these observations should be of reasonable quantity so that they can help a clinician diagnose depression with greater certainty. Unfortunately, there are no tools right now that can satisfy all these constraints. Depression rating scales provided in the clinician's office may not be a true reflection of one's mental health status; a patient may not be truthful about their depressive state of mind because of the stigma about being diagnosed with a mental health disorder. In fact, it is due to stigma that leads to half of the depressed population not seeking any kind of help from the health care providers [4], [127], [150].

Moreover, depression detection and monitoring through heart rate variability found in ECG [11], brain imaging [85], [105], [151], body temperature [40] and speech [3], [123], [147] require hardware resources that needs to be worn by the patient all the time. As of now to the best of our knowledge, there is no exact mechanism developed to ground these modalities into specific depression symptoms as stated in DSM-5.

Therefore, there is a lack of effort for building an inexpensive, interpretable and effective depression detection and monitoring tool based on these modalities. Such kind of tool will be of immense help because it can aid to early detection of depression and can eventually reduce the burden associated with its full fledged form and the related cost.

2.2 Language Speaks of Mind

Previous studies suggest that the words we use in our daily life can express our mental state, mood and emotion [35], [108], [153]. There have been much research which showed that it is possible to identify language markers which are predictive of Depression [27], Anxiety [134], Post Traumatic Stress Disorder (PTSD) [21], Schizophrenia [23], Alzheimer's [38], suicidal thought patterns [6] and many other mental health conditions.

Rude et al. [124] and Resnik et al. [121] conducted psycholinguistic analysis on currently depressed, formerly depressed and never depressed college students' essay on their deepest thoughts and feelings about coming/being in college. They found there is a clear difference in the way of language usage among these groups of students. For example, it is found that depressed individuals use more negative valence words and I-pronoun than the never depressed college students, which is consistent with Beck's cognitive model of negatively valanced bias [125] and with Pyczsinski and Greenberg's self-focus model of depression [115] in Psychology.

According to Smirnova et al. [137] linguistic markers such as: rumination or tautology, rhetorics and similies, multiclause texts, use of indefinite and personal pronouns, and lexical and semantic repetitions are seen markedly higher in mildly depressed people than control based on the essay text analysis these people were provided to write.

According to Mohammed et al. [84], absolutist words, e.g., "absolutely", "everything", "nothing", "never", etc, are better predictive of depression, anxiety and suicidal thoughts based on the observation in anxiety, depression and suicidal forums text compared to control forums.

According to Alhanai et al. [3], language based signals were found to have more predictive value compared to speech, for detecting signs of depression from depressed peoples' interviews.

All these studies have established the fact that, language has a great potential to be used as an inexpensive and effective means for looking into signs of several mental health disorders including depression. However, although most of these studies focused on language usage for a single point of time and in a very controlled setting (e.g., depression forums post, interviews or written essays on a particular topic) there is still a missing and important component for depression detection which is observing temporal depressive mood patterns through the use of language in a longer temporal window and in a much less restricted setting like depression forums or clinician's office. In that particular and also very important aspect, analyzing social media language usage can be of immense help discussed in the next section.

2.3 Depression Detection via Social Media Text

Among the young adults, social media is very popular. They share their day to day activities, and the availability of social media services to do so is growing exponentially year by year [100]. A survey conducted by Pew research showed that around 90% of the teenagers use social media at least several times a day and 45% of them use social media almost constantly. According to some studies [43], [92], [93], it has been found that depressed people who are otherwise socially aloof, show increased use of social media platforms to share their daily struggles, connect with others who might have experienced the same and seek help. So social media can be thought of as a repository of language samples usually over a large temporal window, which makes it an extremely valuable resource for analyzing language markers of different mental health disorders, including depression.

There has been ample research to-date on depression detection through social media. This research is broadly divided into two distinct yet interconnected categories: Signs of Depression Detection in a social media post (SDD) and the same in user level, i.e., through the entire social media timeline of a user, or User-level Depression Detection (UDD). We see that most of these studies are based on public social media datasets, such as Twitter and Reddit. Very few of those are based on other sources which are not public and have increased privacy concerns, such as Facebook, Instagram and the like [83], [116], [128], [130]. In the next section, we will discuss earlier research on these two broad themes and we majorly focus on experiments on Twitter datasets followed by similar platforms like Facebook and Reddit.

We put more emphasis on analyzing Twitter datasets and especially text based depression analysis methods since we are interested in experimenting with accessible and thus reproducibility friendly datasets containing language markers of depression, such as Twitter. Although Reddit qualifies as well in this criteria, we do not consider conducting experiments in Reddit datasets, nevertheless, we will discuss the earlier efforts in those datasets in this chapter to shed light on text based methods of depression detection. The reason we are not interested in Reddit is mainly because Reddit datasets are based on depression forums which are not actual mainstream social media like Twitter or Facebook. Here people only discuss about their struggles and it has less natural settings than Twitter where users are free to discuss anything. We believe Twitter datasets are more representative samples of peoples daily language usage than Reddit.

2.3.1 Signs of Depression Detection in a Social Media Post (SDD)

SDD is a classification task designed for detecting signs of depression from an individual social media post. There are different categories of SDD task, however, those can be divided into three broad categories, such as

1. **Depressive Post Detection (DPD)**: This is a binary classification task with a goal to detect whether a social media post is depression indicative or not [53], [149], [154];also there are an ample earlier work for detection suicidality or suicidal tendency detection in social media [1], [17], [22], [46], [50], [54], [99], which we can think of an special case of signs of depression detection.

2. **Depression Level Detection (DLD)**: This is a multi-class classification task; the goal is to classify whether a particular social media post conveys one of the following “No Depression”, “Mild Depression”, “Moderate Depression”

or “Severe Depression” [79], [154].

3. **Depression Symptoms Detection (DSD)**: This is a multi-class multi-label (or simply multi-label) classification task, the goal here is to determine whether a particular social media post carries one of nine symptoms of depression as pointed out in DSM-5 [15], [87], [88], [152], [155]. Despite of the differences in the task descriptions, ultimately these tasks are all based on the same underlying principle, i.e detecting signs of depression from the linguistic clues contained in a social media post or text. So we will discuss in general the dataset curation process, feature extraction techniques and machine learning models used followed by limitation of the approaches in those dimensions proposed earlier in the literature in the next sections.

Dataset Curation Process

Most of the datasets in the area of DPD and Depression Level Detection (DLD) are either based on:

1. **Keyword based crawling**: In this scheme, tweets are crawled from random/self-disclosing Twitter users’ timeline based on depression/suicide/self-harm related keywords or

2. **Forum membership**: In this scheme, depression forum posts are crawled from different mental health forums, e.g., Reddit Depression or Suicide forums.

In both cases, there were some efforts to ensure the quality of the crawled social media posts either by considering only the users with genuine self disclosures and/or through annotating their posts further with the help of annotators [53], [149]. For the DSD task, Cheng et al. [15] put an effort to curate a Filipino lexicon for depression symptoms based on top-down approach (i.e., extracting symptoms related keywords from depression rating scales) and bottom-up approach, i.e., from interview from relevant cohorts. On the other hand, Mowery et al. [87], [88] put an effort to annotate random tweets largely based on single non-expert annotator

Index	Dataset Name	Social Media Type	Total Posts(D+ND)	Labels	Annotation Scheme	Date of Obtaining	Comment
1	DPD-Vioules [149]	Twitter	507(271+ 236)	Multi-class	Depression keyword based crawling of Twitter profiles and tweets and then expert human annotations for verifying users and their tweets	2017	Collected through signed agreement
2	DPD-Jamil [53]	Twitter	8753(876+ 7877)	Binary	Tweets from self-disclosing depressed users profile further annotated by non-expert single human annotator	2017	Collected through signed agreement
3	D2S	Twitter	3738	Multi-label	Tweets from self-disclosing depressed users profile further annotated by non-expert human annotators	2021	Collected through signed agreement

Table 2.1: SDD available Twitter datasets statistics.

for the same. Yadav et al. [152] although curated the largest amount of DSD annotated datasets, over-all clinician annotated portion of the above mentioned datasets is only hundred samples. Also, majority of Yadav et al.'s tweet samples are not accessible due to missing tweets and limitations of Twitter API.

Summary of Datasets for SDD

In this section, we discuss the Twitter datasets which we could access in our research, for SDD tasks. We provide the details of each dataset in the dimensions of its size, social media type, curation/annotation process, classification task type, date of obtaining and comments about data availability in Table 2.1. More details of these datasets are provided in each corresponding chapters.

Feature Extraction

For DPD and DLD tasks, to represent each social media post, non-deep learning approaches used Bag-of-Words (BoW), off the shelf (e.g., LIWC [106]) and self-curated lexicon based feature extraction, [79], [94], [149], [155]. On the other hand, deep learning approaches used variety of off-the-shelf or self-created pre-trained word/sentence embedding [36], [79]. It is found that, in general deep learning models perform best with word/sentence embedding representation and moderate size of datasets. On the other hand, non-deep learning approaches, such as Support Vector Machine (SVM) worked the best with the same compared to more classic Bag-of-Words (BoW)/Lexical representations and smaller datasets.

For DSD task, very recently a large pre-trained transformer based language model (BERT) showed a lot of promise [152] compared to statistical machine learning approaches, e.g., guided Latent Dirichlet Allocation (LDA) approach for the same [155]. Also, Ji et al. [55] developed Mental-BERT which is a BERT model fine-tuned on several mental health forums in Reddit. Their model showed state-of-the-art performance in depression detection task in Reddit.

Modelling

For DPD and DLD tasks, among non-deep learning approaches, Support Vector Machine (SVM) was found to be performing best especially with rich word/sentence embedding representations [36], [79]. However, in deep learning approaches, Convolutional Neural Network (CNN) performs pretty well [154]. In DSD task, fine-tuning pre-trained BERT based models in multi-task learning settings perform the best [152]. However, most of the performance gain for both families of machine learning models came from the rich feature representations, such as word/sentence embeddings which are pre-trained earlier on a large text corpora.

Limitations

The primary limitations of the earlier approaches are as follows:

1. **Dataset curation process**

   (a) A reasonably large clinician annotated dataset for depression symptoms is extremely rare. Only one that exists only contains 100 samples [152].

   (b) Rigorous sample annotation procedure is absent. Either few hundreds samples are annotated by a clinical expert (not necessarily a practicing clinician) or large amount of samples are annotated by a single non-expert annotators [87], [88].

   (c) The samples which were annotated by humans are most of the times coming from random tweets crawled through keyword matching, not self-disclosing depressed users' social media timelines [87], [88].

   (d) Social media datasets from depression rating scale identified users are unavailable.

2. **Feature extraction**

   (a) Lexicon based approaches require tedious process of curating these lexicons. Their scope is also limited by their lexicon components [15], [149], [155].

   (b) Although pre-trained word and sentence embedding representations have been used in these studies and showed better efficacy for over-all SDD task, there are a very few studies that explored the efficacy of domain specific embedding/language models and their application in detecting signs of clinical depression [55], [102].

   (c) Use of large language models for unsupervised modelling of DSD is missing in earlier research.

3. **Modelling**

   (a) Although the accuracy gain for SDD task ultimately comes from better text representation techniques, there were very little effort in earlier work for better representation learning [152]. Most of the earlier models did not take clear advantage of using domain specific embedding representations, which is also reflected through the fact that they required a lot of samples to learn.

   (b) To detect clinical depression, DSD is one very fundamental task. Only two early studies [152], [155] put emphasis on using DSD for clinical depression modelling. However, models in both of these works were trained on mostly non-expert annotated data.

   (c) Earlier work put much less effort on curating clinically relevant training data and thereby created a gap in learning clinically informed models, instead they focus on increasing model accuracy in their own defined easy version of the task, e.g., most of the samples of these datasets were easy for humans to annotate which was also reflected in their high annotation agreement scores [152].

2.3.2 User Level Depression Detection (UDD)

UDD is generally a binary classification task, where, based on the linguistic components present in all the posts of a user's social media timeline, it is determined whether the user is going through possible clinical depression or not. There are few studies which performed multi-class depression level detection or intensity detection from user-level data [154]. Majority of these works are based on depression forum datasets where post triaging is important. We identify two major approaches for UDD modelling from the earlier literature as follows:

1. **Non-temporal:** This modelling considers all posts of a user's timeline as a single unit of analysis hence these models are temporal information agnostic [20], [27], [53], [95], [102], [110], [120], [121], [133], [152], [154], [155].

2. **Temporal:** This modelling takes into consideration temporal patterns of linguistic components which represents user's mood, emotion and mental health over-time [14], [27], [67]–[69], [117], [130], [138]. Although temporal models make more sense in terms of clinical modelling of depression; non-temporal models over-all shed light on important linguistic clues predictive of a user's depression.

Dataset Curation Process

In both temporal and non-temporal approaches there are three main categories of dataset curation process for user-level depression detection, such as

1. **Self disclosure:** In this curation process, Twitter users who disclosed their diagnosis of depression through a social media post, e.g., "I (am/was/have been) diagnosed with depression", were considered as depressed. The majority of the user-level depression datasets are curated this way [20], [121], [133] because the relative ease of curating large number of users.

2. **Profile analysis:** Some studies put on emphasis on analyzing user's profile name containing keywords related to depression and self-harm and/or the nature of the profile picture and posts containing depression related topics [53], [152].

3. **Depression forum membership:** All the studies based on depression forums, e.g., Reddit datasets are based on depression forum memberships, i.e., a user is identified as depressed if they regularly post in those forums [67], [69], [143].

4. **Depression rating score:** Depression rating score based methods first filter out depressed users and non-depressed users based on their score above a particular threshold of depression rating scale score, then their social media posts are analyzed [27], [117]. Adoption of this type of curation process is extremely rare in the literature due to the difficulty to access social media data of the depression rating scale rated depressed people [47]. Among the two well-known studies [27], [117] which use this strategy, both identified the depressed cohort through paid crowd-sourcing, e.g., with the help of Amazon Mechanical Turk (AMT), which makes these data curation efforts questionable, i.e., it is doubtful whether the users were truthful or not. Handful of earlier studies employed weak human annotation, i.e., single annotator to find out a small subset of users with genuine depression disclosure [53], [133], [152].

For curating control users, most of these datasets curated random Twitter users [18], [20], Twitter users who never posted anything with the character string "depress" [133], Reddit users who are not members of mental health forums [67], [69] and Facebook users who have normal scores in depression rating scales [128]. Most of these datasets tried to maintain the same timeline for extracted control and depressed users. Coppersmith et al. [20] tried to gather an age and gender matched control through an off-the-shelf classifier for detecting age and gender through social media language.

Summary of Datasets for UDD

In this section, we discuss about the Twitter datasets which we could access for UDD tasks. We provide the details of each dataset in the dimensions of their size, social media type, curation/annotation process, temporality: ("Y" means temporal data, i.e., posts are time-stamped, "N" means posts are not time-stamped), whether the dataset contains tweets before (B) or after (A) the disclosure or both (BA), classification task type (Binary/Multi-class/Multi-label), date of obtaining and comments about data availability in Table 2.2.

Feature Extraction

In both non-temporal and temporal modelling approaches, the widely used features include variety of self-curated depression lexicons, e.g., metaphor depression lexicons and off-the-shelf lexicons e.g., LIWC [27], [94], [128], [130], [155] , bag-of-words [27], [53], [67], [69], [128], variations of topic modelling [120], [121], [133], [144], emotion or sentiment features [27], [117], character n-grams, I-pronoun [27], [53], [18], [69], word and sentence embedding based approaches [95], [102], [110], [143], [152], [154], and very recently depression symptom based representation [96]. It has been found that embedding feature representations perform superior than all other hand curated or classic Lexicon/BoW feature representations. It has been also found that in general depression specific vocabulary is more important for these kinds of classification task and depression vocabulary based features perform better than emotion and sentiment features by large margin [27], [133].

Modelling

Non-temporal modelling approaches paired with classic Lexicon/BoW features, SVM and Random Forest performed superior [27], [53]. In deep-learning based efforts, CNN models with optimized word embedding [102] or learned embedding [154] and sparse dictionary learning based approach [133] on multi-modal social media features performed better than classic machine learning approaches.

In temporal modelling approaches, De Choudhury et al. [27] used temporal features such as mean momentum, entropy and mean frequency of posting of a user.

Index	Dataset Name	Social Media Type	Total Users(D+ND)	Total Posts(D+ND)	Avg. Posts	Longitud.	Labels	Annotation Scheme	Before or After Self Disclosure?	Date of Obtaining	Comment
1	UOttawa [53]	Twitter	80(58 + 22)	102072(76721 + 25351)	1322(1145) + 1152(1260)	Y	Binary	Self-disclosing users further verified by human annotators for genuine disclosure	BA	2018	Collected through signed agreement
2	CLPsych-2015 [20]	Twitter	1349(477+ 872)	3110414(1132270+ 1978144)	2373(1032) + 2268(1057)	Y	Binary	Self-disclosing users further verified by human annotators for genuine disclosure	BA	2018	Collected through signed agreement and ethics review
3	IJCAI-2017 [133]	Twitter	8302(2626+ 5676)	4755198(508806+ 4246392)	193(280) + 790(888)	Y	Binary	Self-disclosing users	B	2018	Public

Table 2.2: UDD available Twitter dataset statistics.

Reece et al. [117] used different emotion/sentiment lexicon scores on a user's timeline to find out the correlation between the increase of sad emotions score around the time of self-disclosure of depressed user and they found a positive correlation. They learned a two state Hidden Markov Model (HMM) to detect the differential change between depressed and control groups. Very recently, Zogan et al. [158] proposed Hierarchical Attentional Network (HAN) to extract word level and tweet level features over all the tweets of a user's social media timeline combined with their social media, emotion, depression lexicon and topic features. Later, they used those features for their multi-layer perceptron neural network to predict users depression and achieved state-of-the-art results.

Limitations

Here we also provide the limitations in terms of dataset curation process, feature extraction and modelling as follows:

1. **Dataset curation process**

   (a) It is not possible to ensure whether a user is going through depression genuinely with the data curation strategies followed by majority of the earlier research. These approaches are at best proxy for the every day language samples of possible depression patients.

   (b) Twitter datasets with depression disclosure date and time is almost nonexistent except the datasets curated by Shen et al. [133]. However, temporal information is available in most of the Reddit datasets [67], [69].

2. **Feature extraction**

   (a) Feature extraction approaches proposed in earlier research are not very clinically relevant. Most of these studies do not put emphasis on extracting depression symptoms related features from the social media users timeline and further apply them for clinical depression modelling. Thus these models have generally little value in terms of clinical depression detection.

3. **Modelling**

    (a) Non-temporal modelling proposed in earlier research did not take into account the temporal features, e.g., change of depression level of a user over time at all. These models are based on digest of social media posts and mainly focused on detecting important excerpts or group of words for detecting user-level depression.

    (b) Most of the temporal models are not clinically useful. These models only extracted summary statistics of temporal patterns through a user's timeline, e.g., mean frequency and fluctuation of their posting [27], although these summary statistics may have some value for social media based depression detection, they do not provide insights on the tendency of depression states to linger or recur which are two very important measures related to depression detection and monitoring [61].

    (c) Time series modelling with clinical features is absent in these works. Also, a very few studies are present which attempt to perform deep temporal modelling of clinical depression with the help of state-of-the-art deep learning constructs, such as, Bidirectional Long Short Term Memory (BiLSTM) and Attention, which are found to be performing very well for time series analysis [135]. Few studies, such as [158] created models which are based on too granular units which makes the model easily highly parameterized with risks of over-fitting. Moreover, their end-to-end model does not extract clinical features.

2.4 Research Gaps

Here we discuss the research gaps for dataset curation, feature extraction and modelling efforts from earlier research. We discuss them for both SDD and UDD.

2.4.1 Dataset Curation

The following are the main research gaps in the area of SDD.

1. No framework has been implemented to annotate a large dataset of social media posts carrying clinical signs of depression as vetted by a practicing clinician.

2. No effort was put to annotate samples for depression symptoms from self-disclosing depressed Twitter users with a disclosure statement.

Research gaps in the area of UDD are as follows:

1. A proper framework for collecting social media post samples from genuine depressed users in small time with reasonable quantity is missing.

2. Detection of clinical depression and analysis of depressive mood patterns for a set of users who have possible ongoing depression (as diagnosed by the clinician) in Twitter datasets haven't been explored yet.

3. No effort have been put to automatically identify genuine depressed users based on their profile and social media information. This information which differentiates between a genuine depressed user compared to a non-genuine user can be characterized through analyzing their profile and social media posts collected at 1.

2.4.2 Feature Extraction

The following are the main research gaps in the area of SDD.

1. Although pre-trained word and sentence embedding extractions have been used in these studies and showed better efficacy for over-all SDD task, there are very few that explore the efficacy of domain specific embedding and their application in detecting signs of clinical depression.

2. Although there is an advent of large language models; no previous work explored the efficacy of existing large language models for signs of clinical depression detection through a unsupervised settings, i.e., through designing a Zero-Shot modelling approach.

Research gaps in the area of UDD are as follows:

1. Non-temporal models are inherently not clinical, because they do not take into account the depressive episodes and their occurrence to detect depression.

2. Clinical features, for example depression scores/levels were not extracted in earlier research which results in models which are not clinically interpretable. Presence of depression inertia and recurrence, two important predictors of depression were not analyzed in these studies.

3. Relevant clinically useful analysis is also missing, for example, different thresholds and criteria for clinical depression modelling based on DSM-5 haven't been explored yet.

2.4.3 Modelling

Research gaps of SDD is already discussed earlier in the feature extraction section. We believe in small training data settings, SDD can largely benefit from rich text representation based on word/sentence embeddings. Therefore, the discussion above on research gaps focusing on feature extraction level is sufficient to address the overall research gaps in SDD. Here we discuss the research gaps in UDD as follows:

1. Use of state-of-the-art temporal modelling for example BiLSTM followed by Attention architecture hasn't been rigorously experimented in various feature settings, clinical settings and data distributions.

2.5 Conclusion

Plethora of earlier studies have successfully laid the foundation of social media language based depression detection and analysis. However, all these studies are focused on detecting depression mostly from the presence of depression related keywords and their frequency or semantic representations from a user's timeline. There is a lack of effort in actual clinical depression modelling which requires Depression Symptoms Detection (DSD) from the same. DSD is a supervised learning task which requires a lot of clinician annotated samples. Despite of the advent of large language models there are very few clinical frameworks for DSD modelling, that leveraged these models and clinical insights together. Finally, there haven't been any effort to further employ DSD for user-level clinical depression modelling through relevant clinical feature integration and analyze their contribution for the same. Therefore, there is a lot of opportunity in developing robust clinical depression model that could help clinicians properly analyze a user's social media timeline for their mood patterns predictive of depression.

Chapter 3

Signs of Depression Detection in Text: Experiments with Semantic Representations of Language

In this chapter, we discuss the development of a Depressive Post Detection (DPD) model for detecting signs of depression in social media posts. Unfortunately, DPD is a very low resource task; that means, it is very hard to find enough human annotated depressive post samples to a train DPD model. In such a scenario, utilizing pre-trained resources is a good option because these resources already encode useful knowledge that could be valuable. Here we discuss the efficacy of several pre-trained word embedding representations and their enhancement through the accuracy achieved in DPD task while those are used as the feature representation for the tweets. We learn and evaluate these models in an extensively annotated depression dataset of tweets as well as on another large but noisy dataset for the same. Through quantitative and qualitative analyses in both datasets, we show that domain specific word embedding, i.e., word embedding learned on depressive texts is more effective in detecting signs of depression. Furthermore, we show that the depression specific embedding has a good potential to enhance an existing pre-trained embedding confirming the efficacy and applicability of depression specific semantic representation for the DPD task. We also compare these word embedding representations with more advanced and state-of-the-art sentence embedding representations and found sentence embedding based DPD models are in general better. Finally, we show that, we can construct a majority voting model based on the best embedding models and can achieve significantly better accuracy than any of these models individually.

3.1 Motivation

As stated in Chapter 1 and 2, for detecting clinical depression in user-level, we need to first identify signs of depression from social media posts. Depression detection from social media posts can be specified as a low resource supervised classification task because of the paucity of valid data. Although there is no concrete precise definition of valid data, previous research emphasizes collecting social media posts, which are either validated by annotators as carrying clues of depression, or coming from the people who are clinically diagnosed as depressed, or both. Also from Chapter 2, we know deep learning models in general achieve state-of-art performance in detecting signs of depression from social media posts. However, most of these models require a lot of labelled samples when they do not leverage any pre-trained embedding compared to the ones which leverage pre-trained embeddings [53], [102].

Therefore the motivation of our research comes from the need for a better feature representation specific to depressive language, and reduced dependency on a large set of (human annotated) labelled data for detecting signs of depression in tweets.

3.2 Methodology

In this work, we mainly use existing pre-trained word embedding models. Our goal is to leverage the concepts embedded in the word embedding space and use those to inform DPD models in the face of data scarcity. Here, we start by creating an embedding representation based entirely on depression forums text, which we call Depression Specific Word Embedding (DSE). Later, we use that to make an existing pre-trained General Twitter Word Embedding (TE) more robust through learning a non linear mapping function from TE to DSE, we call it Augmented Twitter Word Embedding (ATE). Finally, we discuss the efficacy of the proposed representation based on the accuracy achieved in the DPD task while using them as feature representation for tweets through various machine learning models compared to DSE, state-of-the-art Sentence Embedding (SE) representations, LIWC [106], Sentiment Lexicon, simple Bag of Words (BoW) based baselines and a relevant early work. We provide an elaborate analysis on different dataset characteristics we use for DSE creation and for our final evaluation. We provide both the quantitative and qualitative analyses of the proposed DPD models. For quantitative analysis, we report the average accuracy of the DPD models trained on a dataset which is rigorously annotated for signs of depression and repeated random sub-samples of held-out sets created from it. We also shed light on different text preprocessing scenarios and their effect on our various embedding representation based models. Finally, to test the generalizibility of the models we report their accuracy in a more challenging setting i.e in a highly imbalanced and noisy Twitter dataset for depression. For qualitative analysis, we provide the PCA projections of positive and negative emotion carrying words from a well-known psycholinguistic lexicon named LIWC in the vector representation space of ATE and TE, to shed light on the effect of our semantic augmentation method.

3.3 Datasets

Here we provide the details of our two datasets, which we use for our experiments and their annotation procedure, the corpus they are curated from and their quality comparisons. We also provide the details of the depression forum dataset curation which we use to create DSE.

3.3.1 DPD-Vioules Dataset

DPD-Vioules dataset is curated by the ADVanced ANalytics for data SciencE (ADVANSE) research team at the University of Montpellier, France [149]. This dataset contains tweets having key-phrases generated from the American Psychiatric Association (APA)'s list of risk factors and the American Association of Suicidology (AAS)'s list of warning signs related to suicide. Furthermore, they randomly investigated the authors of these tweets to identify 60 distressed users who frequently write about depression, suicide and self mutilation. They also randomly collected 60 control users. Finally, they curated a balanced and human annotated dataset of a total of around 500 tweets, of which 50% tweets are from distressed and 50% are from control users, with the help of seven annotators and one professional psychologist. The goal of their annotation was to provide a distress score (0 - 3) for each tweet. They reported a Cohen's kappa agreement score of 69.1% for their annotation task. Finally, they merged tweets showing distress level 0, 1 as control tweets and 2, 3 as distress tweets. Distress tweets carry signs of suicidal ideation, self-harm and depression while control tweets are about daily life occurrences, such as weekend plans, trips and common distress such as exams, deadlines, etc.

3.3.2 DPD-Jamil Dataset

DPD-Jamil dataset is collected by a research group at the University of Ottawa [53]. They first filtered depressive tweets from #BellLetsTalk2015 (a Twitter campaign) based on keywords such as suffer, attempt, suicide, battle, struggle and first person pronouns. Using topic modeling, they removed tweets under the topics of public campaign, mental health awareness, and raising money. They further removed tweets which contain mostly URLs and are very short. Finally, from these tweets they identified 30 depressed users who self-disclosed their own depression, and 30 control users who did not. They employed two annotators to label tweets from 10 users as either depression or control. They found that their annotators labelled most tweets as control. To reduce the number of control tweets, they further removed neutral tweets from their dataset, as they believe neutral tweets surely do not carry any signs of depression. After that, they annotated tweets from the remaining 50 users with the help of two non-clinician annotators with a Cohen's kappa agreement score [140] of 67%. Finally, they labelled a tweet as depressive if any one of their two annotators agree, to gather more depressive tweets. This left them with 8,753 tweets with 706 depressive tweets.

3.3.3 Quality of Datasets

Here we present a comparative analysis of our datasets based on their curation process and the linguistic components present in them relevant to depressive language detection as follows:

Analysis Based on Data Curation Process

DPD-Jamil is different from DPD-Vioules in the following ways: (1) this dataset is collected from the pool of tweets which is a part of a mental health campaign; (2) the words they used for searching depressive tweets are not validated by any depression and/or suicide lexicons; (3) although they used two annotators (none of them are domain experts) to label the tweets, they finally considered a tweet as carrying signs of depression if at least one annotator labelled it as so, hence introduced more noise in the data; (4) it is not confirmed how they identified neutral tweets since their neutral tweets may convey depression as well; (5) they identified a person is depressed if s/he disclose their depression, but they did not mention how they determined these disclosures. Simple regular expression based methods to identify these self disclosures can introduce a lot of noise in the data. In addition, these self disclosures may not be true.

Analysis Based on Linguistic Components Present in the Dataset

For this analysis, we use Linguistic Inquiry and Word Count (LIWC) [141]. LIWC is a tool widely used in psycholinguistic analysis of language. It extracts the percentage of words in a text, across 93 pre-defined categories, e.g., affect, social process, cognitive processes, etc. To analyse the quality of our datasets, we provide scores of few dimensions of LIWC lexicon relevant for depressive language detection [27], [61], [95], such as 1st person pronouns, anger, sadness, negative emotions, etc (Table 3.1) for the depressive tweets present both in our datasets. The bold items in that table shows significant score differences in those dimensions for both datasets and endorses the fact that DPD-Vioules indeed carries more linguistic clues of depression than DPD-Jamil (the higher the score, the more is the percentage of words from that dimension is present in the text). Moreover, depres-

LIWC Category	**Example Words**	**DPD-Vioules Depressive Tweets Score**	**DPD-Jamil Depressive Tweets Score**
1st person pronouns	I, me, mine	12.74	7.06
Negations	no, not, never	3.94	2.63
Positive Emotion	love, nice, sweet	2.79	2.65
Negative Emotion	hurt, ugly, nasty	8.59	6.99
Anxiety	worried, fearful	0.72	1.05
Anger	hate, kill, annoyed	2.86	2.51
Sadness	crying, grief, sad	3.29	1.97
Past Focus	ago, did, talked	2.65	3
Death	suicide, die, overdosed	1.43	0.44
Swear	fuck, damn, shit	1.97	1.39

Table 3.1: Score of DPD-Vioules and DPD-Jamil in few LIWC dimensions relevant to depressive language detection.

sive tweets in DPD-Jamil are mostly about common distress of everyday life unlike those of DPD-Vioules, which are indicative of severe depression. Figure 3.1 depict the word clouds created from DPD-Vioules and DPD-Jamil depressive tweets respectively. We provide few random samples of tweets from DPD-Vioules and DPD-Jamil in Table 3.2 as well.

We use DPD-Vioules as our train dataset because of its consistent presence of linguistic clues of depression. We use DPD-Jamil dataset as a representative dataset for many earlier studies for signs of depression detection task [19], [20], [53] which leverage similar data to report their experiments. Moreover, high prevalence of control tweets compared to depressive tweets is a very common phenomena in this task, and this is very much evident in this particular dataset. Therefore, we use this as our test set.

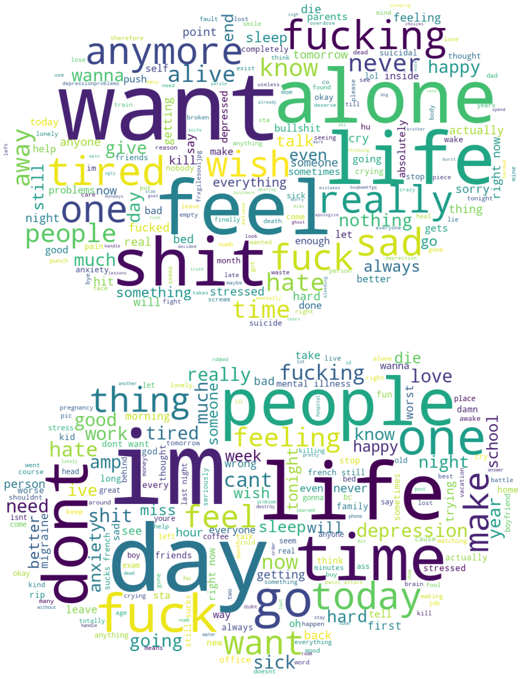


Figure 3.1: DPD-Vioules (top) vs DPD-Jamil (bottom) depressive tweets word clouds.

Datasets	Depressive Tweets
DPD-Vioules	“I wish I could be normal and be happy and feel things like other people”
	“I feel alone even when I’m not”
	“Yesterday was difficult...and so is today and tomorrow and the days after...”
DPD-Jamil	“Last night was not a good night for sleep... so tired And I have a gig tonight... yawnnn”
	“So tired of my @NetflixCA app not working, I hate Android 5”
	“I have been so bad at reading Twitter lately, I don’t know how people keep up, maybe today I’ll do better”

Table 3.2: Sample random tweets from DPD-Vioules and DPD-Jamil.

3.3.4 Creating a Depression Specific Corpus

To create a depression specific word embedding, we curate our own depression corpus. For this, we collect all the posts from the Reddit depression forum: r/depression[1] between 2006 to 2017 and all those from Suicidal Forum[2] and concatenated for a total of 856,897 posts. We choose these forums because people who post anonymously in these forums usually suffer from severe depression and share their struggle with depression and its impact in their personal lives [26]. We believe these forums contain useful semantic components indicative of depressive language.

3.4 Data Preprocessing

We perform the following preprocessing steps for all our Twitter datasets, we use NLTK[3] for tokenizing our tweets and also Ekphrasis[4] for normalizing tweets.

1. Lowercase each words.

2. Remove words starting with @ and “rt”.

3. Remove one character words (except “a”, “i” and “u” (further replaced by you)) and digits.

4. Remove tweets which are less than three words long.

5. Re-contract contracted words in a tweet. For example, “I’ve” is made “I have”.

6. Elongated words are converted to their original form. For example, “Looong” is turned to “Long”.

7. Remove tweets with self-disclosure, i.e., any tweet containing the word “diagnosed” or “diagnosis” is removed.

[1] reddit.com/r/depression/

[2] suicideforum.com/

[3] https://www.nltk.org/book/ch06.html

[4] https://github.com/cbaziotis/ekphrasis

8. Remove all punctuation except period, comma, question mark and exclamation.
9. Remove URLs.
10. Remove non-ascii characters from words.
11. Remove hashtags
12. Remove emojis.

For tweets preprocessing, stop words and few essential punctuations are retained as stated earlier. We find pre-trained word and sentence embedding based models work generally better with stop words and punctuation through our experiments.

However, to curate depression specific corpus to train depression specific embedding, we remove numbers, non-words (e.g., words that start with anything other than letters), stop words, remove any tags specific to Reddit and retain hyphenated words, to preserve important depression related phrases. We make this corpus as clean as possible so that the word embedding algorithm can learn the depression vocabulary specific semantic representation better.

3.5 Feature Representation Methods

3.5.1 Bag-of-Words (BoW)

We represent each tweet as a vector of vocabulary terms and their frequency counts in that tweet, also known as BoW. The vocabulary terms refer to the most frequent 400 terms existing in the training set. Before creating the vocabulary and the vector representation of the tweets, we perform the tweets preprocessing as stated earlier. We also tried Tf-IDf based BoW, however, we find simple frequency based BoW works better in our case.

3.5.2 Lexicons

We experiment with several emotion and sentiment lexicons, such as LabMT [32], VADER [41], Emolex [80], AFINN [97], LIWC [141], NRC-Hashtag-Sentiment-Lexicon (NHSL) [57], NRC Hashtag Emotion Lexicon (NHEL) [81] and CBET [131]. Among these lexicons we find LIWC and NHEL perform the best and hence we report the results of these two lexicons. The following subsections provide a brief description of LIWC, NHEL and lexicon-based representation of tweets.

Linguistic Inquiry and Word Count (LIWC)

LIWC [107] has been widely used as a good baseline for depressive tweet detection in earlier research [18], [95]. We use it to convert a tweet into a fixed length vector representation of 93 dimensions, that is then used as the input for our machine learning models. Each of these dimensions signify the percentage proportion of words related to that dimension out of all the words in a particular text blurb.

NRC Hashtag Emotion Lexicon (NHEL)

In NHEL there are 16,862 unigrams, each of which is associated with a vector of eight scores for eight emotions, such as anger, anticipation, disgust, fear, joy, sadness, surprise and trust. Each of the real valued score indicates how much a particular unigram is associated with each of the eight emotions. In our experiments, after preprocessing, we use the lexicon to determine a score for each token in the tweet; finally, we sum them to get a vector of eight values for each tweet, which represents the expressed emotions in that tweet and their magnitude. Finally, we use that value as a feature for our machine learning models.

3.5.3 Embeddings

We use a number of pre-trained word and sentence embeddings methods to represent the tweets. We provide the technical settings of these methods as follows (also in Appendix A.5):

General Twitter Word Embedding (TE)

We use a pre-trained 400 dimensional skip-gram word embedding learned from 400 million tweets with vocabulary size of $3,039,345$ words [42] as a representative of word embedding learned from a general dataset (in our case, tweets); we believe this captures the most relevant vocabulary for our task. The creator of this word embedding used negative sampling ($k = 5$) with a context window size = 1 and mincount = 5[5]. Since it is pre-trained, we do not have control over the hyperparameters it uses and simply use it as is.

Depression Specific Word Embedding (DSE)

We create a 400 dimensional DSE on our curated depression corpus. First, we identify sentence boundaries in our corpora based on punctuation, such as: "?","!" and ".". We then feed each sentence into a skip-gram based word2vec implementation in gensim[6]. We use negative sampling ($k = 5$) with the context window size = 5 and mincount = 10 for the training of these word embeddings. DSE has a vocabulary size of $29,930$ words. We choose skip-gram for this training because skip-gram learns good embedding from a small corpus [78].

Augmented Twitter Word Embedding (ATE): a non-linear mapping between TE and DSE

In this step, we create a non-linear mapping between TE and DSE. To do this, we use a Multilayer Perceptron Regressor (MLP-Regressor) with a single hidden layer with 400 hidden units and Rectified Linear Unit (ReLU) activations (from hidden to output layer), which attempts to minimize the Minimum Squared Error (MSE) loss function, $\mathcal{F}(\theta)$ in Equation 3.1, using stochastic gradient descent:

$$\mathcal{F}(\theta) = \arg\min_{\theta} \mathcal{L}(\theta) \tag{3.1}$$

where

$$\mathcal{L}(\theta) = \frac{1}{m} \sum_{i=1}^{m} ||g_i(x) - y_i||_2^2 \tag{3.2}$$

[5] radimrehurek.com/gensim/models/word2vec.html

[6] radimrehurek.com/gensim/models/word2vec.html

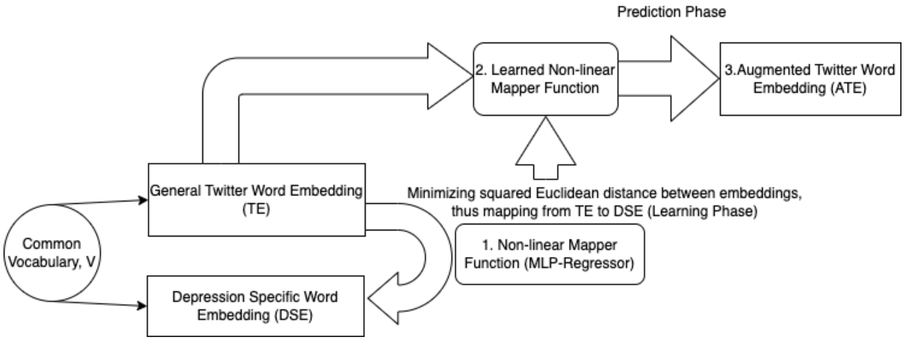

Prediction Phase
2. Learned Non-linear Mapper Function
3.Augmented Twitter Word Embedding (ATE)
General Twitter Word Embedding (TE)
Minimizing squared Euclidean distance between embeddings, thus mapping from TE to DSE (Learning Phase)
Common Vocabulary, V
1. Non-linear Mapper Function (MLP-Regressor)
Depression Specific Word Embedding (DSE)

Figure 3.2: Non-linear mapping of TE to DSE (creation of ATE).

Word Embeddings	**Corpus Type**	**#Posts**	**Vocab. Size**
TE, ATE	Twitter	400M	3M
DSE	Depression Forums	1.5M	30K

Table 3.3: Corpus and vocabulary statistics for word embeddings.

and

$$g(x) = ReLU(b_1 + W_1(b_2 + W_2x)) \tag{3.3}$$

here, $g(x)$ is the non-linear mapping function between the vector x (from TE) and y (from DSE) of a word $w \in V$, where, V is a common vocabulary between TE and DSE; W_1 and W_2 are the hidden-to-output and input-to-hidden layer weight matrices respectively, b_1 is the output layer bias vector and b_2 is the hidden layer bias vector (all these weights and biases are indicated as θ in Equation 3.1). In Equation 3.2, m is the length of V (in our case it is 28,977). Once the MLPR learns the θ that minimizes $\mathcal{F}(\theta)$, it is used to predict the vectors for the words in TE which are not present in DSE (i.e., out of vocabulary (OOV) words for DSE). After this step, we finally get an Augmented Twitter Word Embedding (ATE) which encodes the semantic representation of depression forums as well as word coverage from tweets. A summary of the vocabulary sizes and the corpus our embedding sets are trained on is provided in Table 3.3.

Conditions for Embedding Augmentation/Mapping

Our non-linear mapping between two embeddings works better given that those two embeddings are created from the same word embedding creation algorithm (in our case skip-gram) and have same number of dimensions (i.e., 400). We also find that a non-linear mapping between our TE and DSE produces slightly better ATE than a linear mapping for our task, although the former is a bit slower.

Sentence Embeddings

We use two state-of-the art sentence embedding models, such as, Universal Sentence Encoder (USE) [13] and Sentence Bidirectional Encoder Representations from Transformers (SBERT) [118] to represent the tweets. More technical details of these embeddings are provided in Appendix A.5. The rationales behind choosing these embedding representations are their (1) availability, i.e., they are widely available (2) moderate sized embedding dimension, which does not cause memory problems during experimentation, (3) superior performance in many NLP downstream task, and (4) being good representative of all existing sentence embedding models.

3.5.4 Embedding Representation of Tweets

For word embedding based tweet representation, we take the average of the vector of the individual words in a tweet, ignoring the ones that are Out Of Vocabulary (OOV). For sentence embedding, we simply feed the tweet to the respective sentence embedding model that generates a fixed length sentence vector for that tweet.

3.6 Experimental Setup

We experiment with all the 40 combinations from eight feature representation methods, such as BoW, NHEL, LIWC, TE, DSE, ATE, USE, SBERT and five standard machine learning models, such as Naïve Bayes (NB), Logistic Regression (LR), Linear Support Vector Machine (LSVM), Support Vector Machine with Radial Basis Kernel (RSVM) and Decision Tree (DT)[7]. To show the effects of different preprocessing strategies, we report these results for four kinds of preprocessing strategies, such as (1) **all**: means tweets with stop words and punctuation, (2) **all-punct**:

[7] `https://scikit-learn.org/stable/supervised_learning.html#supervised-learning`

means tweets without punctuation but with stop-words, (3) **all-stop**: means tweets without stop-words but with punctuation and (4) **all-punct-stop**: means tweets without punctuation and stop-words. Along with BoW, NHEL and LIWC, we use few random baselines, such as (1) **All-Majority**: always predicting the majority label and (2) **Random-Uniform**: predicting either depression or control tweets based on random uniform distribution. We do not experiment with deep learning models, because they require a lot of labelled samples in general. Our focus here is to evaluate the efficacy of pre-trained word embedding representations and their enhancements rather the efficacy of the classifier themselves.

Machine Learning Model Settings

For the Support Vector Machines (SVMs) and LR, we tune the hyperparameter, $C \in \{2^{-9}, 2^{-7}, \ldots, 2^{5}\}$ and $C \in \{10^{-9}, 10^{-7}, \ldots, 10^{5}\}$ respectively and additionally, $\gamma \in \{2^{-11}, 2^{-9}, \ldots, 2^{2}\}$ for the RSVM (see scikit-learn Support Vector Machine (SVM)[8] and LR[9] documentations for further description of these hyperparameters). We use min-max feature scaling for all the features.

3.7 Evaluation

3.7.1 Quantitative Performance Analysis

For quantitative performance analysis, we use average results (i.e., average Precision, Recall and F1) for the best performing combinations among all the 40 combinations of our standard machine learning and feature representation methods and preprocessing strategies (as described in Section 3.6).

For a single experiment, we split all our DPD-Vioules data into a disjoint set of training (80% of all the data) and testing (20% of all the data) or the testset-1 (Table 3.4).

We use stratified sampling so that the original distribution of labels is retained in our splits. Furthermore, with the help of 10-fold cross validation in our training set,

[8] `http://scikit-learn.org/stable/modules/svm.html`

[9] `https://scikit-learn.org/stable/modules/linear_model.html#logistic-regression`

Datasets	Train(D)	Test(D)
DPD-Vioules	405(203)	101(50)
DPD-Jamil	-	8753(876)

Table 3.4: Number of tweets in the train and test splits for the two datasets. The number of depressive tweets is in parenthesis.

we learn the best hyperparameter settings for all our model-feature representation combinations, except for those that require no such hyperparameter tuning. We then find the performance of the best model on our test.

We have run 30 such experiments on 30 random train-test splits. Finally, we report the performance of our best model-feature representation combinations based on the Precision, Recall, and F1 score averaged over the test sets (testset-1's) of those 30 experiments and for all preprocessing strategies (Table 3.5).

To establish further generazibility of our pre-trained embedding based ML models, we report their performance in a separate test set created from DPD-Jamil (testset-2). For this, we only report the top models under each embedding representation category, model and preprocessing strategy, found at previous experiment in testset-1. We report a baseline which is trained on DPD-Jamil dataset and random baselines as described earlier in testset-2 to shed a light that in general how challenging it is to score a high accuracy in this set (Table 3.6). We also report the performance of a majority voting model consisting of all our most effective models, i.e., ATE, DSE and USE which are learned on DPD-Vioules dataset and report its performance in testset-2. Through this model, we only predict a tweet as the one carrying signs of depression if majority of the models voted yes for that, otherwise we assume it as a control tweet.

Finally, we report comparison among the models in terms of accuracy for all model-feature-preprocessing combinations to shed light on the impact of different preprocessing strategies (Section 3.8).

3.7.2 Qualitative Performance Analysis

According to earlier research, depression has close connection with abnormal regulation of positive and negative emotion [61] and [130]. So to consider how the

Category	**Model-Feat.**	**Precision**	**Recall**	**F1**
Baselines	NB-NHEL	0.6335(±0.0195)	**0.9285**(±**0.0274**)	0.7529(±0.0173)
	NB-BoW	0.6648(±0.0316)	0.8673(±0.0370)	0.7521(±0.0277)
	LR-LIWC	0.7178(±0.042)	0.7660(±0.0633)	0.7397(±0.0418)
	Majority	0.5392(±0)	1(±0)	0.7006(±0)
	Random-Uniform	0.5323(±0.0363)	0.4897(±0.0588)	0.5093(±0.0422)
	Vioules et al. [149]	0.71	0.71	0.71
Word Embedding based Models	RSVM-TE (all)	0.7793(±0.0481)	0.8169(±0.0638)	0.7957(±0.0385)
	LR-TE (all-punct)	0.7532(±0.0341)	0.8152(±0.0503)	0.7814(±0.0252)
	RSVM-TE (all-stop)	0.7583(±0.0319)	0.8527(±0.0551)	0.8013(±0.0259)
	LR-TE (all-punct-stop)	0.7450(±0.0346)	0.8776(±0.0378)	0.8079(±0.0239)
	LSVM-ATE (all)	0.7783(±0.0352)	0.8442(±0.0396)	0.8089(±0.0237)
	LSVM-ATE (all-punct)	0.7704(±0.0389)	0.8260(±0.0407)	0.7964(±0.0294)
	LR-ATE (all-stop)	0.7627(±0.0381)	0.8697(±0.0341)	0.8118(±0.0253)
	LR-ATE (all-punct-stop)	0.7617(±0.0416)	0.8606(±0.0355)	0.8071(±0.0263)
	LR-DSE (all)	0.7705(±0.0393)	0.8976(±0.0380)	0.8283(±0.0270)
	LR-DSE (all-punct)	0.7703(±0.0397)	0.8969(±0.0400)	0.8280(±0.0285)
	LR-DSE (all-stop)	0.7705(±0.0393)	0.8976(±0.0380)	0.8283(±0.0270)
	LR-DSE (all-punct-stop)	0.7703(±0.0397)	0.8969(±0.0400)	0.8280(±0.0285)
Sentence Embedding based Models	LR-USE (all)	**0.8353**(±**0.0330**)	0.8782(±0.0555)	**0.8549**(±**0.0297**)
	LR-USE (all-punct)	**0.8353**(±**0.0330**)	0.8782(±0.0555)	**0.8549**(±**0.0297**)
	LR-USE (all-stop)	0.7738(±0.0315)	0.8727(±0.0416)	0.8195(±0.0249)
	LR-USE (all-punct-stop)	0.7741(±0.0312)	0.8721(±0.0404)	0.8194(±0.0240)
	LSVM-SBERT (all)	0.8346(±0.0389)	0.8685(±0.0396)	0.8502(±0.0271)
	RSVM-SBERT (all-punct)	0.8293(±0.0418)	0.8612(±0.0452)	0.8438(±0.0309)
	LR-SBERT (all-stop)	0.7744(±0.0325)	0.8297(±0.0575)	0.8001(±0.0343)
	LSVM-SBERT (all-punct-stop)	0.7711(±0.0388)	0.8230(±0.0552)	0.7949(±0.0348)

Table 3.5: Average accuracy on testset-1 best model-feat combinations on various preprocessing strategies.

Category	**Model-Feat.**	**Precision**	**Recall**	**F1**
Baselines	Majority	0.0807	1	0.1493
	Random-Uniform	0.0819	0.5099	0.1412
	Jamil et al. [53]	0.1706	**0.5939**	0.265
Word Embedding based Models	RSVM-TE (all)	0.3102	0.3041	0.3071
	LR-TE (all-punct-stop)	0.201	0.3987	0.2673
	LSVM-ATE (all)	**0.3546**	0.3451	0.3498
	LR-ATE (all-stop)	0.2559	0.4144	0.3164
	LR-DSE (all)	0.2921	0.4512	0.3546
Sentence Embedding based Models	LR-USE (all)	0.3432	0.4243	0.3795
Majority Voting Model	MVC (all)	0.3868	0.3988	**0.3928**

Table 3.6: Average accuracy on testset-2 for contender WE models and best SE model found in testset-1.

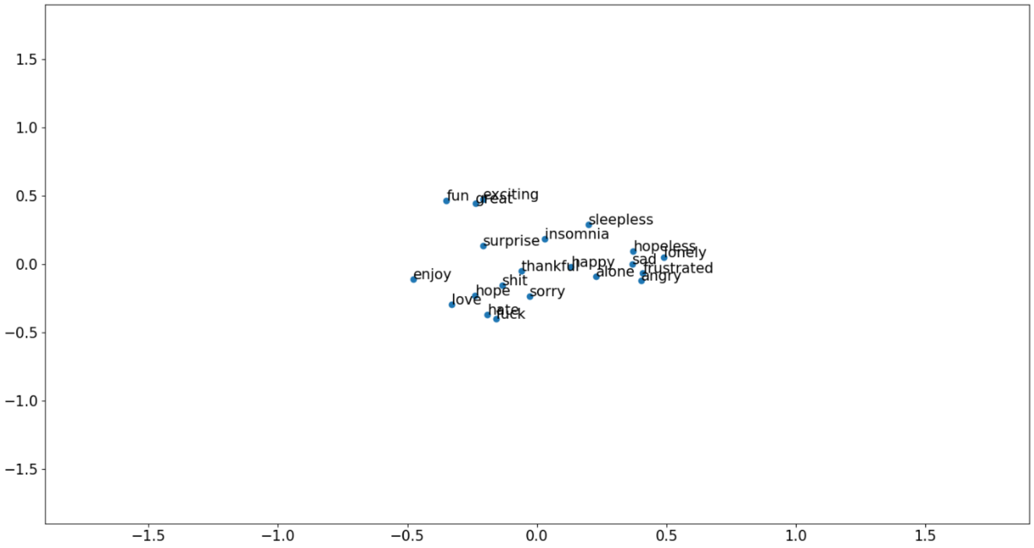


Figure 3.3: Two-dimensional PCA projection of LIWC POSEMO and NEGEMO words (frequently occured in our datasets) in TE.

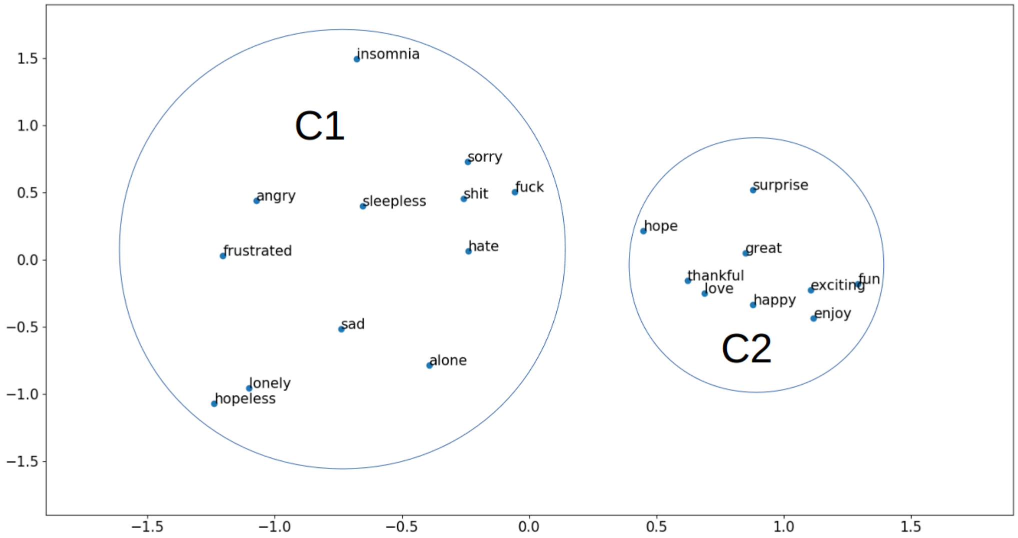


Figure 3.4: Two-dimensional PCA projection of LIWC POSEMO and NEGEMO words (frequently occured in our datasets) in ATE.

words carrying positive and negative sentiment are situated in our augmented vector space, we plot the PCA projections of ATE and TE for the high frequency words used in both datasets that are the members of LIWC positive emotion (POSEMO) and negative emotion (NEGEMO) categories.

3.8 Results Analysis

In quantitative analysis, we observe the following based on F1-scores, unless stated otherwise, significant difference means p-value < 0.05 in a two-tailed paired t-test:

1. Best DSE is better than the best ATE in testset-1 by 1.65% (p-value < 0.01) and in testset-2 it is better by only 0.48%.

2. Since DSE is learned on stop words and punctuation removed sentences, it is not susceptible of preprocessing because it only focuses on important words for depression not the stop words and punctuation.

3. ATE/TE is slightly more sensitive to punctuation than stop-words, i.e., without punctuation their performance degrades, however, they perform slightly better without stop-words, means stop-words may have very slight bad effect on them.

4. In testset-1, difference between ATE and TE is not statistically significant. However in testset-2, ATE is 4% better than TE.

5. Best preprocessing strategies in testset-1 for ATE (i.e "all-stop") and TE (i.e., "all-punct-stop") results in degraded performance in testset-2 compared to "all" preprocessing strategy (Table 3.6).

6. Sentence embedding models perform significantly worse when stop words are removed in testset-1.

7. Most of the model-feature combinations perform slightly better in "all" preprocessing strategy compared to other preprocessing strategies in both testsets.

8. Best sentence embedding based models achieve significantly better accuracy than best word embedding based models in both testsets.

9. F1 scores in testset-2 is lower because of its highly imbalanced nature and not very prominent depressive language components as further confirmed by LIWC analysis (Tables 3.1, 3.2 and Figure 3.1).

10. DSE is much smaller than ATE/TE in vocabulary size, yet has better performance in both testsets.

11. Majority of the models under embedding representation family is significantly better than lexicon and BoW models in testset-1.

12. A Majority Voting Classifier (MVC) performs the best compared to individual classifiers for each embedding representation in testset-2.

In qualitative analysis, we observe that NEGEMO and POSEMO words form two clearly distinctive clusters, i.e., C1 and C2 respectively in ATE. We also notice the word "insomnia" and "sleepless" which represent common sleep problem in depressed people, reside in C1 or NEGEMO cluster. However, we do not see any such clusters in TE (Figure 3.4 and 3.3). We believe this distinctions of affective contents in vector space partially play a role in our overall accuracy. Also the PCA projection gives a glimpse of the semantic relationship of affective words in depressive language. Although its not an exhaustive analysis but a insightful one that we believe would be helpful for further analysis of affect in depressive language.

The above observations lead to the following insights:

1. Overall all the embedding based models are sensitive to preprocessing stopwords and punctuation. Although in testset-1 it is tough to reach to a conclusion on that.

2. Sentence embedding based models are significantly more sensitive to stop words than punctuation. This indicates, stop words contribute in creating contextual representation better (testset-1).

3. Vocabulary wise DSE is much smaller than ATE/TE, yet its powerful performance in DPD task indicates depression specific vocabulary plays important role in depression detection.

4. Since ATE and TE have exactly the same vocabulary, the accuracy increase for ATE compared to TE in both test sets indicates the efficacy of our proposed augmentation method.

5. Sentence embedding representation are more powerful and represent context much better than avg. pooling based word embedding representation of tweets, reflected through their significantly better accuracy gain in both testsets.

6. Semantic augmentation/mapping helps organize the semantic representation space built on samples of source domain, such as tweets in this case (Figure 3.4).

7. Even with small dataset, it is possible to learn a robust DPD model with the help of embedding representations, which is exhibited through the huge performance gap between BoW/Lexicon family of models compared with embedding family of models. This observation is further corroborated by the Majority Voting Classifier (MVC) model's superior performance in testset-2.

3.9 Limitations

We find the following limitations in our work:

1. Due to the limitations of data, we had to conduct our experiments in small datasets.

2. Embedding augmentation does not work well across two different embedding representations based on embedding creation algorithms. That means, source and target embedding should be from same family of embedding creation methods. We verified it for skip-gram to skip-gram mapping, however, we

find glove to skip-gram embedding mapping does not provide any accuracy gain.

3. Semantic augmentation works better in an area where vocabulary plays important role, such as DPD task.

4. Due to small dataset size pairwise student t-test may not be very reliable. However, our testset-2 is big and results in that set provides more reliable insights on the models generalizibility.

5. We test the idea of semantic augmentation in depression detection from language task. Depressive language has well defined semantic representation as backed-up by early literature and found out through our experiments as well. It is still to be properly analyzed it's generalizibility in other domain where this semantic representation is not well defined.

6. Due to the computational expense, we do not experiment with SBERT in testset-2. We think its performance will be close to USE because it was so in the testset-1.

7. How different preprocessing strategies affect the overall accuracy for various embedding based methods needs further analysis.

8. We do not extensively experiment with various other augmentation techniques, including variational auto-encoder based augmentation and also, how sentence embedding could be enhanced for a particular domain based on this process. We will investigate that in future.

3.10 Conclusion

In this chapter, we provide experimental analysis for various embedding based text/tweet representations for the DPD task. We also outline the creation of DSE and its use in further enhancing the predictive capability of a TE in DPD task. We show that, DSE performs the best among the Word Embedding (WE) based models. Despite DSE is much smaller than TE in terms of its vocabulary size, this accuracy gain indicates smaller but relevant vocabulary is very important for DPD task. Also, performance of ATE slightly better than TE corroborates the usefulness of learned semantic representation from depression forum posts for this task. In general, Sentence Embedding (SE) based models are superior than Word Embedding (WE) models, where USE is slightly better than SBERT. We show that, in a majority voting settings, we can achieve the best result in the testset-2, indicating the promise for using a majority voting model for candidate depressive tweet filtering.

Chapter 4

Depression Symptoms Modelling from Text: A Zero-Shot Learning Approach

In this chapter, we describe a Zero-Shot Learning (ZSL) framework for Depression Symptoms Detection (DSD) from tweets. ZSL is a machine learning paradigm allowing between-class attribute transfer at test time, in order to observe and predict the classes of samples from classes that were not observed during training. ZSL models have good potential to alleviate data scarcity problem by helping create initial training dataset for a supervised learning task. In this chapter, we use existing state-of-the-art pre-trained embedding representations and large language model based Natural Language Inference (NLI) systems to represent tweets and clinical descriptions of depression symptoms to formulate a ZSL based depression symptoms detection model. This model mainly leverages semantic similarity between the tweet and the descriptions of the depression symptoms to assign a label to a tweet. We experiment with various combinations of these representation techniques, clinical descriptions as well as few relevant parameters for the ZSL modelling and establish the fact that these models are in general better in DSD task compared to naive baselines and supervised models fine-tuned on a very small training data. We also outline experiments on how to make a explainability friendly DSD system later in the chapter through the proposed ZSL framework.

4.1 Motivation

According to DSM-5, to clinically diagnose depression, a clinician looks at the temporal patterns for depression symptoms for a patient in typically two-weeks time window. As already stated in Chapter 1, in our social media user-level depression modelling framework, we would like to reflect this clinical process of depression diagnosis, where, detecting depression symptoms is a core component. In general there are very few attempts taken for DSD task such as [15], [87], [152]. Here we provide further details of the limitations that were briefly described in Chapter 2, as follows:

1. All these studies were formulated based on pure supervised learning mind set, means, more emphasis were put on bulk social media posts, such as tweets collection based on depression symptoms related keywords, later annotating them directly for depression symptoms and that way curate a dataset to train a DSD model. This process can easily become a burden to the annotators if the majority of the tweets are non-relevant to depression symptoms. Thus affecting badly the total annotation process and results in inferior quality of depression symptoms data.

2. There have been no attempts taken yet in terms of using state-of-the-art word or sentence embeddings, NLI models, and existing clinical resources and insights to make a system that can further filter depression symptoms candidate samples.

3. Most of the curated DSD datasets used in these studies are not accessible. For the only dataset we have access to, we are provided with only the Twitter IDs. Using these IDs, we are able to crawl less than 50% of the tweets because of the limitation of Twitter API and missing tweets. However, we have a huge collection of candidate DSD samples which are yet to be annotated, which means, there is a good potential use of a ZSL approach to further filter them for annotation and thereby curate a good quality of dataset.

Therefore, we formulate a ZSL modelling framework, with which we can create weakly labelled DSD samples and thereby alleviate data scarcity problem to an extent.

4.2 Methodology

We first describe how the DSD task is formulated as a problem of finding semantic similarity between a label and a tweet. To start with, we describe the label description curation process through clinical knowledge of depression symptoms. Later, we describe the ZSL framework, which leverages several pre-trained word and sentence embeddings and NLI models to find out the membership of a tweet with each label through their descriptors. We then evaluate the accuracy of the ZSL models based on well known multi-label classification evaluation measures, such as Macro-F1 and Weighted-F1 scores (Appendix A.1) and compare them with strong supervised and random baselines. We also report the efficacy of the best ZSL model through its performance in DPD task, compared to LIWC and random baselines, to shed light on its ability to identify signs of depression from tweets. Further as an ongoing work and to evaluate the ZSL framework for its efficacy in developing an explainable DSD system, we propose a text explainability algorithm called STEP. STEP may encourage multiple short and hierarchical phrases inside a tweet to explain its labels. In companion with the previous point, we propose an Explainability Index (EI) score which is used to grade the explainability mechanism for various ZSL models proposed in this chapter.

4.3 Datasets

We use mainly two datasets here, such as (1) **DSD-Clinician-Tweets-Original**: This is a subset of IJCAI-2017 dataset. There are 539 tweet samples in this dataset and it is annotated by four annotators for possible depression symptoms, including two clinical experts. Details of the dataset and its curation process is described in Chapters 2 and 5, and (2) **DPD-Vioules Dataset**: This dataset is rigorously annotated for signs of depression and is used for evaluating the performance of DPD models in the last chapter. Details of this dataset and its curation is provided in

Chapter 3.

4.4 Alternative ZSL Approaches

In this section, we elaborately describe different sub-components of ZSL modelling. The basic idea is to find relevance between the tweet and the depression symptom labels. However, before diving deep into the modelling framework, we need to be familiar with few preliminary concepts which is described in the next subsection.

4.4.1 ZSL Model Preliminaries

Given, a tweet, T, it has a label, L_i where, $L_i \in \{L_1, ...L_m\}$, if it has a strong *membership-score* with any of its descriptors, l_j where, $l_j \in L_i$ and L_i = $\{l_1, ...l_n\}$. Here the descriptor l_j is a representation of the label L_i. For example, consider that one of our depression symptoms L_i is "Low Mood" and the descriptors representing L_i is a set, l = $\{Despondency, Gloom, Despair\}$. If T has a strong membership-score with any members of l, we can say T has the label L_i = "Low Mood." Since our DSD task is a multi-label classification task, it is possible for T to have multiple labels at the same time, because it can have a strong membership-score with respect to the descriptor(s) of multiple labels. This is a ZSL paradigm, because we determine the label(s) of T based on its membership-score with respect to any of the descriptors in l at test time, with which our models are probably not familiar with at training time.

We use mainly two broad families of ZSL models in this work, such as (1) Embedding family, which means, both word and sentence embedding models and (2) Natural Language Inference (NLI) models. For embedding models, we represent T and each l_j using various classic and state-of-the-art word and sentence embedding models and measure their membership-scores based on how close they are in the vector space through the cosine distance. For NLI models, we extract the probability type of entailment-scores which shows the membership-score for a T with respect to each l_j.

Figure 4.1 depicts the high level description of the proposed ZSL framework.

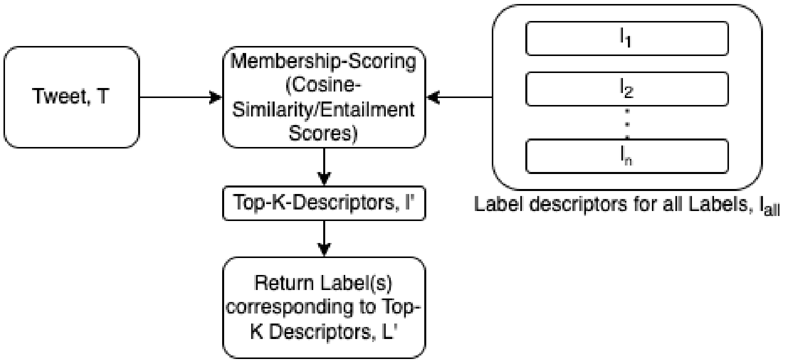

Tweet, T
Membership-Scoring (Cosine-Similarity/Entailment Scores)
Top-K-Descriptors, l'
Return Label(s) corresponding to Top-K Descriptors, L'
l1
l2
ln
Label descriptors for all Labels, lall

Figure 4.1: An overview of ZSL framework.

To further elaborate the framework, we start by discussing depression symptoms label L and their descriptors (l) curation process (Section 4.4.2), later, we describe representation creation of the tweets and labels for embedding families of models (Section 4.4.3), finally we discuss membership-score calculation between tweets and labels (Section 4.4.5). We describe each of these processes, using the notation described above. Please note, in the following sections and through-out the paper, we use the term "Embedding" to define a set of word vectors.

4.4.2 Labels (L) Curation for ZSL

First, we use well known depression rating scales and align the common depression symptoms concepts with the help of a clinician, to better understand the general language used for describing individual clinical symptoms. We then separate the minimal description of the symptoms or **Header** for DSM-5 (**DSM-Header (DH)**) and Montgomery–Åsberg Depression Rating Scale (MADRS) [49] (**MADRS-Header (MH)**) and slightly elaborated description of the symptoms **Lead** for MADRS (**MADRS-Lead (ML)**). We curate a list of elaborated descriptions of depression symptoms concept with the help of all the rating scales (Appendix A.3), such as Patient Health Questionnaire - 9 (PHQ-9), MADRS, Beck Depression Inventory (BDI), Hamilton Depression Rating Scale (HAM-D) and Center for Epidemiologic Studies Depression Scale (CES-D) [76] and discussion with the clinician, which we call **All**. In addition, we combine only the headers of DSM-5 and MADRS for corresponding depression symptoms, which we call **MADRS+DSM-Header**

Sample of depression Symptoms	DSM-Headers (DH)	MADRS-Headers (MH)	MADRS-Leads (ML)
Disturbed sleep	Insomnia, Hypersomnia	Reduced sleep	Reduced duration of sleep, Reduced depth of sleep
Anhedonia	Loss of interest, Loss of pleasure	Inability to feel pleasure	Reduced interest in surroundings, Reduced ability to react with adequate emotion

Table 4.1: A glimpse of few depression symptom labels (L) and some Headers and Leads that constitute l_{all}.

(MH+DH). Finally, we use a hand curated and expert annotated depression symptoms lexicon, named **SSToT** [155]. It is to be noted that, DSM-5 is the manual used for clinical depression detection world-wide, and the basis for most of the other developed depression rating scales. Furthermore, MADRS is a clinician rating scale instead of a typical patient rating scale, meaning, a clinician provides their judgment to rate depression of a patient instead of a patient provides their own rating as in patient rating scales. Since we use annotation advice from the clinician to annotate our data, and we want to analyze the tweets for depression symptoms from the clinician's point of view; MADRS perfectly fits our need. MADRS headers and leads are more easily understandable by the annotators compared to other rating scales. See Table 4.1 for a sample for headers and leads for couple of depression symptoms. All these label descriptors are provided in the Appendix A.6.

4.4.3 Representation of Tweets and Labels for ZSL

Here we separately discuss about various embedding based representation techniques for the tweet, T and depression symptoms label descriptor, l_j.

Word Embedding Family (WE)

We use several classic word embedding models, including Google News (Google)[1], Twitter Glove (Glove)[2], General Twitter Skip-gram Word Embedding (TE) [42]

[1] https://code.google.com/archive/p/word2vec/

[2] https://nlp.stanford.edu/projects/glove/

(Chapter: 3), Depression Specific Word Embedding (DSE) trained on depression specific corpora (Chapter 3) , Depression Specific Embedding Augmented Twitter Word Embedding (ATE) (Chapter 3), NLI pre-trained Roberta Embedding (Roberta-NLI) [66] and Universal Sentence Encoder (USE) embedding [13]. All these embeddings except DSE have been trained on millions of tokens. As USE and Roberta-NLI are sentence embeddings, we take the sentence vector for each word as their word vector. In the following sections, we describe how we leveraged them to create sentence representations.

Average Word Vector Models (WV-AVG)

If we assume a tweet, T or a label descriptor, l_j (Section 4.4.1) as a our sentence, and each sentence, S consists of n words, i.e., $S = \{W_1, ...W_n\}$, *"wv"* is a function that returns the vector representation of a word, then a sentence as an averaged word vector can be expressed as follows:

$$\frac{\sum_{i=0}^{n} wv(W_i)}{n} \tag{4.1}$$

Word Vector Mapper Models (WV-MAPPER)

As originally proposed in Chapter 3, we learn a least square projection matrix, M_w, between the word vectors of the common vocabulary V of both source and target embeddings. This learned matrix is then used to adjust word vectors of source embedding. We call this new adjusted source embedding, an augmented embedding and it is used to create WV-AVG sentence representation as outlined in Equation 4.1. This method has been previously found to be effective for the depressive post detection task [36]. For our WV-MAPPER models, our source embedding is one of the general Twitter pre-trained embeddings, e.g., Glove, Google, USE or Roberta-NLI and the target is DSE. The specification for this mapping is as follows, where M_w^* is the learned projection matrix and V_S and V_T are the common vocabularies of source and target embeddings.

$$M_w^* = \arg\min ||wv(V_S)M_w - wv(V_T)||_2^2 \tag{4.2}$$

Sentence Embedding Family (SE)

We use state-of-the-art Roberta-NLI and USE sentence embeddings which are transformer based models and multi-task pre-trained on NLI and Semantic Textual Similarity Tasks (STS) (i.e., Roberta-NLI) and sentiment analysis tasks as well (i.e USE) (Appendix A.5)

Vanilla Sentence Vector Models (SV)

Provided a sentence, S, its sentence vector is represented as $sv(S)$.

Sentence-to-Word Vector Mapper Models (SV-WV-MAPPER)

We use the same formulation as stated in Equation 4.2, however, while learning the projection matrix M_s^*, here we use the sentence vector of a source word and learn its projection to word vector of the target word for the common vocabulary between the source and target embeddings. All the other notations are the same as noted earlier:

$$M_s^* = \arg\min ||sv(V_S)M_s - wv(V_T)||_2^2 \tag{4.3}$$

Later, we use M_s^* to transform $sv(S)$. The intuition behind this mapping is that a sentence vector may not be good at representing a label descriptor l_j, which are usually short-phrases. So we project a sentence vector for a tweet, T and its corresponding label descriptor, l_j to a word-vector space (in our case "DSE", where we have salient semantic clusters of depression symptoms).

4.4.4 Natural-Language-Inference (NLI) Model

We use the BART-Large-MNLI (or simply Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension (BART)) model[3], which uses BERT and Generative Pre-trained Transformer (GPT) hybrid pre-training, further fine-tuned on NLI task and performs very well in ZSL settings (Appendix A.5). In the NLI task, a model is given a premise and a hypothesis and it needs to predict whether the hypothesis entails the premise or contradicts it. It has

[3] https://huggingface.co/facebook/bart-large-mnli

been found to be a very effective ZSL model, which can pretty accurately predict whether a given label (in our case label descriptor, l_j) entails a particular sample, in our case a tweet, (T). This mechanism provides a probability score $\in [0, 1]$ of entailment for each label.

4.4.5 ZSL Top-k-Label-Membership Formulation

At the heart of our Top-k-Label-Membership formulation is an algorithm that determines the membership of a tweet, T with all the descriptors, l_{all} for all labels, L. We later sort the descriptors based on their membership-scores with T in descending order (assuming higher score means better membership), and get $l_{all-sorted}$ (Algorithm 1). Finally, we return the labels, $L' \subset L$ represented by the top-k descriptors, $l' \subset l_{all-sorted}$ as our candidate labels for the tweet, T (Algorithm 2). In the following sections we describe our membership-scoring details for Embedding and NLI family of models.

Algorithm 1: Sorted-Descriptors

```
Input: $T$, $l_{all}$, $mode$
Output: $l_{all-sorted}$
$l_{all-sorted} \leftarrow \emptyset$ ;
membership-score-dictionary $\leftarrow \emptyset$ ;
if $mode$ is “Embeddings” then
    foreach $l \in l_{all}$ do
        membership-score-dictionary$[l] \leftarrow$ 1 - cosine-distance$(T, l)$ ;
    end
end
else if $mode$ is “NLI” then
    foreach $l \in l_{all}$ do
        membership-score-dictionary$[l] \leftarrow$ entailment-prob-score$(T, l)$ ;
    end
end
$l_{all-sorted} \leftarrow$ descriptors(sort-desc(membership-score-dictionary)) ;
return $l_{all-sorted}$ ;
```

Embedding Family Models

For this family of models, we use cosine similarity or (1 - cosine-distance) to determine the membership of a vector representation of tweet, T to the same of any of

Algorithm 2: Label-Predictor

```
Input: $L$, $l_{all-sorted}$, $k$
Output: $L'$
$L' \leftarrow \emptyset$ ;
$n \leftarrow 0$ ;
while $n < k$ do
    foreach $l' \in l_{all-sorted}$ do
        foreach $L_i \in L$ do
            if $l' \in L_i$ then
                $L' \leftarrow L' \cup L_i$
            end
        end
        $n \leftarrow n + 1$
    end
end
return $L'$ ;
```

the descriptors in l_{all}. **Centroid Membership:** In this scheme, we represent each label, L_i with the average representation vectors of all of its descriptors, which we call "centroid". For example, in the centroid-based method, T has label L_i if T has a strong membership-score with $centroid(L_i)$, where $L_i = l = \{l_1, l_2, ..., l_n\}$ and l is the set of descriptors. Then we return $L' \subset L$, i.e., the top-k labels, based on the descending order of the cosine similarity with T as candidate labels for T. **Top-k Centroid Membership:** Similar to centroid membership, instead of considering all the descriptors of L_i, we use the top-k descriptors based on the cosine similarity. In Figure 4.2, we provide an overview of centroid methods.

NLI Family Models

As mentioned in Section 4.4.4, NLI models provide probability scores for entailment for a tweet, T to its descriptor, l_j. We follow a similar procedure as for the embedding family models except we use the entailment probability scores to find the final candidate labels, L' for T.

4.5 Experimental Setup

We design our experiments to enable analysis with respect to model accuracy and explainability. We report two experiments to confirm the accuracy of our models,

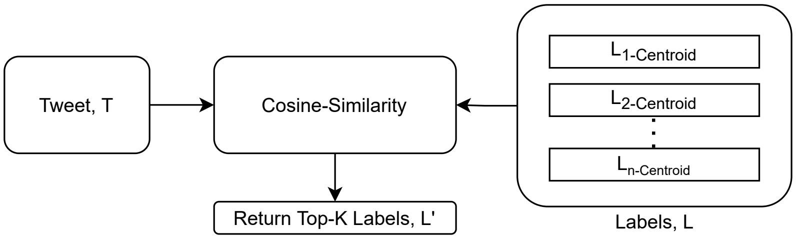


Figure 4.2: An overview of ZSL-Centroid methods.

such as in (1) **DSD** task, which is our core task and (2) **DPD** task, which confirms the predictive capability of our models in general to identify depressive vs. control tweets. In terms of explainability, we formulate an explanation index (EI) score and analyze how different models perform in terms of it.

4.5.1 DSD task

We perform experiments on all the combinations of our ZSL family models and depression label descriptors curation strategies described earlier. In addition, we run these experiments for various configurations of top-k $= \{1, 3, 6, 9\}$ label descriptors.

However, to analyze and discuss the results, we report the best models under each of the ZSL families. We run this experiments in DSD-Clinician-Original-Tweets dataset (Chapter 5). In Tables 4.2 and 4.4, we report these results, where each model is named as: [ZSL -Model-Name(Label-Descriptor-Name)]-[Top-k].

In Table 4.2, we report the performance of the top-3 best models under each family of ZSL models and a set of random baselines. In Table 4.3, we report the models which have similar performance to the best model for each family. Further, we report the results in a subset of DSD-Clinician-Tweets-Original, called DSD-Clinician-Tweets-Original-Test, to compare the best models performance with a strong supervised baseline (Table 4.4). For this baseline, we use Mental-BERT [55] fine-tuned on depression symptoms tweets (Chapter 5).

The three naïve baseline models with which we compare ZSL models are, **All-True:** here we always predict all the labels, **Random-Uniform:** here we predict labels based on random uniform distribution and **All-Majority:** here we always predict from the top-3 majority labels. For all these experiments, we consider “Psychomotor Agitation” and ”Retardation” as two separate symptoms instead of one. This results in framing DSD as a multi-label classification task with total of 10 labels.

4.5.2 DPD Task

For this task, we use membership-scores for various symptoms (i.e their corresponding label-descriptors membership score with the tweet) as the feature representation. We then train an SVM classifier with this representation and create our ZSL based DPD model. We use our top performing model, i.e., BART-All-top-6, for this feature representation. We then compare this model’s performance with All-Majority (always predicting the label which has most number of samples) and Random-Uniform class baselines. To compare and contrast the performance of our ZSL based DPD model with other DPD models described in the last chapter, we use DPD-Vioules dataset here as well. The reason to use an SVM classifier for our ZSL based DPD model is that, it is found to be the best performer given the small size of DPD-Vioules dataset with only ≈ 500 samples of depression and control tweets (Table 4.5).

4.6 Evaluation

Since our DSD task is a multi-label classification task, we report Macro-F1 and Weighted-F1 scores (Appendix A) to evaluate our ZSL models for all the labels and also for the labels with majority samples (because our dataset is highly imbalanced) respectively. For this evaluation we create 10 non-overlapping partition of samples from our DSD-Clinician-Tweets-Original dataset and provide the above accuracy score averaged over those 10 partitions. To compare with a strong supervised baseline based on Mental-BERT, we report the performance of our top contenders in a subset of DSD-Clinician-Tweets-Original dataset called, DSD-Clinician-Tweets-Original-Test. For the binary depressive posts detection task, we use F1-score and we use the same experimental setup as described in Chapter 3. We measure the

ZSL-Family	Model-Name	Macro-F1	Weighted-F1
WE	DSE-(MH+DH)-top-3	**0.2363(±0.0712)**	0.2823(±0.0489)
	Roberta-NLI-MH-top-3	0.2348(±0.0519)	**0.3712(±0.0687)**
	Roberta-NLI-All-top-3	0.2199(±0.0416)	0.2886(±0.0403)
SE	USE-SE-All-top-1	**0.3001(±0.0646)**	0.3592(±0.0497)
	USE-SE-All-top-3	0.2958(±0.0342)	**0.4384(±0.0495)**
	USE-SE-SSToT-top-3	0.2895(±0.0870)	0.3644(±0.0753)
NLI	BART-All-top-6	**0.3613(±0.0624)**	**0.5150(±0.0522)**
	BART-All-top-3	0.3575(±0.0673)	0.5001(±0.0575)
	BART-All-top-3	0.3206(±0.01106)	0.4921(±0.0463)
Random	All-True	**0.1648(±0.001)**	**0.3517(±0.0047)**
	Random-Uniform	0.1422(±0.0060)	0.2876(±0.0147)
	All-Majority	0.1125(±0.0016)	0.3203(±0.0077)

Table 4.2: DSD task macro and weighted F1 scores for the top-3 best models under each ZSL families in DSD-Clinician-Tweets-Original dataset, winner in each categories are bolded.

statistical significance among our various models based on two-tailed paired t-test.

4.7 Results Analysis

We observe the following from our experiments:

1. Based on top-3 models under each family reported in Table 4.2, we find:

    (a) Macro-F1 and Weighted-F1 scores of best Word Embedding (WE) based model are significantly lower than best Sentence Embedding (SE) and NLI based models in all clinician annotated tweets. In the small subset of this dataset, where we compare with the baseline, we also observe this trend.

    (b) SE and NLI based models perform best majorly with “All” descriptors.

    (c) Top-3 label descriptors have good predictability in all the ZSL models.

    (d) MADRS and DSM header based descriptors win slightly by “All” in WE family. However, for other ZSL families we see, “All” always provides

[4] This is based on WV-MAPPER discussed earlier

[5] This is based on SV-WV-MAPPER discussed earlier

ZSL-Family	Model-Name	Label-Descriptors-(Top-k)
WE	DSE	DH-(3,6), MH+DH-(1,6)
	Google	All-(3,6,9)
	Roberta-NLI	All-(1,6), MH-(6), ML-(3), MH+DH-(3,6,9)
	USE-DSE-Mapped[4]	All-(3,6,9), ML-(6,9), MH+DH-(3,6,9)
	USE	All-(9), DH-(9), MH-(6), MH+DH-(6,9)
SE	USE-SE	DH-(1), ML-(3), SSToT-(1,6,9), MH+DH-(1)
	USE-DSE-Mapped[5]	All-(3), ML-(3), MH+DH-(1), SSToT-(1,3,6,9)
NLI	BART	All-(9), ML-(3,6,9), MH-(3), MH+DH-(3)

Table 4.3: ZSL models with similar performance to the top-3 models in each ZSL family.

ZSL-Family	Model-Name	Macro-F1	Weighted-F1
WE	DSE-(MH+DH)-top-3	0.3071	0.3423
SE	USE-SE-All-top-1	0.3534	0.3558
NLI	BART-All-top-6	**0.4156**	0.4943
Supervised	BERT-Finetuned	0.3112	**0.51**

Table 4.4: DSD task macro and weighted F1 scores for the best models in each families in DSD-Clinician-Tweets-Original-Test (test) dataset.

Features	F1-Score
BART-All-top-6	0.7871(±0.0214)
NB-BOW	0.7521(±0.0277)
LR-LIWC	0.7397(±0.0418)
All-Majority	0.7006(±0.0)
Random-Uniform	0.5102(±0.0455)

Table 4.5: F1 scores in DPD task.

better predictability. In SE, we see SSToT performs almost equally good as "All" based on Macro-F1 score. Performance of NLI and WE using SSToT is very poor and are $\approx 0\%$ (not reported in the Table).

2. Based on Table 4.3, where we report other models which do not have statistically significant difference with the top-3 models under each family, we see that, other than DSE, where MH and DH based descriptors are best performing and for Google word embedding where "All" descriptors perform the best, we find all kinds of descriptors can provide similar results. Also, we see for embedding families (i.e., SE and WE), USE (and its augmented version) perform the best. For Top-k, we see that ≥ 3 descriptors provide best results in majority of the cases.

3. Pure depression specific embedding perform better than general embedding and augmented embedding methods, such as TE and ATEs.

4. In the test set (Table 4.4), all ZSL models are better than a strong supervised baseline, i.e., BERT-finetuned model in Macro-F1 score, however, Weighted-F1 score wise none of those models can beat this supervised baseline.

5. In Table 4.5, we see that best ZSL model based representation also helps achieve significantly good accuracy (p-value < 0.05) in the DPD task than other reasonable contenders which use mostly depression vocabulary, such as Logistic Regression with LIWC features (LR-LIWC) and Naïve Bayes with BoW features (NB-BOW), and all the naïve baselines.

We can summarize the above observations into following insights:

1. SE and NLI based ZSL models are superior than word embedding based models because of their pre-training on large dataset and also better algorithm for learning context better than avg. word embedding based models.

2. Number of top-k depends on the underlying ZSL model. Observation finds that usually top-k=3 is enough to provide good accuracy in DSD task.

3. WE performs better with short but informative label descriptors with which it has better vocabulary overlap, such as MADRS and DSM headers, but not SSToT with which it may have less vocabulary overlap. In NLI, we see slightly larger in size (e.g., in “All”) is better and for SE both “All” and short but more number of label descriptors (e.g. in SSToT) are better. It could be the fact that these contextual representation based ZSL models work best when they are provided with more and relevant context.

4. ZSL based representation of tweets achieve reasonably good accuracy in DPD task confirms the overall efficacy of such models in detecting signs of depression from tweets.

4.8 Relationship with Ongoing Work on Explainability

To further push the boundary of ZSL models capability to developing explainable [5] DSD models, we outline an explainability framework. The core idea behind the explainability framework is to find which phrases express the same depression symptoms label as the tweet. Moreover, we would also like to determine how those phrases are surrounded by their neighbouring phrases to contribute to the semantics of a label. To do so, we propose two algorithms as described in the following sections. Our first algorithm “Syntax Tree-Guided Semantic Explanation (STEP)” respects the syntax tree-based compositionality to explore n-grams inside the tweet. This compositionality may also contribute to the tweet having a particular label (Section 4.8.1). Our second algorithm, we call n-gram based explanation (ngramex), naively divides a tweet in its constituent n-grams (where “n” is predetermined) to help explain a tweet for its label (Section 4.8.2). Finally, in Section 4.8.3 we propose an Explanation Index (EI) function that provides higher scores for multiple minimal explanations for a Tweet-Label as opposed to single or lengthy explanations.

4.8.1 Syntax Tree Guided-Semantic Explanation (STEP)

According to generative linguistic theory by Chomsky [103], to understand the meaning of a sentence, humans combine words in at least two levels, such as (1) syntactic level and (2) semantic level. Since explaining a sentence requires understanding its semantics through syntactic composition, we reflect this theory in our explanation algorithm. First, we start by approximating semantic understanding of a tweet, T as a whole (or the label/depression symptoms expressed by it), then we gradually explore the nodes of the syntax tree for T in breadth-first manner and find out which n-grams (children of those nodes) also express the same, until all the nodes have been traversed. Finally, STEP returns the set of n-grams (where "n" is dynamic and $n \in \mathbb{Z}^{+}$) or "explanations," E, in descending order of membership-score with the tweet label. Here, we use the Algorithms 1 and 2 for finding out the candidate label of n-grams at each node of the syntax tree. It is to be noted, we consider the most expressed candidate label (i.e., the candidate label which has the highest membership-score with the label of T) for an n-gram, instead of multiple labels for the ease of understanding our explainability mechanism.

Algorithm 3: STEP

```
Data: $T$
Result: $E$
$Tree \leftarrow$ Syntax-Tree(T) ;
Tweet-Label $\leftarrow$ Label(T) ;
Explanation-Dictionary $\leftarrow \emptyset$ ;
while not Tree.traversedAllNode() do
    ;// Traversing the $Tree$ in Breadth-First order
     and from left-to-right nodes
    foreach $node \in Tree$ do
        node-Label $\leftarrow$ Label(n-gram(node)) ;
        if node-Label == Tweet-Label then
            node-Score $\leftarrow$ Score(n-gram(node), Tweet-Label) ;
            Explanation-Dictionary[n-gram(node)] $\leftarrow$ node-Score
        end
    end
end
$E \leftarrow$ explanations(sort-descending(Explanation-Dictionary) ;
return $E$
```

It is easy to see that we could use this process for each of the candidate labels for explainability analysis if needed. The entire process is described in Algorithm 3. Further for the sake of simplicity, let us assume a function "n-gram" takes the leaves of each node and returns the corresponding n-gram; function "Label" takes an n-gram and returns its most expressed label; function "Score" returns the membership-score of the n-gram and the label, and the function "Syntax-Tree" returns, as the name suggests, a syntax tree of the corresponding text/tweet.

4.8.2 N-gram Based Explanation (ngramex)

In this algorithm, we simply partition T into some pre-defined length of n-grams. Later we identify n-grams which have the same label as T, and return the list of those n-grams according to the descending order membership-score with a label, T.

4.8.3 Explanation Index Score (EI-Score)

We propose an Explanation Index (EI) score to evaluate our ZSL Models in terms of their explainability. We report EI scores for both STEP and ngramex, and analyze their agreement over different samples to compare and contrast. Let us assume a set of explanations, $E = \{e_1, e_2, ...e_n\}$ for a particular tweet for its label. Each e_i corresponds to an n-gram explanation of a tweet for its label. A function "length" returns the number of words in e_i, and the function "rank" returns the rank of a particular e_i in E. Since e_i's are in sorted order under E, the lower the rank the better the explanation. We can express our EI-Score for E as follows,

$$\frac{\sum_{i=0}^{n} EI_i}{n} \tag{4.4}$$

where,

$$EI_i = LengthScore(e_i) * RankScore(e_i) * Relevance(e_i) \tag{4.5}$$

$$LengthScore(e_i) = 1 - (length(e_i)/length(tweet)) \tag{4.6}$$

$$RankScore(e_i) = 1 - (rank(e_i)/n) \tag{4.7}$$

Models	STEP EI-Score (avg.)	ngramex EI-Score (avg.)
DSE-(MH+DH)-top-3	$0.1307(\pm0.1042)$	$0.1870(\pm0.1377)$
USE-SE-all-top-1	$0.1214(\pm0.1247)$	$0.0981(\pm0.1317)$
USE-SE-SSToT-top-3	$0.2325(\pm0.1240)$	$0.1921(0.1434)$
BART-all-top-6	$0.1435(\pm0.1126)$	$0.1580(\pm0.1444)$

Table 4.6: EI-Scores for top ZSL models and a SSToT based USE model reported in the Tables 4.2 and 4.4.

$$Relevance = \begin{cases} 1 & \text{if } Label(tweet) == Label(e_i) \\ 0 & \text{otherwise} \end{cases} \tag{4.8}$$

We can see that EI scores will be higher for multiple explanations over a single explanation, and short explanation over lengthy explanations. It is possible that ngramex with a certain “n” can have a better score according to this scoring system, however, ngramex has a high possibility of returning non-salient explanations which may not useful to humans (Table 4.7). The range of our EI-Score function is between 0 and 1 and the higher score indicates the better explanation. We report two kinds of analyses here, such as (1) regarding average EI-score for both STEP and ngramex (with n=3, because we found it’s often scores better) in DSD-Clinician-Tweets-Original dataset. In Table 4.6 we report these average EI-Scores for our top ZSL models to analyze their explainability performance and (2) regarding the analysis of the rationale behind disagreement between STEP and ngramex EI-score; we sort out few examples where STEP EI-Score is greater than the same for ngramex and vice-versa (Table 4.7). Finally, it should be mentioned that, EI-Score is not a fool proof scoring system. It scores higher for smaller explanations even though those may not make sense. Quality of explanation depends on both underlying ZSL model and the explanation mechanism. Here STEP works as a layer to reduce gibberish explanations as opposed to ngramex.

4.8.4 Explainability Analysis

EI-Score

We observe that STEP EI-score wise, sentence embedding based model (USE-SE-SSToT-top-3) achieves significantly better score (p-value < 0.05 in a paired t-test)

Tweet	Condition	Exps (STEP)	Exps (ngramex)
“No one understands me”	$EI(STEP) > EI(ngramex)$	“No one”, “No one understands me”	“No one understands”, “one understands me”
“I feel like utter shit”	$EI(STEP) < EI(ngramex)$	“feel like utter shit”, “shit”	“I feel like”, “feel like utter”

Table 4.7: Top 2 EI explanations for the label ”Feelings of Worthlessness” for two tweets, where STEP & ngramex disagree for top EI-Scoring ZSL model: (USE-SE-SSToT-top-3).

than all the other models followed by NLI based BART-all-top-6 and word embedding based DSE-(MH+DH)-top-3 model (Table 4.6). Interestingly, BART-all-top-6 and USE-SE-all-top-1 achieve significantly high accuracy for DSD task (Table 4.2), although in-terms of explainability they are as bad as WE based models. It could be due to the fact that, STEP extracted n-grams are sometimes too short for WE and NLI based models to properly represent them. Same is true for SE, but with SSToT lexicon it performs better. It needs further investigation why SE performs better with short descriptors in this case.

We also observe that there is a difference between ngramex and STEP EI-score for a particular model and for WE and BART models ngramex is higher than STEP. This could be due to the fact that, STEP is capable of extracting explanations which are semantically consistent compared to inconsistent n-grams often extracted by ngramex.

In Table 4.7, we see two examples, where in first example STEP explanations provide high score (0.15) than the same for ngramex (0.1), the reason for EI-Score penalization for ngramex is that, the first explanation is almost the same size as the original tweet. In the second example, ngramex EI-Score is higher (0.21) than STEP (0.18), here the EI-score penalization for STEP is because of the same reason as first example, however, if we see the semantic quality of the explanations, STEP explanations are better than ngramex.

4.9 Limitations

In this chapter, we present a ZSL framework as a viable option even in the situation when we have no labelled samples for training, however, evaluating such model has some core limitations, which are as follows:

1. We have experimented with various combinations of ZSL models, label descriptors and several modelling related parameters, such as top-k. In absence of large test data, we cannot come to a final conclusion on the best combinations of these attributes, rather we get a very approximate idea. Unfortunately, human annotated depression symptoms samples are rare to find and also very hard to create in large scale.

2. Explainability score is not fool-proof, which means, models with good explainability may have less score, depending on the underlying models used. We require human judges to find whether the EI-Score goes with human intuition of explainability quality.

3. We do not extensively experiment with different scenarios that can result into different structure of a syntax tree. Also, we depend on the accuracy of the syntax tree in this process. All our analyses are based on a mind-set that each tweet is a single sentence.

4.10 Conclusion

In this chapter, we discuss on a way of tackling the main challenge in DSD task, which is the scarcity of labelled samples for training the same. We show that using various learned representation techniques and their enhancement we can formulate a ZSL approach for this task, which performs better than a strong fine-tuned BERT-based supervised baseline for the same. Moreover as an ongoing work on text explainability, we provide an outline of an algorithm for exploring syntax tree for sub-phrases that explains a particular tweet for its label. Finally as a part of that framework, we propose EI-Score that can be used to evaluate the explainability capability of our Zero-Shot models.

Chapter 5

Depression Symptoms Modelling from Text: A Semi-supervised Learning Approach

In this chapter, we describe a Semi-supervised Learning (SSL) framework (Appendix A.4), which uses an initial supervised learning model that leverages 1) a state-of-the-art large mental health forum text pre-trained language model further fine-tuned on a clinician annotated Depression Symptoms Detection (DSD) dataset, 2) a Zero-Shot Learning (ZSL) model for DSD (Chapter: 4), and couples them together to harvest depression symptoms related samples from our large self-curated Depressive Tweets Repository (DTR). Our clinician annotated dataset is the largest of its kind. Furthermore, DTR is created from the samples of tweets in self-disclosing depressed users Twitter timeline from two datasets, including one of the largest benchmark datasets for user-level depression detection from Twitter. This further helps preserve the depression symptoms distribution of self-disclosing Twitter users' tweets. Subsequently, we iteratively retrain our initial DSD model with the harvested data. We discuss the stopping criteria and limits of this SSL process, and elaborate the underlying constructs which play a vital role in the overall SSL process. We show that, we can produce a final dataset which is the largest of its kind. Furthermore, a DSD and a Depressive Post Detection (DPD) model trained on it achieves significantly better accuracy than their initial version.

5.1 Motivation

The first and foremost challenge for developing a robust DSD model is to curate enough samples of social media posts, such as tweets, carrying signs of clinical depression symptoms, that could help train such a model. Most of the earlier works attempt to annotate samples with the help of primarily non-expert (non-clinician) humans to curate a ground truth [86]–[88], [152]. As mentioned in the last chapter and in Chapter 2, to create the initial candidate sets of samples for annotation, most of these studies crawl random tweets based on depression specific keywords [53], [87], [88], [152]. This sampling does not necessarily reflect the language patterns of the depression population.

Provided we have a lot of users who have disclosed their depression diagnosis in their Twitter timeline, we can use their timeline to gather possible depression candidate tweets with the help of DPD and ZSL models we outlined in the earlier chapters and then annotate them with the help of the expert annotators (in our case the annotators having clinical knowledge). Further, we can learn our very first model with the help of this dataset and iteratively improve it. All these ideas together have not been explored in the earlier research for developing a robust DSD model. Thus the motivation of this work comes mainly from the following points:

1. **Clinician annotated dataset creation from depressed users tweets:** Through leveraging our existing datasets from self-disclosing depressed users and learned DPD model, which is a binary model for detecting signs of depression, we want to curate a clinician annotated dataset for depression symptoms. This is a more in-situ approach for harvesting depression symptoms posts compared to crawled tweets for depression symptoms using depression symptoms keywords, as done in most of the earlier literature [53], [87], [88], [152]. We call it in-situ because this approach respects the natural distribution of depression symptoms samples found in the self-disclosing depressed users timeline. Although Jamil et al. [152] and Yadav et el. [53] collected samples in-situ as well, both of these works did not strictly consider a user is depressed based on their self-disclosure statement of depression diagnosis like us. Moreover, our

clinician annotated dataset is much bigger and annotation is more rigorous than them.

2. **Gather more data that reflects clinical insight:** Starting from the small dataset found at (1) and a learned DSD model on that, we want to iteratively harvest more data and retrain that model for our depression symptoms modelling or DSD task, which is also not explored in the early work.

5.2 Methodology

To achieve the goals mentioned earlier, we divide our depression symptoms modelling into two parts: (1) **Clinician Annotated Dataset Curation:** here we first propose a process to create our annotation candidate dataset from our existing depressive tweets from self-disclosing depressed Twitter users timeline. We later annotate this dataset with the help of a clinician amongst others, that helps us achieve our first goal and (2) **Semi-supervised Learning:** we then describe how we leverage that dataset to learn our first sets of DPD and DSD models and eventually make them robust through iterative data harvesting and retraining or semi-supervised learning.

From our clinician annotated dataset created in step (1), we separate a subset of depression symptoms stratified samples as a test set. After each step of the SSL process, we report Macro-F1 and Weighted-F1 scores (Appendix A.1) to evaluate the efficacy of that step based on that test set.

5.3 Datasets

We use depressed users who disclosed their depression condition through a self-disclosure statement, such as "I (am/was/have) been diagnosed with depression" in IJCAI-2017 dataset [133] (Chapter 2) and separate a portion of it. This portion contains users who have 80% or more of their tweets written in English and minimum 50 posts in their Twitter timeline. The other portion of the IJCAI-2017 dataset where users do not satisfy the above condition and UOttawa depressed users' tweets [53] (Chapter 2), where the users were verified by annotators about their ongoing depressive episodes, are used to develop Depression-Candidate-Tweets. Later, we further filter it for depressive tweets and create Depressive Tweets Repository (DTR) which is used in our SSL process to harvest in-situ tweets for depression symptoms. Next, we separate a portion of DTR for clinician annotation for depression symptoms (Figure 5.3). We also use a dataset curated by Yadav et al. [152] called D2S dataset. This dataset is curated from Twitter users who were identified as depressed based on their profile attributes, such as profile names, pictures and contents. Later, their tweets were analyzed with the help of non-expert annotators for nine depression symptoms and only 100 sub-samples were cross checked with a clinician to find the reliability of their annotation.

5.3.1 Clinician Annotated Dataset Curation

In the overall DSD framework, depicted in Figure 5.1, we are ultimately interested in creating a robust DPD and a DSD model which are initially learned on human annotated samples, called "DPD-Human" model and "DSD-Clinician" model (Figure 5.2). The suffixes with these model names, such as "Human," indicates that this model leverages the annotated samples from both non-clinicians and clinicians; "Clinician" indicates that this model leverages the samples for which the clinician's annotation is taken as more important (more explanation in Section 5.3.4). At the beginning of this process, we have only a small human annotated dataset for depression symptoms and depressive posts from external organizations (i.e., D2S [152] and DPD-Vioules [149] datasets respectively), no clinician annotated depression symptoms samples, and a large dataset from self-disclosing depressed users (i.e., IJCAI-2017 and UOttawa datasets). We take the following steps to create our first set of clinician annotated depression symptoms dataset and DTR which we will use later for our SSL.

1. We start the process with the help of a DPD model, which we call DPD Majority Voting model (DPD-MV). It consists of a group of DPD models (Chapter: 3), where each model leverages pre-trained word embedding (both augmented (ATE) and depression specific (DSE)), and sentence embedding (USE); further trained on a small set of human annotated depressive tweets and a Zero-Shot Learning model (USE-SE-SSToT) (Chapter 4). Subsequently,

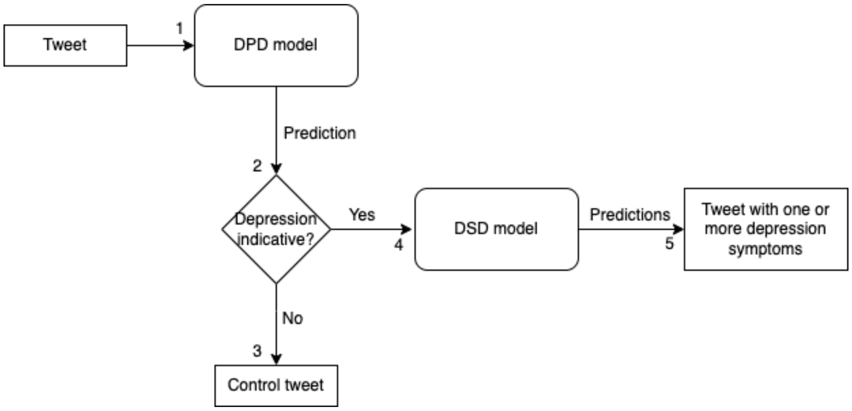


Figure 5.1: DSD modelling algorithm.

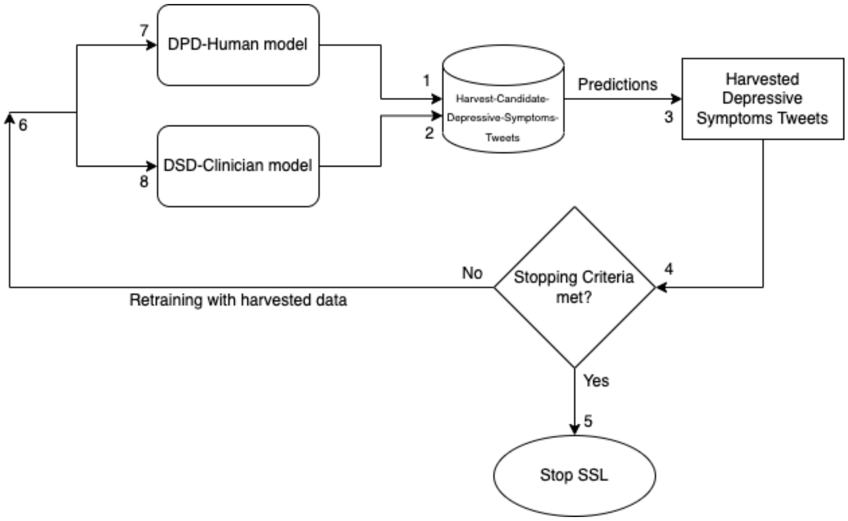

7
DPD-Human model
8
DSD-Clinician model
1
2
Harvest-Candidate-Depressive-Symptoms-Tweets
Predictions
3
Harvested Depressive Symptoms Tweets
4
Stopping Criteria met?
No
Retraining with harvested data
6
Yes
5
Stop SSL

Figure 5.2: Semi-supervised learning process at a high level.

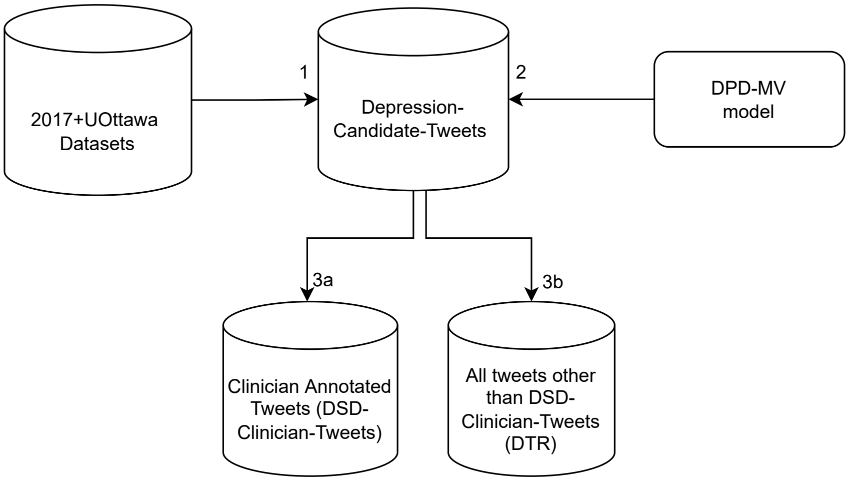

2017+UOttawa Datasets
1
Depression-Candidate-Tweets
2
DPD-MV model
3a
3b
Clinician Annotated Tweets (DSD-Clinician-Tweets)
All tweets other than DSD-Clinician-Tweets (DTR)

Figure 5.3: DSD-Clinician-Tweets and DTR curation process.

Dataset	Sample size	Comment
Depression-Candidate-Tweets	42,691	Depressed users' tweets
DTR	6,077	Depressive Tweets Repository
DSD-Clinician-Tweets	1,500	Clinician annotated tweets

Table 5.1: Datasets

the DPD-MV model takes the majority voting of these models for detecting depressive tweets.

2. We then apply DPD-MV on the sets of tweets collected from depressed users timelines (or **Depression-Candidate-Tweets** (Figure 5.3)) to filter control tweets. The resultant samples, after applying DPD-MV is referred to as Depressive Tweets Repository (**DTR**). We later separate a portion of this dataset, e.g., 1500 depressive tweets for human annotation which we call **DSD-Clinician-Tweets** dataset. Details of the annotation process are described in Section 5.3.4.

3. We learn our first DSD model using this dataset, then use this model to harvest more samples from DTR. An outline of the DTR and DSD-Clinician-Tweets curation process is provided in Figure 5.3. We describe the details of this process in the Semi-supervised Learning section, but describe each of its building block in the next sections. In Table 5.1, we provide relevant datasets description.

5.3.2 Annotation Task Description

Our annotation task consists of labelling a tweet for either 1) one or more of 10 symptoms of depression, 2) No Evidence of Depression (NoED), 3) Evidence of Depression (ED) or 4) Gibberish. We have 10 labels instead of the traditional nine depression symptom labels because we separate the symptom “Agitation/Retardation” into two categories so that our model can separately learn and distinguish these labels, unlike previous research [152]. NoED indicates the absence of any depression symptoms expressed in a tweet. ED indicates multiple symptoms of depression expressed in a tweet in a way so that it’s hard to specifically pinpoint these combined depression symptoms in that tweet. Gibberish is a tweet less than three words long and, due to the result of crawling or data preprocessing, the tweet is not complete and it’s hard to infer any meaningful context.

5.3.3 Annotation Guideline Creation

To create the annotation guideline for the task, we analyze the textual descriptions of depression symptoms from most of the major depression rating scales, such as PHQ-9, CES-D, BDI, MADRS and HAM-D [76]. We also use DSM-5 as our base reference for symptoms description. Based on these descriptions of the symptoms from these resources and several meetings with our clinicians, we consolidate some of the most confusing samples of tweets from DTR and map them to one or more of those depression symptoms. We then create an annotation guideline with clear description of the clinical symptoms of depression that an annotator should look for in the tweets followed by relevant tweets examples for them including the confusing ones previously noted. We then separate a portion of 1500 samples from our DTR and provide it to the annotators along with our annotation guideline. During the annotation we randomly assign a set of tweets multiple times to calculate test-retest reliability scores. We find annotators annotate the tweets consistently with the same annotation with 83% reliability based on the test-retest reliability score. Our detailed guideline description is provided in Appendix A.7.

5.3.4 Depression Symptoms Annotation Process

We provide a portion of 1500 tweets from DTR for depression symptoms annotation by four annotators[1]. Among these annotators two have clinical understanding of depression: one is a practicing clinician and the other one has a PhD in Psychiatry.

In our annotation process, we emphasize the annotation of a tweet based on the clinical understanding of depression which is laid out in our annotation guideline (Appendix A.7). We take majority voting to assign a label for the tweet. In absence of majority, we assign a label based on the clinician's judgment, if present, otherwise, we do not assign a label to that tweet. We call this scheme **MVCP**. Table 5.2 reports the average kappa scores for each label and Annotator-Annotator, Annotator-MVCP and All pairs (i.e., average on both of the previous schemes).

We observe fair to moderate kappa agreement score (0.38 - 0.53) among our annotators for all the labels. We also find, "Suicidal thoughts" and "Change in Sleep Patterns" are the labels for which inter-annotator agreement is the highest and agreement between each annotator and MVCP is substantial for the same. Among the annotators the order of the labels based on descending order of agreement score is as follows: Suicidal Thoughts, Change in Sleep Patterns, Feelings of Worthlessness, Indecisiveness, Anhedonia, Retardation, Weight Change, NoED, Fatigue, Low mood, Gibberish, Agitation and ED. With MVCP, we find moderate to substantial agreement (0.56 - 0.66).

5.3.5 Distribution Analysis of the Depression Symptoms Datasets

In this section, we provide symptoms distribution analysis for our D2S and DSD-Clinician-Tweets datasets. DSD-Clinician-Tweets dataset contains 1500 tweets. We then create a clean subset of this dataset which holds clinician's annotations and only tweets with depression symptoms, which we call DSD-Clinician-Tweets-Original (further detail is in Section 5.5.1). For D2S, we have 1584 tweets with different depression symptom labels. In Figure 5.4, the top 3 most populated labels for DSD dataset are Agitation, Feeling of Worthlessness and Low Mood. However,

[1] Thanks to our annotators, Sudhakar Sivapalan, Jasmine Noble and Katrina Ingram

Depression-Symptom-Labels	**Average(Annots.)**	**Average(Annots.- MVCP)**	**Average(All)**
Suicidal thoughts	$0.5319(\pm0.1045)$	$0.6296(\pm0.1227)$	$0.5710(\pm0.1167)$
Change in Sleep Pattern	$0.5171(\pm0.0770)$	$0.6162(\pm0.1034)$	$0.5568(\pm0.0973)$
Feelings of Worthlessness	$0.4517(\pm0.1978)$	$0.6589(\pm0.2347)$	$0.5346(\pm0.2271)$
Indecisiveness	$0.4475(\pm0.2164)$	$0.6378(\pm0.2479)$	$0.5236(\pm0.2370)$
Anhedonia	$0.4434(\pm0.2383)$	$0.6037(\pm0.0915)$	$0.5076(\pm0.2030)$
Retardation	$0.4382(\pm0.3030)$	$0.5672(\pm0.2446)$	$0.4898(\pm0.2746)$
Weight Change	$0.4358(\pm0.1589)$	$0.6155(\pm0.2149)$	$0.5077(\pm0.1951)$
NoED	$0.4321(\pm0.2119)$	$0.5946(\pm0.2631)$	$0.4971(\pm0.2346)$
Fatigue	$0.4297(\pm0.1136)$	$0.5975(\pm0.2375)$	$0.4968(\pm0.1830)$
Low Mood	$0.4251(\pm0.3041)$	$0.6454(\pm0.3730)$	$0.5132(\pm0.3327)$
Gibberish	$0.4172(\pm0.2606)$	$0.6626(\pm0.3272)$	$0.5154(\pm0.2991)$
Agitation	$0.4008(\pm0.2066)$	$0.6505(\pm0.2571)$	$0.5007(\pm0.2498)$
ED	$0.3877(\pm0.0878)$	$0.5765(\pm0.2742)$	$0.4632(\pm0.1971)$

Table 5.2: Pairwise kappa scores among annotators and MVCP for all the labels.

for D2S dataset Suicidal Thought is the most populated label followed by Feeling of Worthlessness and Low Mood, just like DSD.

We use D2S dataset because D2S dataset curators crawled tweets from self-reported depressed users timeline. Although they did not confirm whether these users have also disclosed their depression diagnosis, they mention that they analyze their profile to ensure that these users are going through depression. Since their annotation process is not as rigorous as ours, i.e., they did not develop an annotation guideline as described in the earlier section and their depressed users dataset may not contain all self-disclosing depressed users, we had to further filter those tweets before we could use them. So we use DSD-Clinician-Original-Tweets for training our very first model in SSL process, later use that model to re-label D2S samples.

To make a contrast with a substantial early work for Twitter symptoms annotation on depression lexicon based keyword crawled tweets from random users, we compare the symptoms distribution of above mentioned datasets with depression symptoms annotated SAD corpus [87]. We find, the most prevalent depression symptoms label in that dataset are Low Mood, Fatigue and Disturbed sleep. There are few to no samples of other depression symptoms in that corpus. This indicates annotated samples in self-disclosing Twitter users timeline is substantially different than that of keyword based crawling from random users timeline. In a later section, we report the depression symptoms distribution on our harvested data and another approach for increasing sample size for least populated labels.

5.4 Data Preprocessing

We use the same data preprocessing for the tweets as described in Chapter 3, Section 3.4.

5.5 Experimental Setup and Evaluation

Our experimental setup consists of iterative data harvesting and re-training of a DSD model, followed by observing its accuracy increase over each iteration coupled with incremental initial dataset size increase. We report the results separately

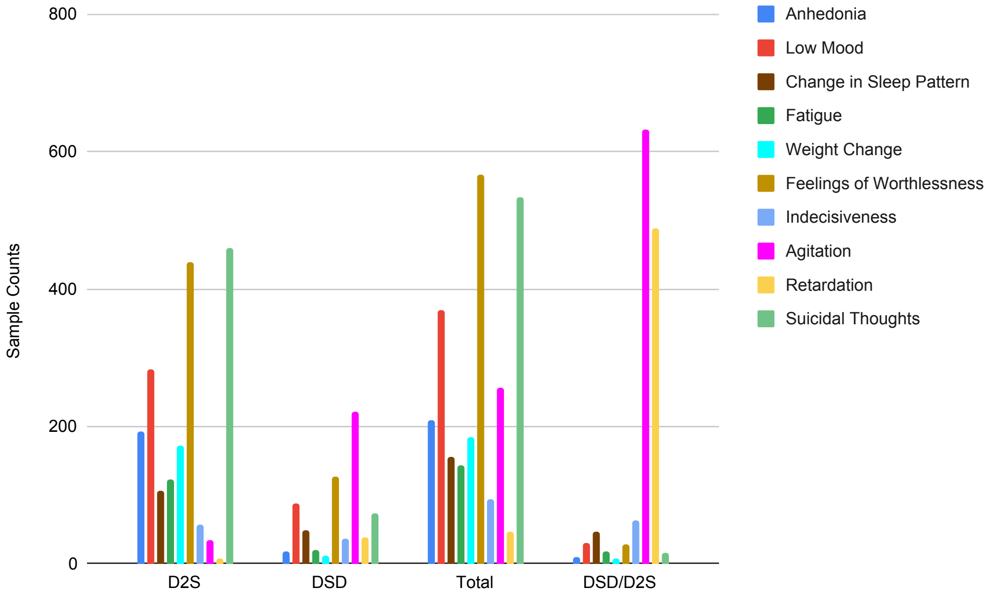


Figure 5.4: Sample distribution and ratio analysis across D2S and DSD-Clinician-Tweets-Original-Train datasets.

for each of the step of SSL in next sections. DSD is a multi-class multi-label problem. We report accuracy measures in Macro-F1 and Weighted-F1 (Appendix A.1).

Our Semi-supervised Learning (SSL) strategy uses the DPD and DSD models and the datasets as described in earlier sections to iteratively harvest more relevant samples and learn robust models (Figure 5.5).

5.5.1 A Semi-supervised Learning (SSL) Framework

In our SSL framework, we iteratively perform data harvesting and retraining of our DSD model, which is a multi-label text classifier utilizing pre-trained Mental-BERT[2], technical details of this model (i.e., the training hyperparameters and setup) is provided in the Appendix A.10. We find Mental-BERT based DSD performs significantly better in terms of Macro-F1 and Weighted-F1 scores compared to base BERT only models in the DSD task (Tables 5.5 and 5.6). In this section, we provide our step by step SSL process description, datasets used at each step and the resulting models and/or datasets. All our steps are depicted in points 11-26 in Figure 5.5 and

[2] `https://huggingface.co/mental/mental-bert-base-uncased`

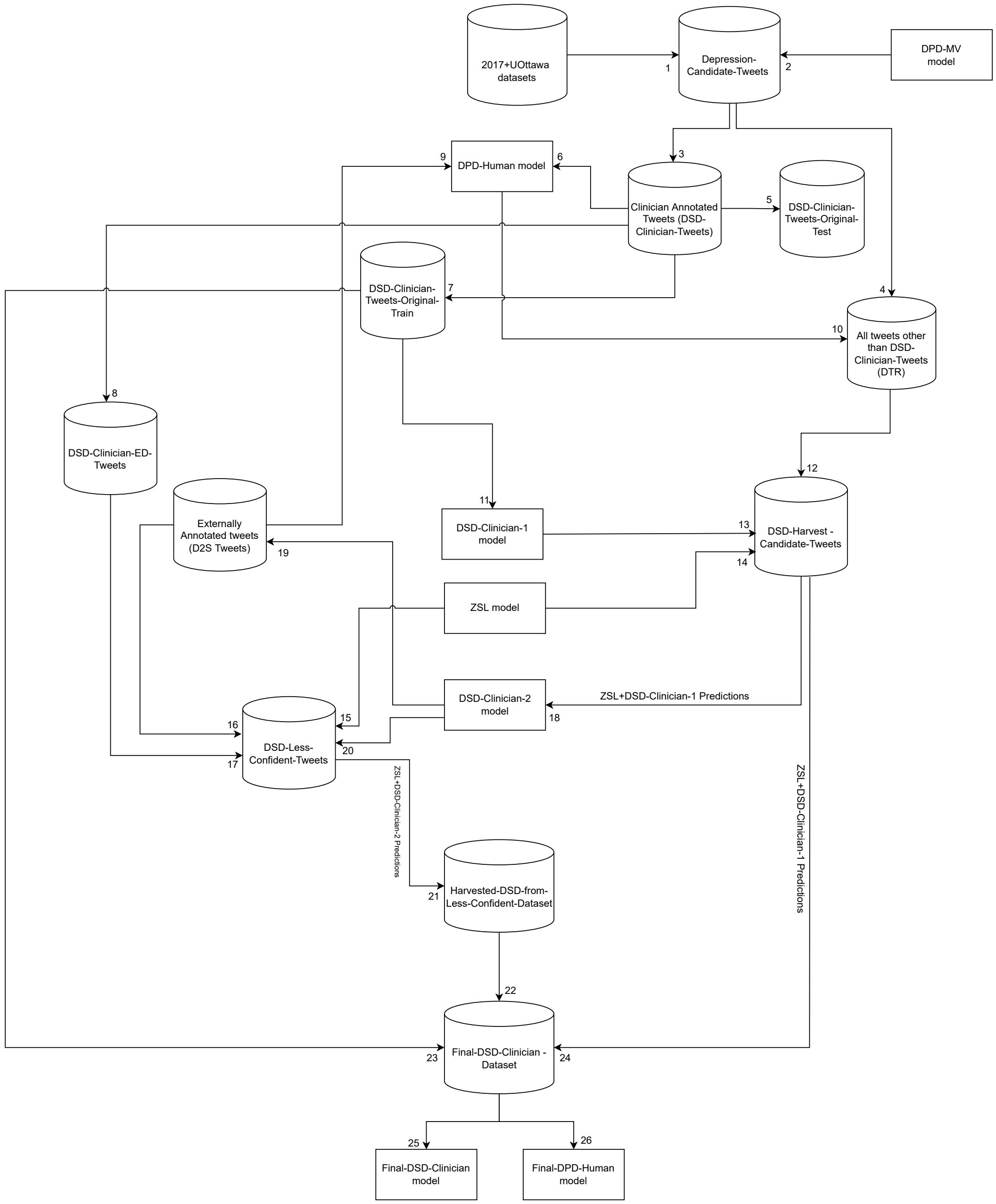

2017+UOttawa datasets
1
Depression-Candidate-Tweets
2
DPD-MV model
3
DPD-Human model
9
6
Clinician Annotated Tweets (DSD-Clinician-Tweets)
5
DSD-Clinician-Tweets-Original-Test
DSD-Clinician-Tweets-Original-Train
7
4
10
All tweets other than DSD-Clinician-Tweets (DTR)
8
DSD-Clinician-ED-Tweets
Externally Annotated tweets (D2S Tweets)
19
11
DSD-Clinician-1 model
12
13
DSD-Harvest - Candidate-Tweets
14
ZSL model
DSD-Clinician-2 model
ZSL+DSD-Clinician-1 Predictions
18
16
15
DSD-Less-Confident-Tweets
20
17
ZSL+DSD-Clinician-2 Predictions
ZSL+DSD-Clinician-1 Predictions
21
Harvested-DSD-from-Less-Confident-Dataset
22
23
Final-DSD-Clinician - Dataset
24
25
26
Final-DSD-Clinician model
Final-DPD-Human model

Figure 5.5: Detailed Semi-supervised Learning (SSL) framework. Here, we show the interaction among our datasets and models. Datasets are shown as cylinders, and models are shown as rectangles. An arrow from a dataset to another dataset represents data subset creation; an arrow to another model means the provision of training data for that model; and an arrow from a model to a dataset means use of that model to harvest samples from the dataset. All the arrow heads are marked, so that these can be easily referred while describing a particular scenario in the SSL framework.

described further below.

Step 1: Creating First DSD Model

In this step, we focus on the creation of a training dataset and a test dataset selected from our clinician annotated samples. This dataset consists of tweets carrying at least one of the 10 depression symptoms. We use this training dataset to create our first DSD model, called **DSD-Clinician-1**. To do so, we follow the steps stated below.

1. We first remove all the tweets with labels "Gibberish," "Evidence of Depression" (ED) and "No Evidence of Depression" (NoED) from a subset of DSD-Clinician-Tweets after applying MVCP. We call this dataset **DSD-Clinician-Tweets-Original**. Details of ED, NoED and Gibberish is provided in Table 5.3.

2. We save the tweets labelled as "Evidence of Depression," which we call **DSD-Clinician-ED-Tweets**, (Arrow 8 in Figure 5.5). We later use those to harvest depression symptoms related tweets.

3. Next, we separate 70% of the tweets from DSD-Clinician-Tweets-Original dataset and create **DSD-Clinician-Tweets-Original-Train** dataset for training our first version of DSD model, called **DSD-Clinician-1** and the rest 30% of the tweets are used as an SSL evaluation set, also called, **DSD-Clinician-Tweets-Original-Test**, (Arrows 5 and 7 in the Figure 5.5). This split is done based on stratified sampling on label distribution. We will use this evaluation set all through our SSL process to measure the performance of SSL, i.e., whether it helps increase accuracy for DSD task or not. Further, we separate 20% of samples from DSD-Clinician-Tweets-Original-Train as a validation set to find out a optimal threshold to predict labels through sigmoid activation functions for each labels. We report the datasets created in this step in Table 5.3, models in Table 5.4 and accuracy scores for each labels and their average in Table 5.6. We also report accuracy for the DPD-Human model in this step in Table 5.7.

Dataset	Sample size	Comment
DSD-Clinician-Tweets-Original	539	Tweets with depression symptoms only
DSD-Clinician-Tweets-Original-Train	377	Initial train dataset
DSD-Clinician-Tweets-Original-Test	162	Overall test dataset
DSD-Clinician-ED-Tweets	135	Depressive tweets
DSD-Clinician-NoED-Tweets	785	Control tweets
DSD-Clinician-Gibberish-Tweets	41	Gibberish tweets

Table 5.3: Datasets in step 1.

Model	Train dataset	Sample size	Comment
DSD-Clinician-1	DSD-Clinician-Tweets-Original-Train	377	DSD-Clinician model at SSL iteration 1
DPD-Human	(DSD-Clinician-Tweets + D2S − (DSD-Gibberish-Tweets + DSD-NoED-Tweets + Tweets with self-disclosure)) + equal number of NoED tweets from DTR	$(1500 + 1584 - (41+785+34)) + 2224 = 4448$	DPD-Human model at SSL iteration 1

Table 5.4: Model details in step 1.

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.00	0.00	0.00	5
Low mood	0.00	0.00	0.00	26
Change in sleep pattern	1.00	0.07	0.12	15
Fatigue	0.00	0.00	0.00	6
Weight change	0.00	0.00	0.00	4
Feelings of worthlessness	0.55	0.16	0.24	38
Indecisiveness	0.00	0.00	0.00	11
Agitation	0.55	0.73	0.62	66
Retardation	0.00	0.00	0.00	12
Suicidal thoughts	1.00	0.14	0.24	22
Macro avg	0.31	0.11	0.12	205
Weighted avg	0.46	0.28	0.28	205

Table 5.5: DSD-Clinician-1 (BERT) model accuracy.

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.00	0.00	0.00	5
Low mood	0.61	0.42	0.50	26
Change in sleep pattern	0.76	0.87	0.81	15
Fatigue	0.00	0.00	0.00	6
Weight change	0.00	0.00	0.00	4
Feelings of worthlessness	0.49	0.53	0.51	38
Indecisiveness	0.00	0.00	0.00	11
Agitation	0.63	0.77	0.69	66
Retardation	0.00	0.00	0.00	12
Suicidal thoughts	0.91	0.45	0.61	22
Macro avg	0.34	0.30	0.31	205
Weighted avg	0.52	0.51	0.51	205

Table 5.6: DSD-Clinician-1 (Mental-BERT) model accuracy in step 1.

Precision	Recall	F1-score	Support
0.84	0.90	0.87	227

Table 5.7: DPD-Human model accuracy in step 1.

Step 2: Harvesting Tweets Using DSD-Clinician-1

In this step, we use DSD-Clinician-1 model created in the previous step to harvest tweets which carry signs of depression symptoms from a set of tweets filtered for carrying signs of depression only by **DPD-Human** model from DTR, we call this dataset **DSD-Harvest-Candidate-Tweets** (Arrows 10 and 12 in Figure 5.5). Our DPD-Human model is trained on all available human annotated datasets, i.e., DSD-Clinician-Tweets-Original and D2S tweets and equal number of control tweets from DTR (Arrows 6 and 9 in Figure 5.5 and more dataset details in Table 5.4). We use this model to leverage human insights to further filter DTR. All the datasets used for training purpose of DPD-Human and its incremental versions are stratified sampled and 90% of the samples are used as train and the rest as test. In this step, we create two more datasets from DSD-Harvest-Candidate-Tweets, (1) **Harvested-DSD-Tweets:** This dataset contains the tweet samples for which the model is confident, i.e., it detects at least one of the 10 depression symptoms and (2) **Harvested-DSD-Tweets-Less-Confident:** This dataset contains the tweet samples for which the model has no confident predictions or it does not predict any depression symptoms.

Dataset	Sample size	Comment
DSD-Harvest-Candidate-Tweets	3145	Harvestable tweets for DSD
Harvested-DSD-Tweets	2491	First harvested dataset
Harvested-DSD-Tweets-Less-Confident	654	First harvested less confident dataset

Table 5.8: Datasets in step 2.

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.00	0.00	0.00	5
Low mood	0.71	0.46	0.56	26
Change in sleep pattern	0.70	0.93	0.80	15
Fatigue	0.00	0.00	0.00	6
Weight change	0.00	0.00	0.00	4
Feelings of worthlessness	0.44	0.63	0.52	38
Indecisiveness	0.00	0.00	0.00	11
Agitation	0.62	0.77	0.69	66
Retardation	0.00	0.00	0.00	12
Suicidal thoughts	0.80	0.55	0.65	22
Macro avg	0.33	0.33	0.32	205
Weighted avg	0.51	0.55	0.52	205

Table 5.9: DSD-Clinician-1 model accuracy in step 2.

Harvested dataset statistics is provided in Table 5.8.

Step 3: Harvesting Tweets Using Best ZSL Model

In this step, we use a ZSL model (USE-SE-SSToT) described in Chapter 4, to harvest tweets carrying signs of depression symptoms from the DSD-Harvest-Candidate-Tweets. We choose this model because it has reasonable accuracy in the DSD task and it is fast. We also set a threshold while finding semantic similarity between the tweet and the label descriptor to be more on a conservative side so that we reduce the number of false positive tweets. We find that a threshold < 1 is a reasonable choice because cosine-distance < 1 indicates higher semantic similarity. In this step, we create two datasets: (1) **Only-ZSL-Pred-on-Harvested-DSD-Tweets (step: 3a):**

Dataset	Sample size	Comment
ZSL-and-Harvested-DSD-Tweets	2491	Second harvest, sample size is same as Harvested-DSD-Tweets because harvesting is done on the same data
Only-ZSL-Pred-on-Harvested-DSD-Tweets	2248	Sample size less than the above because we are not using samples with no labels predicted

Table 5.10: Datasets in step 3.

This dataset is only ZSL predictions on DSD-Harvest-Candidate-Tweets. (2) **ZSL-and-Harvested-DSD-Tweets (step: 3b):** This dataset is a combination of ZSL predictions and DSD-Clinician-1 predictions on DSD-Harvest-Candidate-Tweets. This means, if either DSD-Clinician-1 or ZSL detects a depression symptom label for a sample in DSD-Harvest-Candidate-Tweets, then we assign that label to that sample. We follow steps: 3a and 3b to compare whether datasets produced through these steps help in accuracy gain after using them to retrain DSD-Clinician-1.

Compared to step: 1 (Table 5.6), we achieve 4% gain in Macro-F1 and 5% gain in Weighted-F1 using the combined dataset in step: 3b (Table 5.12). We achieve 1% gain in both the measures using Harvested-DSD-Tweets only in step: 2. With ZSL only in step: 3a (Table 5.11), we actually lose 3% in Macro-F1 and 15% in Weighted-F1. We also provide our produced datasets description in Table 5.10.

Step 4: Creating a Second DSD Model

From the previous experiments, we now create our second DSD model by retraining it with DSD-Clinician-Tweets-Original-Train and ZSL-and-Harvested-DSD-Tweets. This results in our second DSD model (Table 5.13).

Step 5: Creating Final DSD Model

In this final step, we do the following:

1. We create a combined dataset from D2S and DSD-Clinician-ED-Tweets and we call this combined dataset **DSD-Less-Confident-Tweets** dataset (Arrows

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.00	0.00	0.00	5
Low mood	0.56	0.85	0.68	26
Change in sleep pattern	0.72	0.87	0.79	15
Fatigue	0.00	0.00	0.00	6
Weight change	0.00	0.00	0.00	4
Feelings of worthlessness	0.33	0.55	0.42	38
Indecisiveness	0.00	0.00	0.00	11
Agitation	1.00	0.11	0.19	66
Retardation	0.00	0.00	0.00	12
Suicidal thoughts	0.82	0.64	0.72	22
Macro avg	0.34	0.30	0.28	205
Weighted avg	0.60	0.38	0.36	205

Table 5.11: DSD-Clinician-1 model accuracy in step 3a.

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.00	0.00	0.00	5
Low mood	0.71	0.92	0.80	26
Change in sleep pattern	0.68	0.87	0.76	15
Fatigue	0.00	0.00	0.00	6
Weight change	0.00	0.00	0.00	4
Feelings of worthlessness	0.34	0.82	0.48	38
Indecisiveness	0.00	0.00	0.00	11
Agitation	0.65	0.82	0.72	66
Retardation	0.00	0.00	0.00	12
Suicidal thoughts	0.76	0.73	0.74	22
Macro avg	0.31	0.42	0.35	205
Weighted avg	0.49	0.67	0.56	205

Table 5.12: DSD-Clinician-1 model accuracy in step 3b.

Model	Train dataset	Sample size	Comment
DSD-Clinician-2	DSD-Clinician-Tweets-Original-Train + ZSL-and-Harvested-DSD-Tweets	$(377 + 2491) = 2868$	DSD model at SSL iteration 2

Table 5.13: Model details in step 4.

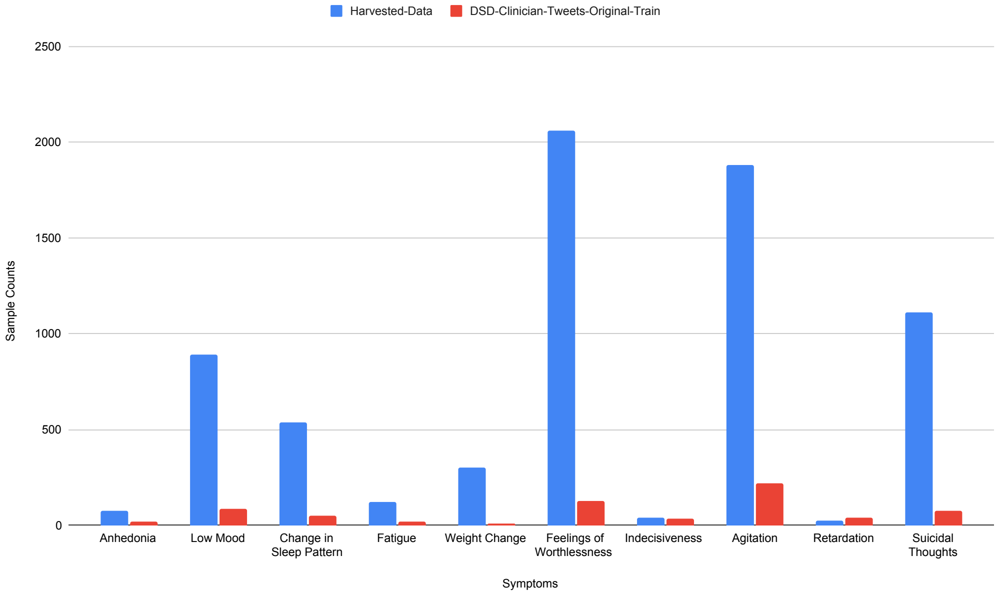


Figure 5.6: Sample distribution in harvested dataset vs original clinician annotated dataset.

15, 16, 17, 20 in Figure 5.2). D2S tweets are used here because the dataset was annotated externally with a weak clinical annotation guideline. We use our model to further filter this dataset.

2. We use DSD-Clinician-2 model and ZSL to harvest depression symptoms tweets from DSD-Less-Confident-Tweets, we call this dataset **ZSL-and -Harvested-DSD-from-Less-Confident-Tweets**. Finally with this harvested data and the datasets used to train DSD-Clinician-2 model, we create our final dataset called Final-DSD-Clinician-Tweets and by training with it, we learn our final DSD model called, Final-DSD-Clinician. We also retrain our DPD-Human model to create Final-DPD-Human model. Datasets, models and the relevant statistics are reported in Tables 5.14, 5.15, 5.16, and 5.17.

   We report the symptoms distribution for our DSD-Clinician-Tweets-Original-Train dataset earlier, and here report depression symptoms distribution in our SSL model harvested datasets (ZSL-and-Harvested-DSD-Tweets + ZSL-and-Harvested-DSD-from-Less-Confident-Tweets) only. We see that sample size for all the labels generally increased and the samples reflect almost the same

Dataset	Constituent datasets	Sample size
Final-DSD-Clinician-Tweets	DSD-Clinician-Tweets-Original-Train + ZSL-and-Harvested-DSD-Tweets + ZSL-and-Harvested-DSD-from-Less-Confident-Tweets	$(377 + 2491 + 1699) = 4567$
Final-DPD-Human-Tweets	Final-DSD-Clinician-Tweets which are not in DPD-Human test set + DPD-Human trainset which are not in Final-DSD-Clinician-Tweets + Equal number of NoED tweets from DSD-Harvest-Candidates	$(2743 + 1997) \times 2 = 9480$

Table 5.14: Datasets in step 5.

Model	Train dataset	Comment
Final-DSD-Clinician	Final-DSD-Clinician-Tweets	DSD model at SSL Step 5
Final-DPD-Human	Final-DPD-Human-Tweets	DPD model at SSL step 5

Table 5.15: Model details in step 5.

label distribution as our DSD-Clinician-Tweets-Original-Train dataset. Interestingly, data harvesting increases the sample size of “Feelings of Worthlessness” and “Suicidal thoughts” while still maintains the distribution of our original clinician annotated dataset (DSD-Clinician-Tweets-Original-Train) (Figure 5.6).

We also report the top-10 bi-grams for each of the symptom for our Final-DSD-Clinician-Tweets dataset in Table 5.19. In that table, we report the salient bi-grams highlighted, i.e., the bi-grams exclusive to each of the symptom. We see that top bi-grams and the salient ones convey the concepts of each symptoms.

Step 6: Combating Low Accuracy for Less Populated Labels

Here we attempt to combat the low accuracy for the labels which have very small sample size. In these cases, we analyze the co-occurrence of those labels with other labels through an associative rule mining (Apriori) algorithm [2]. Our idea is to use

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.00	0.00	0.00	5
Low mood	0.57	0.96	0.71	26
Change in sleep pattern	0.68	0.87	0.76	15
Fatigue	1.00	0.17	0.29	6
Weight change	1.00	0.75	0.86	4
Feelings of worthlessness	0.35	0.76	0.48	38
Indecisiveness	0.00	0.00	0.00	11
Agitation	0.62	0.77	0.69	66
Retardation	0.00	0.00	0.00	12
Suicidal thoughts	0.64	0.82	0.72	22
Macro avg	0.49	0.51	0.45	205
Weighted avg	0.51	0.68	0.56	205

Table 5.16: Final-DSD-Clinician model accuracy in step 5.

Precision	Recall	F1-score	Support
0.83	0.97	0.89	227

Table 5.17: Final-DPD-Human model accuracy in step 5.

Step	Model	Macro-F1	Weighted-F1	F1
1	DSD	0.31	0.51	-
1	DPD	-	-	0.87
2	DSD	0.32	0.52	-
3a	DSD	0.28	0.36	-
3b	DSD	0.35	0.56	-
Final	DSD	0.45	0.56	-
Final	DPD	-	-	0.89

Table 5.18: Summary of accuracy improvements (DSD and DPD correspond to DSD-Clinician and DPD-Human models respectively).

Depression-Symptoms	**Bi-grams**
Anhedonia	**want go**, **dont care**, **go work**, **motivation anything**, want die, **want live**, **go away**, **im done**, **tired bored**, **getting bed**
Low Mood	feel like, **want cry**, depression anxiety, feeling like, mental illness, want die, like shit, **want someone**, **feel alone**, **feels like**
Change in Sleep Pattern	**want sleep**, **go sleep**, im tired, **hours sleep**, **fall asleep**, **cant sleep**, **need sleep**, **back sleep**, **could sleep**, **going sleep**
Fatigue	im tired, **f*cking tired**, **physically mentally**, **tired everything**, **tired tired**, feel tired, **im f*cking**, **need break**, **tired yall**, **sad tired**
Weight Change	eating disorder, fat fat, **stop eating**, feel like, **keep eating**, **im gonna**, **lose weight**, **eating disorders**, **fat body**, wish could
Feelings of Worthlessness	feel like, like shit, feeling like, fat fat, wish could, f*cking hate, **good enough**, **ibs hate**, **hate ibs**, **makes feel**
Indecisiveness	**cant even**, **even know**, **says better**, **thoughts brain**, **seems like**, feel like, better dead, **assistant remember**, **remember things**, **time like**
Agitation	feel like, mental illness, f*ck f*ck, depression anxiety, **f*ck life**, f*cking hate, fat fat, **panic attacks**, **every time**, **hate body**
Retardation	feel like, **lay bed**, **ever get**, **committed bettering**, **sleepy kind**, im tired, **one moods**, **talking going**, **well mind**, **motherf*ckers prove**
Suicidal thoughts	want die, feel like, wanna die, **want kill**, **want cut**, **f*cking die**, better dead, **self harm**, **hope die**, want f*cking

Table 5.19: Top-10 bi-grams for each symptoms for Final-DSD-Clinician-Tweets dataset with the ones bolded occur exclusively to each symptoms.

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.03	0.80	0.06	5
Low mood	0.59	0.92	0.72	26
Change in sleep pattern	0.71	1.00	0.83	15
Fatigue	0.04	0.83	0.08	6
Weight change	1.00	0.50	0.67	4
Feelings of worthlessness	0.34	0.79	0.47	38
Indecisiveness	0.09	1.00	0.16	11
Agitation	0.61	0.76	0.68	66
Retardation	0.07	0.75	0.12	12
Suicidal thoughts	0.72	0.82	0.77	22
Macro avg	0.42	0.82	0.45	205
Weighted avg	0.49	0.82	0.57	205

Table 5.20: Final-DSD-Clinician model with applied label association rules accuracy in step 6.

Comment	Precision	Recall	F1-score	Support
Anhedonia	0.00	0.00	0.00	5
Low mood	0.52	0.96	0.68	26
Change in sleep pattern	0.71	1.00	0.83	15
Fatigue	1.00	0.17	0.29	6
Weight change	1.00	0.75	0.8	4
Feelings of worthlessness	0.32	0.82	0.46	38
Indecisiveness	0.00	0.00	0.00	11
Agitation	0.64	0.76	0.69	66
Retardation	0.00	0.00	0.00	12
Suicidal thoughts	0.60	0.82	0.69	22
Macro avg	0.48	0.53	0.45	205
Weighted avg	0.50	0.70	0.56	205

Table 5.21: DSD-Clinician model trained on IJCAI-2017-Unlabelled and all the harvested dataset.

significant co-occurring labels and artificially predict one label if the other occurs. For that, we analyze a small human annotated train dataset (DSD-Clinician-Tweets-Original-Train). However, since the support and confidence for association rules are not significant due to the small sample size, we consider all the "strong" rules with non-zero support and confidence score for those labels. The rules we consider have the form: (strong-label $\rightarrow$ weak-label), where the weak label (such as Anhedonia, Fatigue, Indecisiveness and Retardation) means, the labels for which our model achieves either 0 F1 score or very low recall). These are the candidate labels for which we would like to have increased accuracy. On the other hand strong labels are those for which we have at least a good recall. By emphasizing high recall, we intend to not miss a depression symptom from being detected by our model. All the extracted strong rules are reported in Appendix A.9. When we compare the sample distribution for Apriori based harvested data and plain harvested data, we see for the least populated class we have more samples (Figure 5.7). This makes the classification task more sensitive towards the weak labels. However, with this method, we do not achieve better Macro-F1 score compared to our Final-DSD-Clinician model (Table 5.20).

Stopping Criteria for SSL:

The following two observations lead us to stop the SSL:

1. Our DTR consists of total 6,077 samples and we have finally harvested 4,567 samples, so for $(6,077 - 4,567) = 1,510$ samples neither ZSL nor any version of DSD models have any predictions. We exhausted all our depression candidate tweets from all sources we have, therefore, we do not have any more depression symptoms candidate tweets for moving on with SSL.

2. We have another very noisy dataset, called IJCAI-2017-Unlabelled [133], where we have tweets from possible depressed users, i.e., their self-disclosure contains the character string "depress" but it is not verified whether those are genuine self-disclosures of depression. Using our Final-DSD-Clinician model we harvest $\approx 22K$ depression symptoms tweets from $\approx 0.4M$ de-

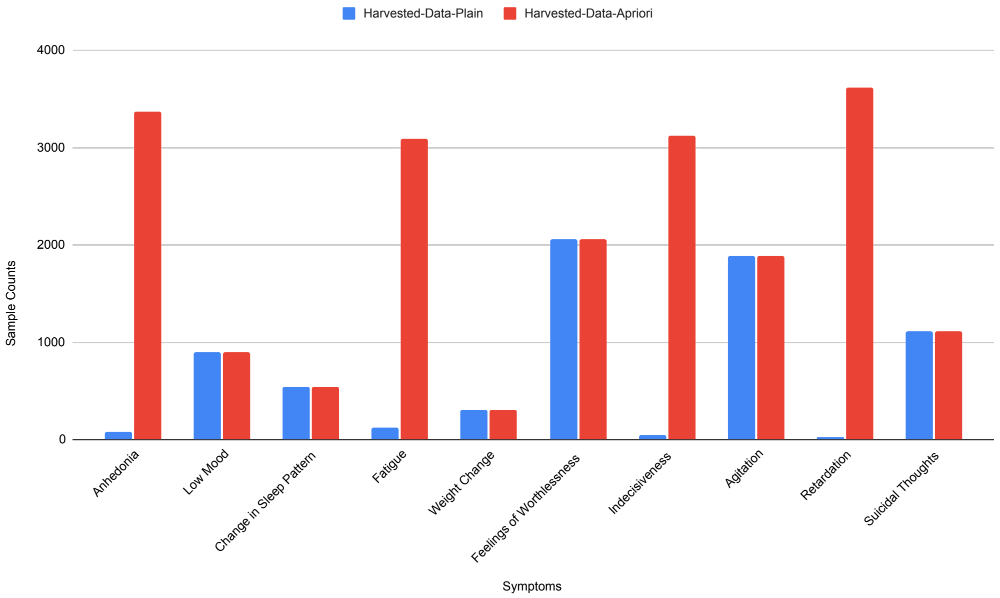


Figure 5.7: Sample distribution in Apriori harvested dataset vs plain harvested dataset.

pression candidate tweets identified by Final-DPD-Human model from that dataset. We then retrain the Final-DSD-Clinician model on all the samples previously we harvested combined with the newly harvested $\approx 22K$ tweets, which results in a total of $\approx 26K$ tweets (≈ 6 times larger than the samples DSD-Final-model was trained on). However, we do not see any significant accuracy increase, so we do not proceed (Table 5.21).

5.6 Results Analysis

Here we analyze the efficacy of our semi-supervised learning framework on three dimensions, as follows:

5.6.1 Dataset Size Increase

Through the data harvesting process, we are able to increase our initial clinician annotated 377 samples to 4567 samples, which is 12 times bigger than our initial dataset. In addition, we have access to a external organization collected dataset (i.e., D2S dataset), for which we could access around ≈ 1600 samples. Our final dataset is more than double the size of that dataset.

5.6.2 Accuracy Improvement

Our Final-DSD-Clinician model has Macro-F1 score of 45% which is 14% more than that of our initial model and Weighted-F1 score increased by 5% from 51% to 56% (Table 5.18). The substantial gain in Macro-F1 score indicates the efficacy of our data harvesting in increasing F1 scores for all the labels. We also find that the combination of DSD-Clinician-1 and ZSL models in step 3a helps achieve more accuracy than individually; specifically, using only ZSL harvested data for training is not very ideal. Macro-F1 has slow growth and does not increase after step 3b. We also find that the combined harvesting process on D2S sampled helped us achieve further accuracy in a few classes for which D2S had more samples, such as "Fatigue," "Weight Change" and "Suicidal Thoughts."

5.6.3 Linguistic Components Distribution

In Table 5.19, we see that our harvested dataset contains important clues of depression symptoms. Interestingly, there are some bi-grams, such as "feel like" occurs in most of the labels; this signifies the frequent usage of that bi-gram in various language based expressions of depression symptoms. This also shows a pattern of how people describe their depression.

5.6.4 Sample Distribution

Compared with the original clinician annotated dataset distribution (Figure 5.6), we see similar trends in our harvested dataset, i.e., in Final-DSD-Clinician-Tweets. However, instead of "Agitation" we have some more samples on "Feeling of Worthlessness," although those are not surpassed by "Suicidal Thoughts" as in D2S dataset. "Suicidal Thoughts" samples have also strong presence which is the result of integrating D2S dataset in our harvesting process. Since the majority of our samples are coming from self-disclosing users tweets, and we apply our DSD model learned on the clinician annotated portion of that dataset to the D2S dataset to harvest tweets, our final harvested dataset reflects mainly the distribution of symptoms from the self-disclosing depressed users and insights from clinical experts. However, D2S has some impact which results in more samples in few labels of the final harvested dataset.

5.6.5 Data Harvesting in the Wild

We use our final model on a bigger set of very loosely related data, i.e., IJCAI-2017-Unlabelled, but we do not see any increase of accuracy, which suggests that harvesting from irrelevant data is of no use.

5.7 Limitations

1. Our overall dataset size is still small, i.e., for some labels we have very small amount of data both for training and testing.

2. We do not attempt to artificially increase sample amount for the small populated labels.

3. To start with, we use a small expert (i.e., clinician) annotated dataset because human annotation is expensive.

4. We haven't explored stratified sample generation using state-of-the-art generative models.

5. Our SSL process depends on the iteratively learned model's efficacy for labelling new samples. So it is possible that the error this model makes can be propagated further in next iteration; we tried to tackle it to some extent through integrating a ZSL model in this process. However, one of the solution to this problem is to employ Active Learning (AL) based on human in the loop labelling, unfortunately, it is almost impossible to acquire expert human annotation in an iterative basis as it is very expensive.

5.8 Conclusion

We have described a Semi-supervised Learning (SSL) framework more specifically semi-supervised self-training (Appendix A.4) for gathering depression symptoms samples in-situ from one of the largest benchmark datasets of self-disclosing users' Twitter timeline. We articulate each step of our data harvesting process and model re-training process. We also discuss our integration of Zero-Shot Learning model in this process and its contribution. We show that each of these steps provides moderate to significant accuracy gains. We discuss the effect of harvesting from the samples of an externally curated dataset, and we also try harvesting samples in the wild, i.e., a large noisy dataset with our Final-DSD-Clinician model. In the former case we find good improvement in Macro-F1 score. In the latter, we do not see any improvements indicating that there is room for further progress to improve accuracy in those samples. Finally, we discuss the effect of our SSL process for curating small but distributionally relevant samples through both sample distribution and bi-gram distribution for all the labels.

Chapter 6

Deep Temporal Modelling of User-Level Clinical Depression Through Text

In this chapter, we use learned Depression Symptoms Detection (DSD) model as described in the last chapter and develop a deep learning based depression detection model to detect user-level clinical depression through their temporal social media posts. This chapter provides an insight on the strengths and weaknesses of the underlying DSD model to extract clinically relevant features, e.g., depression scores and their consequent temporal patterns, as well as user posting activity patterns, i.e., quantifying "no activity" or "silence" of a user. To evaluate the efficacy of these extracted features, we create three kinds of datasets and a test set from the two existing well-known benchmark datasets for user-level depression detection task. Later, we provide accuracy measures through single features, baseline features and feature ablation tests, in several temporal granularity, data distributions and clinical depression detection related settings to draw a complete picture on the impact of these features across our created datasets. We show that, in general only semantic representation based models perform well. However, clinical features may enhance overall performance very slightly provided the training and testing distribution is same and there is more data in a user's timeline. Further, we show that the predictive capability of depression score increases significantly while used in a more sensitive clinical depression detection settings.

6.1 Motivation

Most of the earlier studies in the area of user-level depression modelling through social media posts do not attempt to align with the clinical framework (Chapter 2). By clinical framework, we mean conforming to the definition of clinical depression as defined in DSM-5[1], i.e., looking for signs of depression in at least a two-week episode of a user. Developing such a model is very challenging because it requires a Depression Symptoms Detection (DSD) model and a framework to calculate depression scores over the temporal episodes in a user's social media timeline. In this work, we mainly focus on using our learned DSD model and clinical insights for depression detection for extracting depression scores. We subsequently represent a user's timeline as a temporal series of depression scores then use that representation for our deep Temporal User-level Clinical Depression Detection (TUD) model.

According to earlier research, social media posting activity patterns and language specific clues are very important for user-level depression modelling. Most current research has focused on these features in a non-temporal manner, i.e., on digests of tweets, meaning, taking all the tweets of a user's timeline and concatenate them together to represent that user, where temporal sequence of these tweets were not considered [18], [20], [27], [96], [155]. Very recently, a very closely related work was carried out by Nguyen et al. [96], who inferred the presence of depression symptoms from individual Reddit posts of a user. They extracted summaries of symptoms from an arbitrary number of posts through different kernel sizes of a CNN classifier and use those as non-temporal feature representations for user-level depression detection. Their depression presence calculation is based on looking for hand crafted text patterns of depression symptoms in a Reddit post. Relatively few studies have considered temporal modelling; but showed a different focus from depression detection, e.g., finding correlation between depression score of the patients from depression rating scales and their underlying mood patterns through social media text [68] or tracking change of the same before the date of depression diagnosis [117]. Recently, [158] proposed a multi-modal social media depression

[1] https://www.psychiatry.org/psychiatrists/practice/dsm/feedback-and-questions/frequently-asked-questions

detection algorithm which uses a hierarchical attention layer that leverages each tweet to learn word level and tweet level compositions. The main criticism of most of these studies is about the value of extracted features: they fail to follow clinical depression modelling criteria (Appendix A.3) and are primarily based on topical and lexical representations which are not as clinically useful as the clinical representations, i.e., depression scores. In addition temporal patterns analysis is missing.

Unlike earlier research, we extract depression scores for each of the two-week depressive episodes in a user's timeline and provide it to the temporal deep learning model, thus enabling the consideration of temporal modelling for user-level clinical depression. We also integrate user posting activity patterns through the proportion of the number of days they have posting activities out of all the days in an episode. This helps us distinguish between an episode without any signs of depression and the same period with no activity.

Earlier research was also not concerned about varying levels of granularity in a user's timeline. In our approach, we provide our model with a sliding two-week time window of all possible depressive episodes with various sliding lengths over a user's social media timeline, e.g., sliding lengths of 1, 7 and 14 days.

In addition, and absent in the earlier research, we consider two different kinds of important depression modelling strategies: one follows strictly the clinical definition of depression, i.e., there must be social media posts that carry signs of either "Anhedonia" or "Low mood" in an episode to qualify it as an episode of depression; and the other does not.

Depression scoring depends on the thresholds used by the clinicians for determining whether a depression symptom is expressed either "not at all," "for several days," "more than half of the days," or "nearly everyday," we experiment with a more sensitive threshold, that qualifies an episode to be expressing depression even when a user has exhibited a symptom for at least a day in that episode.

Therefore, the main motivation of this work comes from user level clinical depression modelling, which means, following clinical criteria of depression detection as laid out in DSM-5 and in clinical practice [29] (Appendix A.3).

6.2 Methodology

We begin with an extensive analysis of our datasets (Chapter 2). First we report distributions of different user specific statistics related to social media usage behavior, demography and linguistic components analysis based on a well-known psycholinguistic lexicon named, LIWC [106]. Next, we describe different clinical features based on depression scores and social media usage behavior of the users and how we extract them from our datasets. Later, we describe these feature distributions across our datasets. We then describe our deep learning model followed by the experimental setups, where we describe our sets of feature-ablated models and single-feature models compared to the all-feature model and relevant baselines.

We experiment with three types of depressive episode analysis, starting from most granular to least granular. To do this, we slide a two-week temporal window in a user's time line from their earliest post in the history to the latest. We experiment with various slide length = (1, 7, 14). Slide length=1 provides us with the most granular temporal analysis to slide length=14 which is the least granular settings of the same. We keep the temporal window as two weeks to conform with the DSM-5 criteria of depression detection which defines depressive episodes to be of two weeks long. Moreover, it is found that temporal mood patterns are best captured through a two-week time window [68], and weekly windows are better than per-day analysis [117].

We experiment with two kinds of clinical depression detection settings–one strictly follows the clinical definition of depression and the other does not. We also experiment with two kinds of clinical analyses based on two different depression scoring strategies- one reflecting traditional clinical scoring approach, another reflecting more sensitive approach for depression detection. We create three main datasets for training purpose and separate a portion of each for testing the performance of the model. We also create a separate test set from one of the datasets which is annotated for ongoing depressed users and then evaluate all the models in that set.

Finally, we provide detailed analysis on how different clinical features contribute to the user-level depression detection task in each of those datasets across various level of granularity and clinical settings.

6.3 Datasets

We have created balanced data subsets from the CLPsych-2015 and IJCAI-2017 datasets (Chapter 2). Both of these datasets are from Twitter users who self-disclosed their diagnosis of depression through a self-disclosing statement. In both of these datasets, depressed users are identified from Twitter users' self-disclosure and control users are the users who do not have such disclosures. We use balanced depressed and control subsets of users for our experiments, as it is found to be the most effective strategy to build robust user-level depression detection model by Shen et al. [133], the curators of the largest benchmark dataset (IJCAI-2017 dataset) for the same task.

CLPsych-2015 users have markedly longer tweets history compared to IJCAI-2017 users. Moreover, IJCAI-2017 users have data preceding only one month of their self-disclosure. So analyzing IJCAI-2017 data in contrast to CLPsych-2015 provides a clear idea whether recency of self-disclosure has any effect on temporal user-level depression detection. In addition, our experiments are heavily based on social media posts from Twitter instead of Reddit or any other depression forums alike. The reason is that, we would like to use an unbiased representative of social media text, as opposed to using the datasets which have strong self-reporting bias such as depression forums (Chapter 2).

6.3.1 Experiment Datasets Creation

We run experiments on three datasets. As described earlier, these datasets are extracted from two publicly available datasets: CLPsych-2015 and IJCAI-2017, which are similar to most of the datasets previously reported: they used public social media posts from self-disclosing users (i.e., Twitter users) for their depression condition. We describe the curation of these datasets as follows:

CLPsych-2015-Users Dataset

CLPsych-2015-Users dataset is a balanced subset of the CLPsych-2015 dataset (Chapter 2, Table 2.2). We ensure each user has minimum 50 posts and 30 days of Twitter history. This dataset does not include any self-disclosing statements. The original dataset from Twitter was created from users with the disclosure statement “I was just diagnosed with depression”. Further, the original dataset curators employed human annotators to verify the authenticity of these self-disclosing statements for most of the users in that dataset. In addition, for a control population, random users were selected without such disclosing statements. The timeline for this dataset collection is in between the years 2008 and 2013.

IJCAI-2017-Users Dataset

We use a subset of the IJCAI-2017 dataset (Chapter 2, Table 2.2) with users who have minimum 50 posts and 30 days of Twitter history, to train our temporal deep learning model with enough data. We ensure that our datasets do not contain the users whose tweets were used for creating training datasets for DSD and DPD tasks (Chapter 5). Note further, this is a multi-lingual dataset with users producing Tweets in different languages. To initially avoid the need for multi-lingual analysis, we discard user records which have more than 20% non-English tweets. Even with this filter, we still find close to 1000 users. This dataset does include the self-disclosure statements from the users. For this dataset, the self-disclosure looks like the following text: “I (am/was/have been) diagnosed with depression.” Many of these disclosures also include the exact time of such diagnosis. Control users were identified based on the Twitter users who do not have any tweets with the character string “depress.” Because the Twitter API could return a huge number of tweets, the curators of this dataset restricted their collection of control tweets from the month of December, 2016. The timeline for collecting depressed users is in between the years 2009 and 2016. Note further that this dataset contains the most recent one month of Tweets from the disclosure for depressed users; for control, it is just the recent one months of posts.

Since for this dataset we have self-disclosure statement and the timeline of depression diagnosis, by analyzing each user's self-disclosure, we identify genuine users and create two types of user datasets based on the recency of their diagnosis:

1. **IJCAI-2017-Ongoing-Users**: these users declared that their depression diagnosis is recent.

2. **IJCAI-2017-Today-Users**: these users declared that they were diagnosed with depression exactly at the day of the disclosure.

We identify genuine users based on the criteria that the user is talking about their own depression and not using sarcasm, lyrics or any other text that does not directly indicate the user's depression diagnosis. Whenever a user expressed any doubt about their depression diagnosis, we also consider them as not genuine. Details on the annotation task for finding out users with current/ongoing depression is provided in the work by McAvaney et al. [69]. We find only 20% of our IJCAI-2017 users to be genuine ongoing depression candidate users. Moreover, we find only 9% of those users who disclosed their exact date of depression diagnosis.

Mixed-Users Dataset

Mixed-Users dataset is a derived dataset created by combining CLPsych-2015-Users and IJCAI-2017-Ongoing-Users datasets described earlier. This dataset is created to see whether combining both datasets help in depression detection in our test set described in the next sections. We do not separately report the feature and linguistic analysis for this dataset because it is the aggregate of our two main training datasets, i.e., CLPsych-2015-Users and IJCAI-2017-Ongoing-Users.

The choice of minimum number of posts, days and proportion of non-English tweets to curate the above datasets is largely influenced by an earlier research which curated one of the very well-known benchmark datasets for user-level depression detection through Twitter timeline [20]. The authors of that paper used users with maximum 25% non-English tweets and minimum 25 posts, where, we adopted a more strict strategy for non-English tweets proportion (i.e., 20%) and minimum number of posts (i.e., 50) to facilitate more data per user for our deep learning model and thereby learning better models.

6.3.2 Dataset Statistics

In this section, we provide user-level social media behavior statistics and their demographic profile. We also provide linguistic component distribution analysis for the above mentioned datasets. For this linguistic analysis, we use a well-known psycholinguistic lexicon named, LIWC [106], which is popularly used in user-level mood fluctuation, emotion and sentiment analysis in temporal social media data [27], [36].

We do not have demographic information for the IJCAI-2017 dataset, nor do we have any information of the geographic location of the users. For the CLPsych-2015 dataset only, we have demographic information available.

User Specific Statistics:

In our dataset statistics tables, we provide the following user specific statistics:

1. **#Users:** Total number of users.

2. **Avg. Frequency. of Posting (AFP):** Time difference between two consecutive user activities, here activity means tweet post by a user. AFP is the average of these differences. The lower the number, the higher the activity or posting frequency of the user.

3. **Fluctuation of Posting Frequency (FPF):** This is standard deviation of AFP, which means, how much irregular a user's posting frequency is.

4. **#Tweets:** Total number of tweets in a user's profile.

5. **#Proper-Tweets:** Total number of proper tweets, i.e., tweets after preprocessing in a user's timeline.

6. **#Days:** Total number of days a user has Twitter history.

7. **Age:** Age of a user. Only available for CLPsych-2015 dataset, inferred by a third party machine learning model for detecting age [20].

8. **Gender:** Gender of a user. Only available for CLPsych-2015 dataset, inferred by a third party machine learning model for detecting gender [20].

9. **Avg. Tweets Length:** We report the average length of Tweets, i.e., average number of tokens in all tweets in a user's timeline.

10. **Avg. Sents :** We also report the average number of sentences in a tweet. To calculate this we simply split a tweet based on period/question mark/exclamation.

For all these statistics, we report the average and standard deviations across depressed and control population except #Users (Tables 6.2 and 6.3). We use Welch's two-tailed unpaired t-test to find statistical significance among the means of these features across depressed vs. control population (statistically significant means p-value < 0.05). Welch's unpaired t-test is widely used for comparing means between two populations [39].

We observe that IJCAI-2017-Ongoing/Today-Users datasets are smaller than CLPsych-2015-Users: the average number of posts in CLPsych-2015-Users is higher in their timeline compared to IJCAI-2017-Ongoing/Today-Users. However, average Tweet length and average number of sentences are the same across these datasets (Table 6.1).

In IJCAI-2017-Ongoing/Today-Users, number of posts for control population is higher than the depressed population, however, there is no such difference for the same in CLPsych-2015-Users. #Tweets and #Proper-Tweets are significantly higher in control than depressed in IJCAI-2017-Ongoing/Today-Users, in CLPsych-2015-Users there is no such difference in the same. In all three datasets, avg. tweets length is significantly higher in depressed population compared to control population.

For both depressed and control CLPsych-2015-Users, we find there are more females than males and most of them are young adults (Table 6.2), with the control population significantly older than the depressed population, by 4 years.

The Twitter timeline of CLPsych-2015-Users is significantly longer than IJCAI-2017-Ongoing/Test-Users. Moreover, in the CLPsych-2015-Users dataset, control users have significantly longer timelines than depressed users. For the IJCAI-2017-Ongoing/Test-Users, they are same, because, IJCAI-2017-Ongoing/Test-Users datasets are collected for a window of 1 month only. In the CLPsych-2015-Users, depressed users post more frequently and show less fluctuation than control users; it's just the opposite for IJCAI-2017-Ongoing/Test Users. However, in both datasets, both control and depressed users are very active which is reflected through their AFP which is less than two days (Table 6.3).

Linguistic Components Distribution:

Here we provide the linguistic component analysis with the help of LIWC. We create a digest of all Twitter posts from both depressed and control users' Twitter timelines. Later, we apply LIWC on these digests. For a given digest, LIWC finds the proportion percentage of lexicon items under each lexicon components. We call this proportion percentage, "Lexicon Component Intensity (LCI)". We follow the steps provided below to perform our linguistic component distribution analysis:

1. We find the deviations between the LCIs (we call LCI_{dev}) for depression (LCI_d) and control (LCI_c) population for each dataset. All the positive values (or deviations) mean those components have high LCI in depressed population compared to the control population; negative means vice-versa, and zero means equal (Equation 6.1).

2. Finally, we report LCI_{dev} for all the common LIWC components where $LCI_{dev} > 0$ for depressed population and control population for all three datasets. For the control population, we make negative deviation positive. We then report the average and standard deviation of those in the Tables 6.4 and 6.5 in descending order of the average LCI_{dev} across all datasets (Equations 6.1, 6.2 and 6.3).

   This analysis provides us with the LIWC components that are mostly expressed in depressed population compared to control population and vice-versa.

Datasets	**Sample-Size (D)**	**Sample-Size (C)**	**#Tweets (D)**	**#Tweets (C)**	**Avg. Tweets Length(D)**	**Avg. Tweets Length(C)**	**Avg. Sents(D)**	**Avg. Sents(C)**	**#Proper-Tweets (D)**	**#Proper-Tweets (C)**
CLPsych-2015-Users	273	264	$1067(\pm745)$	$1044(\pm793)$	$13.2(\pm3.3)^*$	$12.4(\pm3.4)$	$1.5(\pm0.4)$	$1.6(\pm0.4)^*$	$1020(\pm724)$	$999(\pm770)$
IJCAI-2017-Ongoing-Users	196	196	$187(\pm160)$	$425(\pm464)^*$	$13.8(\pm3.1)^*$	$12.5(\pm3.4)$	$1.5(\pm0.3)^*$	$1.4(\pm0.4)$	$167(\pm142)$	$372(\pm403)^*$
IJCAI-2017-Today-Users	18	18	$155(\pm131)$	$362(\pm310)^*$	$14.6(\pm3.8)^*$	$12.1(\pm2.6)$	$1.6(\pm0.5)$	$1.5(\pm0.3)$	$142(\pm120)$	$292(\pm240)^*$

Table 6.1: Dataset statistics for all datasets (* indicates significantly higher with p-value < 0.05 in Welch's two-tailed unpaired t-test).

Class	**#Male**	**#Female**	**Age (Mean)**
Control	74	190	$25.2(\pm 6.33)^*$
Depression	54	219	$21.6(\pm 4.92)$

Table 6.2: CLPsych-2015 demographic statistics (* indicates significantly higher with p-value < 0.05 in Welch's two-tailed unpaired t-test.)

$$LCI_{dev} = |LCI_d - LCI_c| \tag{6.1}$$

$$LCI_{dev-avg} = \mu(LCI_{dev}) \tag{6.2}$$

$$LCI_{dev-std} = \sigma(LCI_{dev}) \tag{6.3}$$

These tables show that language used by depressed population has more use of personal pronouns, negative emotion and anxiety related words compared to control population (bold items in Table 6.4). This observation aligns with an earlier research [27].

6.4 Data Preprocessing

6.4.1 Tweets Preprocessing

We use the same data preprocessing for the tweets as described in Chapter 3, Section 3.4.

All the tweets which are excluded after preprocessing are counted towards user posting activity but they don't carry signs of depression. "No posting activity" or absence is represented differently than absence of depression, so that our modelling can distinguish between these two.

6.4.2 User Level Filtering

Our datasets are derived from two widely used benchmark datasets used by numerous established studies [18], [53], [102], [133], [152] without any user filtering. One reason could be that, the original data curators already verified the users

Datasets	**#HistoryDays (D)**	**#HistoryDays (C)**	**AFP (D)**	**AFP (C)**	**FPF (D)**	**FPF (C)**
CLPsych-2015-Users	$366(\pm355)$	$495(\pm479)^*$	$0.62(\pm1.20)$	$1.139(\pm2.66)^*$	$3.18(\pm5.93)$	$5.04(\pm12.05)^*$
IJCAI-2017-Ongoing-Users	$30(\pm0)$	$30(\pm0.2)$	$0.09(\pm0.08)^*$	$0.04(\pm0.06)$	$0.35(\pm0.28)^*$	$0.17(\pm0.19)$
IJCAI-2017-Today-Users	$30(\pm0)$	$30(\pm0)$	$0.13(\pm0.11)^*$	$0.05(\pm0.11)$	$0.49(\pm0.44)$	$0.23(\pm0.36)$

Table 6.3: User posting related statistics for all datasets (* indicates significantly higher with p-value < 0.05 in Welch's two-tailed unpaired t-test).

LIWC Components	**CLPsych-2015**	**IJCAI-2017-Ongoing-Users**	**IJCAI-2017-Today-Users**	$\mathrm{LCI}_{\mathbf{dev-avg}}$	$\mathrm{LCI}_{\mathbf{dev-std.dev.}}$
Authentic	13.74	7.40	7.61	9.58	3.60
Linguistic	8.1	3.50	0.29	3.96	3.93
function	7.61	2.97	0.97	3.85	3.41
Dic	6.17	3.74	0.97	3.63	2.60
i	2.92	1.16	1.65	1.91	0.91
pronoun	3.93	1.46	0.23	1.87	1.88
ppron	3.27	1.24	0.39	1.63	1.48
Cognition	1.9	1.55	0.30	1.25	0.84
cogproc	1.68	1.45	0.54	1.22	0.60
auxverb	2.08	0.50	0.24	0.94	1.00
conj	1.37	0.81	0.46	0.88	0.46
Period	0.07	1.36	1.14	0.86	0.69
adverb	1.37	0.43	0.51	0.77	0.52
emotion	0.57	0.51	0.60	0.56	0.05
focuspresent	1.14	0.12	0.20	0.49	0.57
tone_neg	0.49	0.44	0.33	0.42	0.08
emo_neg	0.39	0.37	0.32	0.36	0.04
health	0.34	0.65	0.05	0.35	0.30
insight	0.16	0.15	0.48	0.26	0.19
cause	0.19	0.27	0.23	0.23	0.04
emo_pos	0.18	0.13	0.24	0.18	0.06
emo_sad	0.11	0.07	0.20	0.13	0.07
certitude	0.16	0.05	0.17	0.13	0.07
illness	0.07	0.28	0.02	0.12	0.14
mental	0.16	0.14	0.03	0.11	0.07
family	0.06	0.16	0.10	0.11	0.05
want	0.13	0.14	0.04	0.10	0.06
feeling	0.12	0.08	0.11	0.10	0.02
assent	0.06	0.08	0.03	0.06	0.03
friend	0.05	0.04	0.08	0.06	0.02
sexual	0.06	0.07	0.01	0.05	0.03
emo_anx	0.08	0.04	0.01	0.04	0.04
lack	0.02	0.02	0.01	0.02	0.01

Table 6.4: Depression deviations for all three datasets (bolded components are usually high in depressed population than control indicated in earlier research).

LIWC Components	**CLPsych-2015-Users**	**IJCAI-2017-Ongoing-Users**	**IJCAI-2017-Today-Users**	$\mathbf{LCI_{dev-avg}}$	$\mathbf{LCI_{dev-std.dev.}}$
Analytic	27.89	8.93	2.14	12.99	13.35
Clout	17.88	5.89	9.66	11.14	6.13
Tone	12.63	12.41	7.35	10.80	2.99
Lifestyle	1.22	0.34	0.14	0.57	0.57
Perception	0.71	0.46	0.50	0.56	0.13
netspeak	0.13	0.31	0.91	0.45	0.41
Drives	0.81	0.03	0.42	0.42	0.39
space	0.53	0.24	0.38	0.38	0.15
Conversation	0.04	0.24	0.79	0.36	0.39
food	0.29	0.07	0.45	0.27	0.19
leisure	0.35	0.24	0.07	0.22	0.14
tone_pos	0.26	0.29	0.11	0.22	0.10
Culture	0.31	0.26	0.02	0.20	0.16
power	0.26	0.04	0.28	0.19	0.13
motion	0.17	0.06	0.31	0.18	0.13
we	0.24	0.26	0.01	0.17	0.14
affiliation	0.31	0.04	0.04	0.13	0.16
reward	0.11	0.24	0.01	0.12	0.12
relig	0.00	0.10	0.14	0.08	0.07
politic	0.15	0.04	0.03	0.07	0.07
ethnicity	0.03	0.01	0.03	0.02	0.01

Table 6.5: Control deviations for all three datasets.

through human annotation and analysing the genuineness of their disclosure [20], [133]. In addition, our own Tweets preprocessing and minimum 50 number of posts constraint also removes users with excessive gibberish posts. Finally, we manually reviewed each user's timeline to verify the quality of the users based on the content of their posts, i.e., whether they have at least a few post regarding their struggles related to depression.

6.5 Clinically Relevant Features Extraction

Here we describe how we calculate several clinically relevant features for an episode (i.e., for a two-week time window). Later, we use those to learn temporal patterns using our deep Temporal User-level Clinical Depression Detection (TUD) model.

6.5.1 Depression Score (DS)

One of the major contributions of our research is to employ the DSD model to guide extraction of depression scores for an episode. We extract such depression scores over all such episodes in a user's Twitter timeline and then use TUD model to learn useful temporal patterns of depression.

To enable this feature extraction process we take the following steps:

1. We first sort the posts of a user based on their Twitter post timestamp information, in an ascending order of recency.

2. We then create day-wise chunks of tweets.

3. For each day of tweet chunks we find out Depression Symptoms Expression Vector (DSEV), where $DSEV \in \{0, 1\}^d$, and d = #depression-symptoms. Each indices of this vector corresponds to each of the 10 depression symptoms we are interested in. DSEV is initialized as all 0s at the beginning indicating no symptoms is expressed; then if any of the tweets in the chunk has expressed symptoms, a particular index of DSEV vector is made 1, which signifies corresponding depression symptom is expressed for that day (Algorithm 4).

Algorithm 4: Depression-Symptoms-Expression-Vector Algorithm (DSEVA).

```
**Input:** A day chunk of tweets, $D$, Depression symptoms, $S$
**Output:** $DSEV$
$DSEV \leftarrow \{0\}^{\#symptoms}$ ;
foreach $tweet \in day$ do
    $depSymptsIDs \leftarrow DSD(tweet)$ ;
    foreach $symptIndex \in range(|S|)$ do
        if $symptIndex \in depSymptsIDs$ then
            $DSEV[symptIndex] \leftarrow 1$
        end
    end
end
return $DSEV$ ;
```

4. Later, in the first layer of TUD model, we extract all the DSEVs in an episode, aggregate them and calculate the percentage of days each depression symptoms is expressed (Lines (5-12) in Algorithm 5). We calculate this percentage on the number of days the user has activity, i.e., Twitter posts.

5. In an episode, a user may not have tweets for all of its days. So we also keep track of the days for which a user has no activity (i.e., no tweets), which we call Absence Ratio (AR). This is separately discussed in Section 6.5.3.

6. Finally, Depression Score (DS) is calculated based on the percentage of days for each depression symptoms in an episode. In this calculation we consider "Agitation" and "Retardation" as one symptoms instead of two separate ones to conform with PHQ-9 (for the sake of brevity this is not included in the algorithm). If this is within a predefined range of thresholds[2], as defined in PHQ-9, we assign a corresponding score (or symptomScore) for that symptom in an episode. Aggregating all these scores for all the symptoms provide us with the final depression score for an episode (Lines (5-28) in Algorithm 5).

7. Since to identify clinical depression, a user must have either "Low Mood" or "Anhedonia", we enable our scoring Algorithm 5 (Lines 29-39), so that we

[2] It is to be noted that, the predefined thresholds in PHQ-9 are not concrete, they are roughly described, so we use a clinician's advice to ground those descriptions to numerical values.

Algorithm 5: Depression-Score Algorithm (DSA).

```
Input: An episode, $E$, depression symptoms, $S$, Mode, $M$
Output: Depression Score, $depScore$
$depScore \leftarrow 0$ ;
$symptomScore \leftarrow 0$ ;
$DSEV_{Sum} \leftarrow 0$ ;
$symptomScoreSum \leftarrow 0$ ;
foreach $day \in E$ do
    foreach $tweet \in day$ do
        $DSEV_{sum} \leftarrow DSEV_{sum} + DSEVA(day, S)$
    end
end
foreach $symptIndex \in range(|S|)$ do
    $symptomScoreSum \leftarrow DSEV_{sum}[symptIndex]$
     $percentOfDays \leftarrow (symptomScoreSum/|E|) \times 100$
    if $(percentOfDays \geq 50)$ and (S is ("Anhedonia" or "Low Mood")) then
        $isClinicallyDepressed \leftarrow True$
    end
    if $(percentOfDays \geq 0)$ and $(percentOfDays < 20)$ then
        $symptomScore \leftarrow 0$
    end
    else if $(percentOfDays \geq 20)$ and $(percentOfDays < 50)$ then
        $symptomScore \leftarrow 1$
    end
    else if $(percentOfDays \geq 50$) and $percentOfDays < 85$ then
        $symptomScore \leftarrow 2$
    end
    else if $(percentOfDays \geq 85)$ then
        $symptomScore \leftarrow 3$
    end
    $depScore \leftarrow depScore + symptomScore$ ;
end
if $M$ is "Clinical" then
    if $isClinicallyDepressed$ then
        return $depScore$ ;
    end
    else
        return 0 ;
    end
end
else if $M$ is "Non-Clinical" then
    return $depScore$ ;
end
```

can calculate depression scores by fulfilling the clinical criteria or relaxing it. The former option is called **Clinical Scoring (CS)**, the latter is called **Non-clinical scoring (NCS)**. In Clinical Scoring (CS) criteria, a depression score of 0 is assigned for an episode, if none of the above depression symptoms are expressed in that episode, otherwise, we move on with the depression score calculation. We report our TUD model's performance for both options.

8. We also consider a much more sensitive version of depression scoring. So, instead of considering all the thresholds stated in lines (12-26) in Algorithm 5, we only consider one threshold, i.e., whenever there is a tweet carrying signs of depression, we consider a symptom score (symptomScore) of 1, and an episode will be considered as a minimal depressive episode whenever, it has depression score $(depScore) > 0$, otherwise the episode will not be considered as a depressive episode, we call this **Minimal Depression Expression (MDE)** based Temporal Modelling.

6.5.2 Semantic Information

To create a representation that holds semantic information corresponding to a depressive episode, we first take the average of the sentence embeddings for all the tweets in a day to represent that day, we call this **Day Level Sentence Embedding Average (DLSEA)**. Subsequently, based on this day level semantic representation, we calculate the episode level semantic representation by again taking average embedding for all the DLSEAs in an episode, we call this **Episode Level Sentence Embedding Average (ELSEA)**. We also take all the tweets and the average of their sentence embeddings, we call this **All Tweets Embedding Average (ATEA)**. We use Universal Sentence Encoder (USE) based sentence embedding for all these representations, as USE embedding has been found out to be very effective and compact representation for detecting signs of depression detection (Chapter 3).

6.5.3 User Posting Activity Pattern

As mentioned earlier, we find out posting activity patterns for each episode. For this, we calculate, number of days a user has no activity (or no social media posts in a day) out of all days in an episode, we call this **Absence Ratio (AR)**.

6.5.4 Temporal Depression Patterns

We extract two kinds of temporal depression patterns among all the episodes with user activity. These are (1) Depression Recurrence Frequency and (2) Inertia. Depression Recurrence Frequency has been found to be an important predictor of clinical depression as it is usually highly recurrent in nature [10], and Inertia has been found by early research as an important trait for depressed social media users [61].

To calculate those, we first binarize the temporal series of episodic depression scores. We call this series, **Binarized Temporal Episodes (BTE)**. Through binarization we convert the depression scores to 1 if those correspond to minimal or higher level of depression, otherwise we convert it to 0 (Algorithm 6). Later we take the following steps to calculate **Depression Recurrence Frequency** and **Inertia** scores.

Algorithm 6: Binarized-Temporal-Episodes Algorithm (BTEA).

```
Input: Temporal Episodes, $TE$
Output: Binarized Temporal Episodes, $BTE$
$BTE \leftarrow \{0\}^{|TE|}$ ;
foreach $E \in TE$ do
    if $depLevel(E) \geq$ “$MINIMAL''$ then
        $BTE \leftarrow 1$
    end
end
return $BTE$ ;
```

Depression Recurrence Frequency Score (DRFS):

Depression recurrence means repetition of depressive mood. Here, we track whether a user's depression shows up in a recurring manner. To calculate this, we first compress BTE, or remove consecutive repetitive binary scores from it, we call this series **Compressed Binarized Temporal Episodes (CBTE)**. Later, we find the cyclic pattern “1-0-1” (or a cycle), which means, a user starts with depression, gets better but again falls into depression. We count all such patterns in CBTE and normalize them with the number of items or binary scores in CBTE. We call this score **DRFS** (Algorithms 7, 8 and 9).

Algorithm 7: Compressed-Binarized-Temporal-Episodes Algorithm (CBTEA).

```
Input: Binarized-Temporal-Episodes, $BTE$
Output: Compressed-Binarized-Temporal-Episodes, $CBTE$
$CBTE \leftarrow \emptyset$ ;
$CBTE.insert(BTE[0])$ ;
foreach $TE \in BTE$ do
    if $TE \neq CBTE[|CBTE| - 1]$ then
        $CBTE.insert(TE)$
    end
end
return $CBTE$ ;
```

Algorithm 8: Cycle-Count Algorithm (CCA).

```
Input: Compressed-Binarized-Temporal-Episodes, $CBTE$
Output: Cycle-Count, $CC$
$CC \leftarrow 0$ ;
if $|CBTE| > 2$ then
    foreach $index \in range(|CBTE|$ do
        if $0 < index < (|CBTE| - 1)$ then
            if $CBTE[index] = 0$ then
                if $(CBTE[index - 1] = 1)$ and $(CBTE[index + 1] = 1)$ then
                    $CC \leftarrow CC + 1$
                end
            end
        end
    end
end
return $CC$ ;
```

Inertia Score (IS):

Inertia means the tendency of staying in depressive mood for a while. To calculate this, we take BTE and find how many consecutive episodes have values 1, which means how many consecutive depressive episodes are there in a user’s timeline. We

Algorithm 9: Depression-Recurrence-Frequency-Score Algorithm (DRFSA).

```
Input: Compressed-Binarized-Temporal-Episodes, $CBTE$
Output: Depression-Recurrence-Frequency-Score, $DRFS$
$cycles \leftarrow CCA(CBTE)$ ;
$DRFS \leftarrow (|cycles|/|CBTE|)$ ;
return $DRFS$ ;
```

then normalize this count with the total episode counts of that user. We call this score **Inertia Score (IS)** (Algorithm 10).

Algorithm 10: Inertia-Score Algorithm (ISA).

```
Input: Binarized-Temporal-Episodes, $BTE$
Output: Inertia-Score, $IS$
$consecutivenessCount \leftarrow 0$ ;
foreach $index \in range(|BTE| - 1)$ do
    if $BTE[index] - BTE[index + 1] = 0$ then
        if $BTE[index] \times BTE[index + 1] = 1$ then
            $consecutivenessCount \leftarrow consecutivenessCount + 1$
        end
    end
end
$IS \leftarrow (consecutivenessCount/|BTE|)$ ;
return $IS$ ;
```

6.6 Clinically Relevant Feature Distribution in the Datasets

Here we report the extracted feature distributions, such as depression levels[3], depression score related temporal patterns (i.e., DRFS and IS) and user-activity patterns for our three datasets (i.e., CLPsych-2015-Users, IJCAI-2017-Ongoing-Users and IJCAI-2017-Today-Users). To calculate this distribution, we first determine the proportion of episodes out of all the episodes in a user's Twitter timeline. We then find out the average and standard deviation of these measures for all the users in depressed and control population. These numbers are reported in Tables 6.6, 6.7 and 6.8. We report differences among these features across depression versus control

[3] Depression levels are calculated based on depression score ranges. These ranges are reported in Section A.8 and Table A.2

Depression-Level	**Control**	**Depression**
None	$0.9763(\pm 0.1167)^*$	$0.9258(\pm 0.1885)$
None(MDE)	$0.6701(\pm 0.3405)^*$	$0.4808(\pm 0.3281)$
Minimal	$0.0125(\pm 0.0577)$	$0.03649(\pm 0.0949)^*$
Minimal(MDE)	$0.3299(\pm 0.3405)$	$0.5192(\pm 0.3281)^*$
Mild	$0.0111(\pm 0.0717)$	$0.0375(\pm 0.1519)^*$
Moderate	$0(\pm 0)$	$0.0002(0.0026)$
Moderately-Severe	$0(\pm 0$	$0 \pm 0)$
Severe	$0(\pm 0)$	$0(\pm 0)$
AR	$0.3616(\pm 0.2771)$	$0.3600(\pm 0.2660)$
IS	$0.0211(\pm 0.1099)$	$0.0664(\pm 0.1750)^*$
IS(MDE)	$0.3160(\pm 0.3310)$	$0.4996(\pm 0.3216)^*$
DRFS	$0.0013(\pm 0.0068)$	$0.0039(\pm 0.0105)^*$
DRFS(MDE)	$0.0083(\pm 0.0105)$	$0.0104(\pm 0.0098)^*$

Table 6.6: CLPsych-2015-Users features distribution (* indicates significantly higher with p-value < 0.05 in Welch's two-tailed unpaired t-test).

populations, based on Welch's two-tailed unpaired t-test (statistically significant means p-value < 0.05).

We find that, in CLPsych-2015-Users dataset, depression levels, such as "Minimal" and "Mild" and temporal patterns, such as "IS" and "DRFS" are significantly higher in depressed population compared to control population. Alternatively note that, instances labelled as "None" are significantly higher in the control population than in the depressed population. These distributions are expected according to the earlier research and clinical criteria of depression [27], [61] and Appendix A.3.

In the IJCAI-2017-Ongoing-Users dataset, we note that the depressed population has a significantly higher Absence-Ratio than the control population. However, for all other features and in both IJCAI-2017-Ongoing-Users and IJCAI-2017-Today-Users datasets, we do not see any statistically significant difference.

6.7 Experimental Setup

Figure 6.1 illustrates the diagram of the over-all temporal deep-learning model. The model is provided with day-level aggregate of depression score and semantic representation of tweets based on day-level average embedding. Later, the model based on a flexible settings for different sliding day lengths calculates episode level ag-

Depression-Level	**Control**	**Depression**
None	$0.9271(\pm0.2258)$	$0.9232(\pm0.2114)$
None(MDE)	$0.3328(\pm0.4117)^*$	$0.2290(\pm0.3433)$
Minimal	$0.0396(\pm0.1484)$	$0.0528(\pm0.1553)$
Minimal(MDE)	$0.6671(\pm0.4116)$	$0.7710(\pm0.3433)^*$
Mild	$0.0327(\pm0.1624)$	$0.024(\pm0.1279)$
Moderate	$0.0006(\pm0.0084)$	$0(\pm0)$
Moderately-Severe	$0(\pm0)$	$0(\pm0)$
Severe	$0(\pm0)$	$0(\pm0)$
AR	$0.0827(\pm0.1177)$	$0.1691(\pm0.1585)^*$
IS	$0.0621(\pm0.2063)$	$0.0615(\pm0.1878)$
IS(MDE)	$0.6164(\pm0.3903)$	$0.7113(\pm0.3318)^*$
DRFS	$0.0021(\pm0.015)$	$0.0045(\pm0.020)$
DRFS(MDE)	$0.0048(\pm0.0161)$	$0.0066(\pm0.0186)$

Table 6.7: IJCAI-2017-Ongoing-Users features distribution (* indicates significantly higher with p-value < 0.05 in Welch's two-tailed unpaired t-test).

Depression-Level	**Control**	**Depression**
None	$0.9935(\pm0.0278)$	$1(\pm0)$
None(MDE)	$0.2549(\pm0.3406)$	$0.2974(\pm0.3909)$
Minimal	$0.0065(\pm0.0278)$	$0(\pm0)$
Minimal(MDE)	$0.7451(\pm0.3406)$	$0.7026(\pm0.3909)$
Mild	$0(\pm0)$	$0(\pm0)$
Moderate	$0(\pm0)$	$0(\pm0)$
Moderately-Severe	$0(\pm0)$	$0(\pm0)$
Severe	$0(\pm0)$	$0(\pm0)$
AR	$0.1127(\pm0.1902)$	$0.2374(\pm0.2293)$
IS	$0.0033(\pm0.014)$	$0(\pm0)$
IS(MDE)	$0.6797(\pm0.3248)$	$0.6470(\pm0.3775)$
DRFS	$0(\pm0)$	$0(\pm0)$
DRFS(MDE)	$0.0131(\pm0.0252)$	$0.0065(\pm0.0190)$

Table 6.8: IJCAI-2017-Today-Users features distribution (* indicates significantly higher with p-value < 0.05 in Welch's two-tailed unpaired t-test).

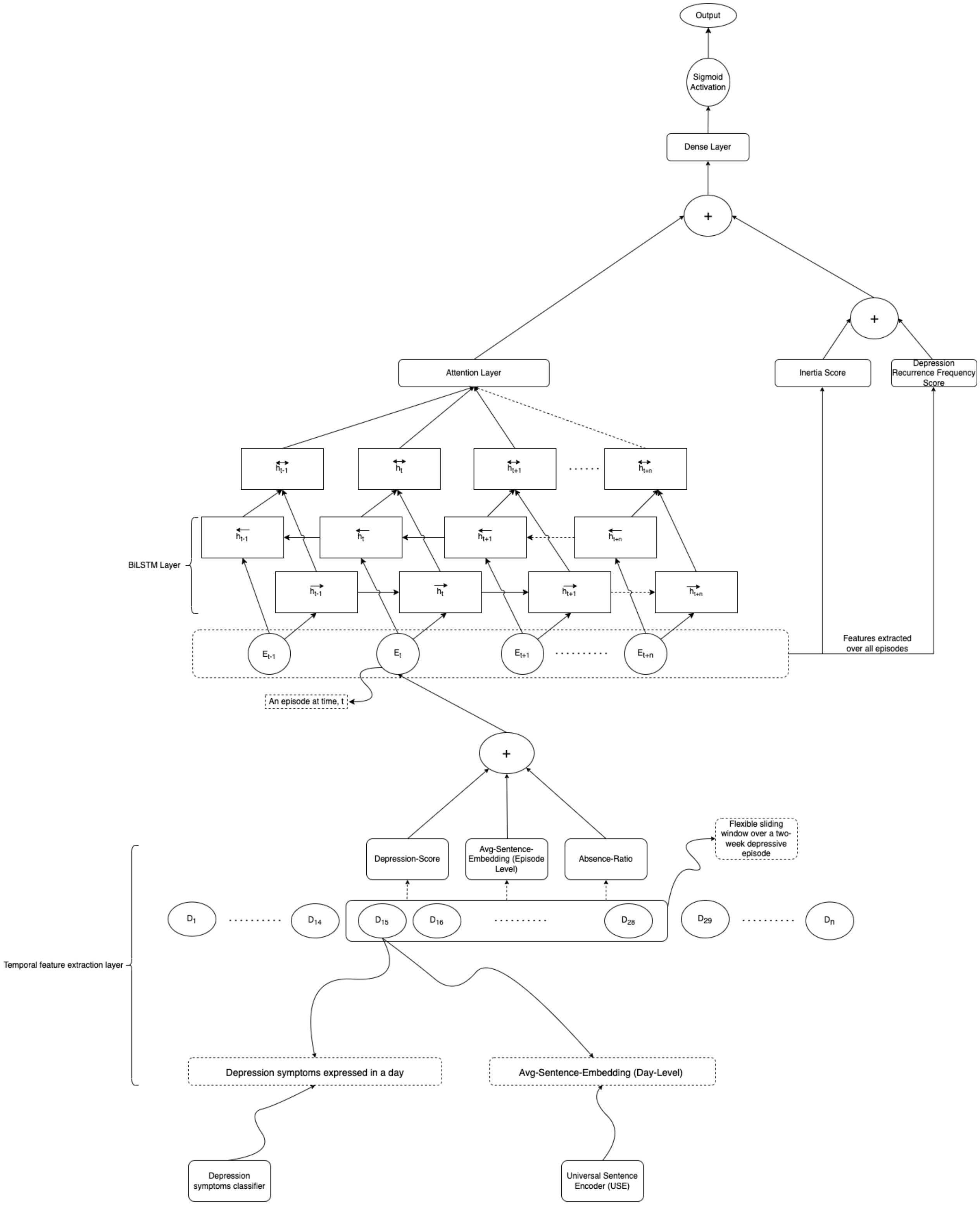

Output
Sigmoid Activation
Dense Layer
+
+
Attention Layer
Inertia Score
Depression Recurrence Frequency Score
$\overleftrightarrow{h}_{t-1}$
$\overleftrightarrow{h}_{t}$
$\overleftrightarrow{h}_{t+1}$
$\overleftrightarrow{h}_{t+n}$
$\overleftarrow{h}_{t-1}$
$\overleftarrow{h}_{t}$
$\overleftarrow{h}_{t+1}$
$\overleftarrow{h}_{t+n}$
BiLSTM Layer
$\overrightarrow{h}_{t-1}$
$\overrightarrow{h}_{t}$
$\overrightarrow{h}_{t+1}$
$\overrightarrow{h}_{t+n}$
E_{t-1}
E_t
E_{t+1}
E_{t+n}
Features extracted over all episodes
An episode at time, t
+
Depression-Score
Avg-Sentence-Embedding (Episode Level)
Absence-Ratio
Flexible sliding window over a two-week depressive episode
D_1
D_{14}
D_{15}
D_{16}
D_{28}
D_{29}
D_n
Temporal feature extraction layer
Depression symptoms expressed in a day
Avg-Sentence-Embedding (Day-Level)
Depression symptoms classifier
Universal Sentence Encoder (USE)

Figure 6.1: Detailed TUD model architecture (+ means concatenation, curved arrow followed by dashed box provides description of a component, solid arrow means data/process flow and dashed arrow means the same from "n" number of items).

gregates of depression score and semantic representations. This flexibility helps us doing three kinds of granular analysis over a user's depressive episodes based on sliding lengths of 1, 7 and 14. Later, these temporal episode level feature representations are concatenated and further fed to a BiLSTM (Appendix A.5) encoder to learn necessary temporal patterns of depression. This step produces encoder output, h_i for each episode and is further multiplied with final BiLSTM hidden representation, h_{final}. This is done for the entire temporal episode sequence to determine an attention weight, w_i for each episode (Equations A.2 and 6.5). Each w_i is then normalized based on a softmax function which turns it to an attention score α_i. This attention mechanism has been proposed by Bahdanau et al. [7] and is often called, "Global Attention" or "Bahdanau Attention." Finally, we calculate a fixed length Attention score weighted sum of encoder outputs or episodes, C (Equation 6.6), which is further fed to a fully connected or dense layer followed by a sigmoid activation function outputting a binary value, "1" indicating presence or "0" indicating absence of depression. Hyperparameter settings for training TUD model is provided in the Appendix A.11.

$$w_i = attention(h_i, h_{final}) \tag{6.4}$$

$$\alpha_i = \frac{\exp(w_i)}{\sum_{k=1}^{|sequence|} \exp(w_k)} \tag{6.5}$$

$$C = \sum_{i=1}^{|sequence|} \alpha_i h_i \tag{6.6}$$

We report the accuracy scores (described in next section) for user level depression detection task individually for each of our three datasets (i.e., CLPsych-2015-Users, IJCAI-2017-Ongoing-Users and Mixed-Users) for slide length=1 (because this provides us with the best results) and for the traditional clinical depression detection settings (alternative settings do not provide better results) described earlier for following experiments:

1. **Ablation tests:** We start with a model with all features (all-feats), then compare this model's accuracy scores with all the other feature ablated versions

of it.

2. **Single feature tests:** We report the model's performance for each individual feature to asses that single feature's discriminatory power.

3. **Best sliding length configurations:** We report the best model for other sliding configurations, i.e., for slide lengths = 7 and 14.

4. **Baselines:** We create two baselines, such as: (1) **Episodic Semantic Representation based model (ES)**: This model uses Episode Level Sentence Embedding Average (ELSEA) (Section 6.5.2) and the BiLSTM-Attention model for depression detection and (2) **All Historic Tweets Semantic Representation based model (HTS)**: This model uses All Tweets Embedding Average (ATEA) representation (Section 6.5.2) followed by a fully connected layer for depression detection task.

5. **Non-Clinical vs Clinical Setting:** We also report whether following strict clinical criteria for depression detection, i.e., verifying the presence of either "Anhedonia" or "Low Mood," makes any difference in user level depression detection compared to non-clinical settings (described in Section 6.5.1).

6. **Apriori vs Non-Apriori DSD model Settings:** We use a version of DSD model which use Apriori to predict samples for the labels for which DSD has weak performance (Chapter 5). We then use this model for TUD task.

7. **Minimal Depression Expression (MDE) based temporal modelling:** Based on the depression level features distribution, we confirm that the "None" level is higher in control than depression (Section 6.5), which indicates that we may try MDE to observe any increase the accuracy for DS.

6.8 Evaluation

Since our task is a binary classification task, for accuracy analysis, we report Precision, Recall and F1 scores for each of our three datasets across the corresponding held-out sets and a test set. To enable 10 fold cross validation (CV), we create

10 (train set, held-out set) pairs. We then report average Precision, Recall and F1 scores and their standard deviations across this 10 folds. We also report, how our models trained on each folds do on a separate test set, i.e., in IJCAI-2017-Today-Users dataset (Section 6.3.1).

This provides us with the information on how generalizable our model is in a dataset with totally different data distribution. We use two-tailed paired t-test and consider the difference between two accuracy scores as significant if p-value is < 0.05.

6.9 Results Analysis

In this section, we provide results analysis in the following dimensions (corresponding experiments are reported in Tables 6.9, 6.10, 6.11, 6.12, 6.13 and 6.14). Underline signifies the score is significantly worse than that of the model which uses all the features (all-feats model).

1. **Feature ablation study:** For all three dataset experiments, we do not see any significant accuracy difference among the ablated models and all-feats in both the held-out and test sets except avg-embedding ablated model, which performs significantly worse.

2. **Single feature study:** We report single features, i.e., depression-score (DS), absence ratio (AR) and temporal patterns (TP) for all the experiment datasets. We use Temporal Patterns (TP), which is a vector of two scores, i.e., IS and Depression Recurrence Frequency Score (DRFS). TP is calculated over a user's timeline unlike a series of scores like DS and AR. We see that these models are highly unstable, i.e., they have high variability in accuracy scores across different folds in held-out and test sets. Performance becomes exclusively worse when the train and test sets are from different distribution, i.e., number of episodes vary by a large margin (Tables 6.9, 6.10, 6.11, 6.12, 6.13 and 6.14).

   For the same data distribution (i.e., in held-out set of CLPsych-2015-Users

Datasets	**Category**	**Experiment-Name**	**Precision (Mean)**	**Precision (STD-Dev)**	**Recall (Mean)**	**Recall (STD-Dev)**	**F1 (Mean)**	**F1 (STD-Dev)**
CLPsych-2015-Users	feature ablation tests	all-feats	0.7121	0.1397	0.7114	0.1792	0.7021	0.1362
		all-feats (MDE)	0.7636	0.1173	0.6659	0.122	0.7055	0.1012
		- DS	0.6749	0.1467	0.7317	0.1319	0.6815	0.0747
		- IS	0.7253	0.1208	0.7615	0.0635	0.736	0.0739
		- DRFS	0.7199	0.1194	0.7137	0.139	0.7005	0.0824
		- TP	0.7405	0.1118	0.7289	0.1229	0.7230	0.0837
		- AR	0.7176	0.1176	0.7718	0.0732	0.7371	0.0727
		- ES	0.7649	0.1334	0.3521	0.1488	0.4671	0.1492
	single features	DS	0.6834	0.1279	0.3915	0.2105	0.467	0.1651
		DS (MDE)	0.623	0.1238	0.7142	0.1064	0.659	0.0943
		TP	0.6826	0.2244	0.4603	0.3842	0.4125	0.1661
		TP (MDE)	0.6248	0.1251	0.7214	0.0966	0.6592	0.0764
		AR	0.5327	0.1121	0.4520	0.1803	0.4593	0.1088
	best slides	- DRFS (Slide-1)	See above	See above	See above	See above	See above	See above
	best clinical settings	- AR (Slide-1)	See above	See above	See above	See above	See above	See above
	baselines	ES	0.6839	0.1438	0.7256	0.1323	0.6836	0.0771
		HTS	0.7057	0.149	0.7485	0.1057	0.7091	0.0672
		Yadav et al. [152]	-	-	-	-	0.7079	-

Table 6.9: CLPsysch-2015-Users dataset accuracy scores in held-out dataset.

Datasets	Category	Experiment-Name	Precision (Mean)	Precision (STD-Dev)	Recall (Mean)	Recall (STD-Dev)	F1 (Mean)	F1 (STD-Dev)
CLPsych-2015-Users	feature ablation tests	all-feats	0.5358	0.0156	0.9611	0.07	0.6872	0.025
		all-feats (MDE)	0.5218	0.0294	0.8166	0.0527	0.6363	0.033
		- DS	0.5093	0.0205	0.9722	0.0878	0.6659	0.0123
		- IS	0.5302	0.0158	0.9944	0.0176	0.6913	0.1002
		- DRFS	0.5509	0.068	0.95	0.1213	0.6877	0.0235
		- TP	0.5377	0.0283	0.9833	0.0268	0.6947	0.0238
		- AR	0.5341	0.0283	0.9944	0.0176	0.6944	0.0231
		- ES	0	0	0	0	0	0
	single features	DS	0.0514	0.1626	0.1000	0.3162	0.0679	0.2148
		DS (MDE)	0.4913	0.0091	0.8000	0.0287	0.6087	0.0145
		TP	0.1500	0.2415	0.3000	0.4830	0.2000	0.3220
		TP (MDE)	0.4764	0.0084	0.7833	0.0175	0.5924	0.0093
		AR	0	0	0	0	0	0
	best slides	- AR (Slide-1)	See above	See above	See above	See above	See above	See above
	best clinical settings	- AR (Slide-1)(NCS)	0.5368	0.0264	0.9889	0.0228	0.6953	0.0203
	baselines	ES	0.5101	0.0070	0.9944	0.0176	0.6742	0.0065
		HTS	0.5037	0.0093	0.9944	0.0171	0.6686	0.0094

Table 6.10: CLPsych-2015-Users dataset accuracy scores in test dataset.

Datasets	**Category**	**Experiment-Name**	**Precision (Mean)**	**Precision (STD-Dev)**	**Recall (Mean)**	**Recall (STD-Dev)**	**F1 (Mean)**	**F1 (STD-Dev)**
IJCAI-2017-Ongoing-Users	feature ablation tests	all-feats	0.7669	0.1229	0.7943	0.0818	0.7770	0.0930
		all-feats (MDE)	0.7757	0.1074	0.7450	0.1157	0.7529	0.0789
		- DS	0.7903	0.1256	0.7766	0.1034	0.7779	0.0937
		- IS	0.7722	0.1478	0.7713	0.1245	0.7703	0.1319
		- DRFS	0.7594	0.1358	0.7480	0.1104	0.7502	0.1105
		- TP	0.7513	0.1504	0.7055	0.0998	0.7200	0.0988
		- AR	0.7662	0.1511	0.6852	0.1367	0.7174	0.1213
		- ES	0.6263	0.0784	0.4487	0.1879	0.5031	0.1342
	single features	DS	0.5481	0.2989	0.3002	0.3628	0.2854	0.2052
		DS (MDE)	0.5261	0.1094	0.7893	0.0979	0.6259	0.0930
		TP	0.3647	0.2000	0.6579	0.4521	0.4317	0.2827
		TP (MDE)	0.5378	0.1039	0.7890	0.1087	0.6297	0.0796
		AR	0.6328	0.1171	0.5690	0.1230	0.5863	0.0818
	best slides	- DS (Slide-1)	See above	See above	See above	See above	See above	See above
	best clinical settings	- DS (Slide-1)	See above	See above	See above	See above	See above	See above
	baselines	ES	0.7830	0.0969	0.7746	0.0846	0.7763	0.0785
		HTS	0.7447	0.1384	0.7001	0.0804	0.7138	0.0757

Table 6.11: IJCAI-2017-Ongoing-Users dataset accuracy scores in held-out dataset.

Datasets	**Category**	**Experiment-Name**	**Precision (Mean)**	**Precision (STD-Dev)**	**Recall (Mean)**	**Recall (STD-Dev)**	**F1 (Mean)**	**F1 (STD-Dev)**
IJCAI-2017-Ongoing-Users	feature ablation tests	all-feats	0.7632	0.0284	0.8389	0.0805	0.7975	0.0439
		all-feats (MDE)	0.7498	0.0804	0.7444	0.0652	0.7429	0.0413
		- DS	0.7762	0.0488	0.8444	0.0438	0.8067	0.0149
		- IS	0.7661	0.0357	0.8278	0.0715	0.7937	0.0376
		- DRFS	0.7639	0.0160	0.8222	0.0631	0.7905	0.0252
		- TP	0.7645	0.0305	0.8056	0.0705	0.7828	0.0377
		- AR	0.7543	0.0742	0.6889	0.0951	0.7129	0.0446
		- ES	0.6535	0.0399	0.5056	0.1321	0.5613	0.09738
	single features	DS	0.1000	0.2108	0.2000	0.4216	0.1333	0.2811
		DS (MDE)	0.4976	0.0155	0.7500	0.0705	0.5964	0.0178
		TP	0.3529	0.2436	0.7000	0.4830	0.4692	0.3238
		TP (MDE)	0.4899	0.0285	0.7166	0.1094	0.5771	0.0338
		AR	0.6792	0.0170	0.6444	0.0537	0.6599	0.0299
	best slides	- DS (Slide-1)	See above	See above	See above	See above	See above	See above
	best clinical settings	- DS (Slide-1)	See above	See above	See above	See above	See above	See above
	baselines	ES	0.7721	0.0662	0.7611	0.0644	0.7632	0.0391
		HTS	0.7473	0.0496	0.7278	0.1184	0.7317	0.0726

Table 6.12: IJCAI-2017-Ongoing-Users dataset accuracy scores in test dataset.

Datasets	Category	Experiment-Name	Precision (Mean)	Precision (STD-Dev)	Recall (Mean)	Recall (STD-Dev)	F1 (Mean)	F1 (STD-Dev)
Mixed-Users	feature ablation tests	all-feats	0.6249	0.0800	0.7938	0.0680	0.6951	0.0515
		all-feats (MDE)	0.7783	0.0991	0.6592	0.0885	0.7072	0.0664
		- DS	0.6098	0.0618	0.8315	0.0806	0.7003	0.0482
		- IS	0.6260	0.0638	0.7078	0.0958	0.6612	0.0629
		- DRFS	0.6267	0.0783	0.7930	0.1003	0.6950	0.0604
		- TP	0.6279	0.0592	0.7783	0.0893	0.6923	0.0556
		- AR	0.6353	0.0646	0.7745	0.0831	0.6951	0.0549
		- ES	0.6787	0.0901	0.3402	0.1510	0.4386	0.1411
	single features	DS	0.6728	0.1444	0.2793	0.1520	0.3852	0.1647
		DS (MDE)	0.5864	0.0634	0.7518	0.0839	0.6565	0.0565
		TP	0.5725	0.1440	0.5697	0.4325	0.4395	0.2035
		TP (MDE)	0.5762	0.0631	0.7234	0.0878	0.6376	0.0495
		AR	0.6223	0.1115	0.4550	0.2290	0.4840	0.1974
	best slides	ES (Slide-7)	0.6005	0.0491	0.8579	0.0561	0.7051	0.0412
	best clinical settings	HTS (Slide-1)(NCS)	0.6957	0.0642	0.7266	0.1029	0.7071	0.0661
	baselines	ES	0.6088	0.0569	0.7984	0.0733	0.6874	0.0363
		HTS	0.6935	0.0578	0.7207	0.1083	0.7037	0.0692

Table 6.13: Mixed-Users dataset accuracy scores in held-out dataset.

Datasets	**Category**	**Experiment-Name**	**Precision (Mean)**	**Precision (STD-Dev)**	**Recall (Mean)**	**Recall (STD-Dev)**	**F1 (Mean)**	**F1 (STD-Dev)**
Mixed-Users	feature ablation tests	all-feats	0.5906	0.0316	0.9611	0.0457	0.7305	0.0239
		all-feats (MDE)	0.7677	0.0621	0.6278	0.0743	0.6855	0.0268
		- DS	0.5648	0.0535	0.9556	0.0574	0.7066	0.0240
		- IS	0.6251	0.0768	0.9167	0.1119	0.7343	0.0349
		- DRFS	0.5797	0.0414	0.9556	0.0631	0.7190	0.0220
		- TP	0.5808	0.0403	0.9556	0.0631	0.7201	0.0221
		- AR	0.5992	0.0466	0.9500	0.0715	0.7317	0.0269
		- ES	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	single features	DS	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
		DS (MDE)	0.4862	0.0073	0.7889	0.0234	0.6016	0.0123
		TP	0.2500	0.2635	0.5000	0.5270	0.3333	0.3514
		TP (MDE)	0.4811	0.0093	0.7667	0.0234	0.5910	0.0000
		AR	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	best slides	HTS (Slide-1)	See below	See below	See below	See below	See below	See below
	best clinical settings	HTS (Slide-1)	See below	See below	See below	See below	See below	See below
	baselines	ES	0.5555	0.0234	0.9889	0.0234	0.7110	0.0187
		HTS	0.7583	0.0335	0.8278	0.0410	0.7904	0.0215

Table 6.14: Mixed-Users dataset accuracy scores in test dataset.

and both held-out and test set for IJCAI-2017-Ongoing-Users), AR has the most predictive value.

In general, it is clear that single features are data distribution sensitive, however, TP is a bit less sensitive compared to others. Moreover, DS has better performance in a dataset which has more depressive episodes compared to other one with less number of depressive episodes (Table 6.3) and vice versa is true for TP. Except for AR in IJCAI-2017-Ongoing-Users dataset, all the other single features perform poorly and under chance level. It is as expected as we can see there is no significant difference between the control and depressed population across other features than AR in IJCAI-2017-Ongoing-Users dataset (Section 6.6).

Also, all the single features are significantly worse than the baselines, all ablated models (except in some cases Episodic-Semantic Representation (ES) ablated model) and all-feats models. We can also see, DS performs worst for the model trained in CLPsych-2015-Users and tested in IJCAI-2017-Today-Users. Moreover, DS performance in IJCAI-2017-Ongoing-Users is not as good as it is in CLPsych-2015-Users. This can be due to the fact that overall, DS score is not a significantly important factor to discriminate between depressed and control population in IJCAI-2017-Ongoing/Today-Users as we have observed in Section 6.6.

3. **Comparison with baseline models:** Both ES and All Historic Tweets Semantic Representation (HTS) are significantly better than avg-embedding ablated model and DS model across all datasets and in both held-out and test sets. In Mixed-Users test set, HTS is significantly better than all-feats and ablated models.

   Moreover, HTS is slightly (although not always significantly) better than ES in both held-out and test sets across all datasets except IJCAI-2017-Ongoing-Users dataset where the other way round is true. This somewhat signifies IJCAI-2017-Ongoing-Users dataset has more prominent temporal signals.

   In general, the power of embedding representation confirms that deep learning based methods learn better with high dimensional feature representations compared to low dimensional depression score based feature representations.

4. **Comparison with an early work baseline on same dataset** Yadav et al. [152] used their fine-tuned DSD model to detect user-level depression in CLPsych-2015 dataset, which yielded 0.7079 F1 accuracy (Table 6.9). Although, without more information it is hard to compare their accuracy with ours, we see that our least and best performing TUD models in all-feats and feature ablated categories achieve 0.6815 and 0.7371 mean F1 accuracy which are only $\approx 2\%$ less and $\approx 3\%$ more than theirs, confirming the efficacy of our underlying DSD model for user-level clinical depression modelling (Table 6.9).

5. **Sliding lengths contribution:** Sliding length adjustments do not have statistically significant effect on accuracy gain for all our experiments, except in the Mixed-Users dataset where sliding length 14 has significantly worse performance.

6. **Clinical vs non-clinical setting analysis:** Non-clinical setting does not have statistically significant effect on accuracy gain compared to clinical setting for all our experiments, confirms they have no effect in our clinical depression modelling. Distribution wise most of the depression episodes have "None" level of depression which represents depression score $0- < 4$; and this is present in abundant in both depressed vs. control population classes (Tables: 6.6, 6.7 and 6.8). Clinical setting does not introduce any significant change to this distribution; could be because of the lack of the symptoms "Anhedonia" and "Low Mood" being expressed in our datasets.

7. **Minimal Depression Expression (MDE) based temporal modelling:** We observe statistically significant increase for DS and TP in this mode (i.e., DS(MDE) and TP(MDE)) for all the experiments in test sets (Tables 6.10, 6.12 and 6.14). We also find, instead of concatenating DS if we element wise multiply it with temporal embedding representation (ES) to create all-feats

model (i.e., all-feats(MDE)), then there is some accuracy improvement over the original concatenation based all-feats model. This accuracy increase is also vetted by the statistically significant difference in “None”, “Minimal” and “IS” level for this mode, where except IJCAI-2017-Today-Users dataset, we see “None” level is lower in depressed population and higher in control population; “Minimal” and “IS” levels are higher in depressed population and lower in control population in all other datasets (Tables 6.6, 6.7 and 6.8). In IJCAI-2017-Today-Users dataset, we see the statistically non-significant reverse distribution patterns for “None”, “Minimal” and “IS” levels between these two populations. “DRFS” was found out to be majorly statistically non-significantly higher in depressed population compared to control population for all datasets except IJCAI-2017-Today-Users dataset, where the other way round is true. It is interesting to see still in IJCAI-2017-Today-Users dataset the models perform well despite they were trained on a reverse distribution, this could be due to the fact that depressed population have specific temporal pattern of depression scores which plays a discriminatory role here. We need to perform further investigation to confirm this in future.

Compared to Non-MDE, single feature based models seem to be more stable in this mode across held-out and test sets for all the datasets (Tables 6.9, 6.10, 6.11, 6.12, 6.13 and 6.14). However, through comparing the best models in each datasets in both held-out and test set for MDE and Non-MDE modes, we do not see their difference is statistically significantly different.

8. **Precision vs. recall:** We see in held-out sets precision and recall scores are somewhat close. However, in test sets, recall becomes higher and precision becomes lower, resulting in more sensitive models. Change in training data distribution (i.e., trained in more temporal episodes) results in sensitive models (as evaluated in test data).

All the above observations can be summarized into the following facts:

1. Performance of DS depends on the dataset characteristic, if in a particular

dataset, DS has significantly more discriminatory power then in that dataset DS might add more value.

2. In general, single feature based models perform worse than all features and ablated features based models. However, MDE mode leads to more stable and in many cases significantly better single feature based models compared to the Non-MDE mode.

3. Language only models (i.e., baseline models) are over-all pretty good in-terms of user-level depression detection compared to posting behavior of the users, expressed depression in the posts through their depression scores and relevant temporal patterns. Although, those features can positively effect the model performance provided that the data distribution is same in train and test sets.

4. Mixing two datasets with different distribution makes temporal modelling worse which is indicated by the fact that HTS performs better than all other temporal models in the Mixed-Users dataset.

5. Larger sliding length can result in similar model performance than more granular sliding lengths, indicating the promise for building more compact models in future.

6.10 Limitations

Some limitations of our work are provided below:

1. Our model uses depression score calculated based on the output of our Depression Symptoms Detection (DSD) model. This model is trained on a highly imbalanced dataset and is not robust to identify all the symptoms of depression from text.

2. We do not consider pure transformer models because earlier research do not show any extra benefit for using transformer for this kind of temporal modelling. The amount of memory needed by the Self-attention (Appendix A.5)

in the Transformer is quadratic on the length of the input, means there is a limitation on the input size. Another shortcoming for using a transformer is that, to represent a sequence, an explicit mechanism to inform the model on the order of episodes is needed which is not necessary in our architecture. There is a state-of-art transformer model called Temporal Fusion Transformer (TFT) for temporal modelling [64], however, it is not yet established whether TFT architecture is highly superior in the same. Interestingly, TFT has close connection to BiLSTM-Attention in its model architecture. Our future work will consider other Attention mechanisms to see if there is any improvements.

3. We follow clinical criteria of depression detection, which limits us from experimenting with various lengths of depressive episodes, i.e., episodes larger than two weeks or less. Likewise, we emphasize on the expression of depression symptoms in a tweet; if a candidate tweet expresses depression but no particular symptoms is detected (which is a rare possibility), that tweet do not contribute in the depression scoring. Since, we see with a sensitive setting in MDE mode (Section 6.9), we have some improvement in our clinical modelling. In our modelling, we do not explicitly account for other mental health conditions and bereavement that can resemble depression symptoms. We also do not ensure whether any depressive symptom causes significant change in a user's daily life functioning. In future, we would like to investigate more in this line to establish optimal thresholds and other depression criteria mentioned above in our clinical depression modelling.

4. Although Long Short Term Memory (LSTM) might not perform good for longer sequence, we use BiLSTM followed by Attention which helps alleviate problems with longer sequence to some extent.

5. Due to the datasets size, we do not try further stratification of users based on #Days in their timeline. IJCAI-2017-Ongoing-Users experiment however sheds some light of fixed timeline size, although the accuracy we achieve there is higher than that of CLPsych-2015, the general trend found in different features contribution is similar.

6. We largely follow the machine learning evaluation framework used in the seminal work of De Choudhury et al. [27] for social media based depression detection and their sample size is also similar to us[4]. However, they only reported avg. accuracy scores to compare different model-feature combinations. They did not report any std. deviation or t-test to compare their models. Moreover, most of the experiment results we report are not statistically significant, paired t-test in 10 fold cross validation is robust against Type-2 error (Appendix A.1 and [31]), which means, when there is no significant difference between two model's accuracy, we can confidently assume that their accuracy values are similar. We believe, our experiment results in an independent test set (i.e., IJCAI-2017-Today-Users) complement the analysis with the held-out set. Moreover, we find those clinical features have some discriminatory power which also have significant difference across depressed and control population (Section 6.6), this further corroborates the efficacy of our extracted clinical features. We also focus on the nature of change in accuracy scores rather than only comparing them by their value in our analysis, which also shed light on the performance of our various model-feature combinations.

6.11 Conclusion

In this chapter, we have described the construction of a deep Temporal User-level Clinical Depression Detection (TUD) model, using Twitter posts and all of their sub-components. These sub-components are created to help extract depression score (and few clinically relevant features based on it) from temporal social media posts. Later, we find their efficacy based on their accuracy for user-level depression

[4] There is some chance of Type-1 error in the paired t-test we used (i.e., it can reject null hypothesis when it is actually true), one solution to this is repeating cross-validation many times which is extremely time and resource intensive for the deep learning models and therefore is not suggested [31]. Also, another option is trying 5×2 fold CV, unfortunately, this has potential to largely decrease train set size which may render deep learning models to be ineffective. Non-parametric significance test, such as McNemer's test also has less statistical power in general when used in small dataset like ours. We tried non-parametric pairwise Wilcoxon's signed ranked test. However, we did not see any difference of this test with our pair-wise student's t-test in terms of statistical significance finding.

detection through several temporal and clinical depression related modes of analysis. We observe that, clinical features are more useful in same data distribution and some of the features are dataset specific. Also, semantic embedding representation is the most effective among all.

Chapter 7

Conclusion

In this chapter, we provide the summary findings of this research in the main areas of contribution mentioned in Chapter 1, i.e., (1) creating building blocks of Temporal User-level Clinical Depression Detection (TUD) model, and (2) analyzing the TUD model through various feature analyses in several data distributions and clinical depression detection settings. Later, we describe the limitations of our research in terms of the validity and reliability of the dataset curation and modelling approaches. We also discuss how we conform to the existing best practices of ethics in social media based mental health monitoring research. Finally, we discuss future research directions, including an outline of an actual depression monitoring system that might be useful for practicing clinicians to monitor their patients in clinical white space, i.e., the time between the visits to the clinician's office.

7.1 Summary of Contributions

In Chapter 1, we posed the main research question, described four goals we aimed to achieve to answer the research question, and discussed our two major contributions. These contributions have been actualized through the efforts to achieve four goals towards natural language modelling of clinical depression. In the following sections, we discuss the summary findings for each goal.

Developing Sub-modules of TUD Model

Following are the major findings while developing sub-components of the TUD model:

1. **Depressive Post Detection (DPD) Model:** We started by developing a DPD model, one of the most important sub-components for TUD model (Chapter 3). We faced a major obstacle in developing such a model because of the lack of expert human annotated datasets for the DPD task. To overcome this, we resorted to learning rich feature representations, such as word embeddings from depression forum posts, to represent our tweets. We showed that our learned Depression Specific Embedding (DSE) performed better than an existing TE representation for DPD task. We showed that vocabulary size does not matter provided the vocabulary is relevant to the task, e.g., DSE vocabulary is much smaller but relevant than TE with large and many irrelevant vocabulary words. We demonstrated that a non-linear mapping between the word vectors of common vocabulary can help improve the predictive power of TE, which further corroborates the efficacy of our learned depression specific embedding representation. We also experimented with two state-of-the-art sentence/contextual embedding representations and showed those perform better than all our word embedding representations. Finally, we developed a majority voting model (i.e., DPD-MV) using all our best word and sentence embedding based DPD models, and showed that in a large test set, DPD-MV performs the best.

2. **Zero-Shot Learning (ZSL) based Depression Symptoms Detection (DSD) Model:** Next, we described the application of state-of-the-art sentence embedding, word embedding, and natural language inference (NLI) models to develop a ZSL based DSD model (4). We showed that with this model, we do not need annotated samples for depression symptoms. In this case, we can leverage already existing clinical depression symptoms description found in various clinical resources, e.g., clinical handbook of depression diagnosis (e.g., DSM-5), depression rating scales (e.g., PHQ-9), and insights from a

practicing clinician. We demonstrated that such a model can achieve better accuracy than strong supervised models fine-tuned on small clinician annotated samples. We further showed that with the help of the constituency parse/syntax tree, we can make our model explainable, i.e., we can find out sub-phrases (or explanations) inside a tweet that explains its label for the depression symptom. Finally, we showed that we can develop a text explainability scoring module that can rank our explanations.

3. **Semi-supervised Learning (SSL) enhanced DSD model:** Finally, with the help of the DPD-MV and ZSL models, we outlined a Semi-supervised Learning (SSL) framework, that started from a small clinician annotated DSD samples and gathered more relevant samples for DSD (Chapter 5). In DSD sample annotation phase, we found the most agreement among the annotators were for the depression symptom labels “Suicidal Thoughts” and “Change in Sleep Patterns”. Further, we showed that our SSL process helps achieve a better DSD model in a couple of data harvest and model re-training iterations. In this process, ZSL based DSD model had a positive contribution. However, harvesting more samples this way from a large unlabelled sample repository for candidate depression symptom related tweets did not help to increase DSD model accuracy any further.

Design and Development of TUD Model and its Experimental Evaluation

In this section, we describe the findings during the design and development of the TUD model and its experimental evaluation as follows:

1. With the help of our DSD model trained with harvested samples of depression symptoms which is a by-product of the SSL framework, we finally designed and developed a TUD model which conforms to the criteria for clinical depression detection (Chapter 6). In this modelling process, we integrated temporal depression scores and patterns therein helpful to detect and monitor clinical depression, such as continuation of depressive mood through our

proposed Depression Inertia Score (IS) and recurrence of depressive mood through Depression Recurrence Frequency Score (DRFS). We reported experiment results based on all these features together through ablation study and their individual performances. We also reported the performance of the TUD model in the various granularity of temporal feature representations and clinical depression detection settings. Our clinical depression detection settings include (1) adherence to the clinical definition of depression and the lack thereof and (2) two thresholds for clinical depression scoring.

2. We found that only semantic representation based models provide the best accuracy. Furthermore, we observed that, depression score based patterns are not generalizable across different data distributions in train and test sets; however, for the same distribution, they perform better. Also, we did not see much difference among the performances of our models based on different granularity of temporal feature representations and clinical depression detection settings.

3. We also found depression score performs significantly better when we have a more sensitive threshold for depression scoring per episode.

7.2 Limitations and Efforts to Overcome

Earlier, we discussed the limitations of the various approaches relevant to each chapter. Here we discuss the limitations of our overall research in terms of the validity and reliability of our clinical depression modelling. This includes a discussion about our sample size, quality and biases, and annotation procedure. Further, we shed light on the underlying biases in the large language models and embedding representations that are used for this modelling. We also provide a description of our effort to overcome those limitations.

Validity of TUD Modelling

Validity refers to how accurately a psychometric test measures what it is supposed to measure. In our case, how accurately our TUD model measures the true depression of a user. We discuss the validity of our TUD model in terms of three well-known types of validity measures [146], such as the following:

1. **Construct Validity**: This refers to whether we can infer that a social media user has depression, provided that our TUD model has detected it. To ensure such validity, we require social media users who have ongoing depression as diagnosed by a clinician. Unfortunately, we do not have access to such datasets. So, we used social media users who self-disclosed their depression diagnosis. Research by De Choudhury et al. [27] showed that people rated as depressed by depression rating scales use more depression related language patterns compared to the control population in their Twitter timeline. Through leveraging linguistic clues, our TUD model detected user-level depression with good accuracy. This particular fact vouches, although weakly, for the construct validity of our modelling.

2. **Content Validity**: This refers to the extent to which our TUD modelling covers all the aspects necessary for measuring clinical depression. We followed the clinical depression diagnosis criteria mentioned in DSM-5 in our TUD modelling, which is absent in earlier research. Although it is not possible to cover all the aspects through social media language to measure clinical depression, our adherence to the depression diagnosis criteria helped us develop a model as content valid as possible.

3. **Criterion Validity:** This refers to the correlation of our clinical modelling to other valid measures for depression detection. At the heart of our TUD modelling is depression scoring, in which we follow clinical guidelines. We showed that in a more sensitive threshold for depression detection in an episode, the depression score provides good discriminatory power in depression detection, indicating that our depression scoring has some criterion validity.

Reliability of TUD Modelling

Reliability refers to how consistently TUD modelling discriminates between depression and control populations in different datasets from similar demographics.

Although we found semantic information alone is a reliable measure, our proposed individual clinical features are not so much so, i.e., they do not generalize well in held-out and test sets. Furthermore, the demographic information across our datasets is not very accurate; for example, for the CLPsych-2015-Users dataset, the demographic information was predicted through a machine learning model for detecting gender and age from their tweets, which may not be perfect. For IJCAI-2017-Users dataset, demographic information was not available; however, based on the majority of the users' demographic information on Twitter, it is assumed that most of the users in that dataset were around 25-34 years old [145].

Sample Size for Signs of Depression Detection Through Social Media Text

For Depression Symptoms Detection (DSD) task, we harvested a moderate sized sample of total $\approx 4.5K$ tweets with depression symptoms and the same amount of control tweets. Moreover, we harvested another set of $22K$ tweets from $\approx 0.4M$ tweets and showed it did not help increase accuracy for the DSD task. As mentioned earlier, our final Depressive Post Detection (DPD) model is also trained on the harvested DSD data. Samples of such well curated tweets (both properly filtered and carrying insights from expert annotation) for DPD and DSD task is rare in earlier work [149], [152], [155].

Sample Size for TUD Modelling

The sample size for TUD modelling is not very large. Our sample size for our two benchmark datasets, i.e., 537 users in CLPsych-2015-Users dataset and 392 users in IJCAI-2017-Users dataset, can be identified as moderate sized, compared to the similar work by De Choudhury et al. with a sample size of 476 users [27]. However, our dataset is still not sufficient to better deep-learning models. Therefore, we had to use as many samples as possible to train the TUD model, leaving a small subset for testing.

Sample Quality

Data quality is measured based on 12 dimensions, including data accessibility, data believability and completeness [51], [112]. Although social media datasets can be unreliable in terms of truthfulness and credibility, they are well known for their high availability. However, social media language is full of spelling errors and culture specific jargon and symbols, which makes it challenging to infer a complete picture of users' lives from it.

To reduce inherent noise in social media samples/tweets, we used a preprocessing pipeline, which considers only clean tweet samples devoid of hashtags and symbols for further analysis. Further, we normalized the words with repetitive characters to their original forms. Finally, we excluded depression diagnosis statements to avoid overfitting.

For DPD and DSD tasks, we ensured the quality of the respective samples through the SSL approach to harvest samples from depressed users' timelines. Further, we filtered these samples with the help of our DSD model, which utilized a state-of-the-art large pre-trained language model fine-tuned in clinician annotated samples. For DSD samples annotation, we put efforts into developing a depression symptoms annotation guideline based on clinical criteria of depression detection outlined in DSM-5, a well-known clinician rating scale for depression, and consultation with a practicing clinician.

For TUD modelling, it is not possible to verify the truthfulness of the users. However, we self-annotated a set of users for the genuineness of their depression self-disclosure and its recency. Based on this annotation process, we separated a set of users who have ongoing depression with a very high possibility and reported our TUD modelling performance in that dataset.

Sample Bias

Our datasets represent specific demographics (i.e., young adults aged around 25-34 years), with a majority of those being English speakers. Indeed our modelling outcomes did not provide general insight into depressive language characteristics across all demographics. On top of that, we depended on a user's self-disclosure, which in itself may not be true or a biased representation of a user's current mental health status. However, our annotation procedure to find the genuineness of self-disclosures and rigorous SSL process ensures the believability of our datasets to a great extent.

Annotation Procedure

It is very challenging to determine depression symptom labels for a tweet in absence of a proper context, which is the case for our Twitter samples. To overcome this challenge, we developed an annotation guideline with the help of a clinician rating scale for depression (i.e., MADRS [29]) instead of a patient rating for the same. The clinician rating scale is based on how a clinician would rate a patient's depression depending on the patient's answers to the questions related to depression symptoms. We found this rating scale has sufficient clues for depression symptoms detection through language. Moreover, we consulted with a practicing clinician and went through examples of confusing tweets and their possible labels. We use those samples as the examples in the annotation guideline to provide indications to the annotators to guide their annotation decisions. Thus, our adherence to clinical depression symptoms detection criteria alleviated confusion among the annotators for the task of labelling DSD. Further, we employed four annotators, including one clinical expert and one practicing clinician, to ensure the overall validity of our annotated dataset.

To ensure the reliability of the annotation procedure, we reported a test-retest reliability score of 83%, which is very high and shows the proper understanding of the clinical depression guideline by the annotators and its reflection on the annotation process.

Bias in Pre-trained Representations

Large pre-trained language models and word embeddings contain inherent biases learned from different stereotypes hidden in the datasets these models are trained on [9], [90], [157]. Although detecting and mitigating these biases is itself a research topic, we found that sometimes these inherent biases in the pre-trained representations help in achieving better accuracy. For example, in DPD task, we achieve better accuracy through using Depression Specific Embedding (DSE) (Chapter 3). On the other hand, in DSD task, we achieved better accuracy through the use of Mental-BERT language model [55], which was pre-trained on large mental health corpus (Chapter 5).

7.3 Ethical Challenges

Although in public social media like Twitter, people publicly broadcast social media posts, they still do not perceive it as a purely public space [74]. Mikal et al. [77], who employed focus groups to understand the perceived ethics of utilizing Twitter for mental health research, found that Twitter users often fail to understand (1) "data permanence," (2) "data reach," and (3) "big data computation tools that can be used to analyze social media posts." Thus, Benton et al. [8] proposed an ethical research protocol for social media health research. In the following section, we discuss how we adopted the guidelines outlined in that protocol in our research.

Institutional Research Board (IRB) Review

Research on public social media posts that do not have population intervention or interaction may qualify for IRB ethics review exempt status. This research may still require an application to the IRB but we can expect a substantially simplified review process. When a research project does not include any human subjects an IRB review is not required. As of the writing of the thesis manuscript, we have obtained three IRB approvals from the University of Alberta Research Ethics Office covering all aspects of our overall research, including data collection and annotation.

Informed consent

It is not feasible to obtain informed consent individually from millions of users [101]. However, informed consent should be obtained whenever possible, especially for the users of private social media like Facebook or Twitter private accounts

[12]. Normally in such scenarios, users have more expectation on respecting their privacy [142]. Therefore in our research, we used all the datasets that are based on public Twitter accounts. Moreover, we obtained these datasets from external sources by signing a Data Use Agreement (DUA) that ensures that we have the permission to use the dataset and we will respect all ethical aspects regarding the dataset use, as outlined there.

User Interventions

Research involving user intervention may not qualify for IRB exemption. In our research, we do not have any user intervention. An example of such kind of intervention is by Kramer et al. [60], in which they manipulated Facebook users' feeds through varying emotional content and monitored the feeds' influence on users' emotional states. Kramer et al. did not obtain any informed consent from the users, which raised ethical reservations. In our research, we simply analyzed depressed users' Twitter timelines instead of creating any user interaction and/or intervention.

Protections for Sensitive Data

Appropriate protection measures should be taken to protect sensitive data, even if it is public. For example, Twitter data from users self-disclosing their mental health status may have sensitive information regarding them. In that case, data can be stored in a safe storage that can restrict access to users' sensitive information, as done previously by [19]. In our research, only the cosigners of the Data Use Agreement (DUA) have access to the datasets. DUA ensures the privacy of the social media users of those datasets. A portion of one of these datasets has been used for further annotation for Depression Symptoms; in that case, we took signed consent from our annotators to ensure that the samples would not be searched through Internet for possible linking or identifying users by the annotators. We did not reveal any direct user identification to the public, nor did we try to find any user's identification over Internet or by any means in any stage of our research. Also, among the three major datasets we used in our research, two were anonymized for user identity by their curators.

User Attribution

Although public social media posts are freely accessible to anyone, users may not intend their posts to have such a broad audience. For that reason, Benton's guidelines [8] proposed three measures as follows:

1. **Removing usernames and profile pictures from publications:** We do not expose usernames, profile pictures, or any direct user identifiers in any papers or presentations.

2. **Paraphrasing original social media message:** We paraphrased original message, especially if it contains uniquely user identifying sensitive information. In general, our machine learning performance has been measured mostly quantitatively, meaning, we don't need to expose any such messages unless absolutely needed.

3. **Use of synthetic examples:** We did not have to use any synthetic examples to obfuscate user messages in our research.

User De-identification in Analysis

Researchers should remove the identity of a user and other sensitive personal information if it is not needed. Since our clinical depression model leverages language to detect depression, we only use language samples or tweets and no other user identifiers. Our algorithms further use an encoded text representation, i.e., word/sentence embeddings for representing tweets, which works as a layer of tweets obfuscation.

Sharing Data

Sharing datasets with external research groups to encourage reproducibility is important, but at the same time, we should ensure that these external research groups respect ethical and privacy concerns. In our research, we haven't yet finalized a protocol for safe and secure data sharing. We are still in the phase of developing one. However, we can readily share the encoded version of our datasets, e.g., word/sentence embeddings of tweets. Although this encoding has obvious drawbacks, such as limiting other research groups to use some fixed text representations, we believe this is the first step for us in terms of sharing our annotated datasets with external researchers. Of course, in any case, we will obtain consent from external research groups so that they do not try to identify any user should there be a way to do so.

Data Linkage Across Sites

It is possible to link users from their multiple public social media profiles [33], [91]. Researchers should be cautious about this and refrain from doing so even if there is an opportunity. In our case, we did not try to link any user across their public social media or other available profiles. In the future, when we share our datasets, we will enforce this practice through a DUA that needs to be signed by external research groups.

7.4 Future Directions and Practical Implications

Here we provide future directions under the following themes:

Signs of Depression Detection

For signs of depression detection task, i.e., DPD (Chapter 3), in our research, we focused on utilizing learned representations and their augmentation. We found this is the very first step for developing a good supervised DPD model when training data is scarce. In this augmentation process, we use an off-the-shelf embedding representation for a particular social media domain, for us, it is Twitter, and we further enhance its capability for signs of depression detection from text. To some extent, this augmentation technique alleviates the need for learning a social media domain specific, i.e., Twitter or Facebook embedding from large number of social media posts.

Particularly, our intention was to bring attention to the fact that, we can learn a good task domain embedding representation, e.g., we can learn DSE from more regular and less noisy depression forums text (Chapter 3). Later, we can transfer the domain knowledge embedded in DSE to another social media domain through semantic augmentation technique. However, our augmented semantic representation, i.e., ATE did not perform better than DSE. Also, it performed non-significantly better than unaugmented TE, indicating there is extensive research needed to uncover what semantic information are crucial and what adds noise. DSE's performance was best compared to the other representations indicates atleast depression specific vocabulary identified by our word embedding creation algorithm is invaluable. On the other hand, ATE's performance slightly better than TE indicates semantic knowledge transfer has some positive impact in DPD modelling.

Moreover, our semantic augmentation was done at the vocabulary level, meaning, it is possible to use the same technique to develop better representations for other natural language text domains. So in this sense, our method is not restricted to social media language based analysis only. In light of the above discussion, we can see the following research directions for signs of depression detection:

1. We will develop an extensive framework to analyze why semantic augmentation works in terms of semantic knowledge transfer. As a part of this exercise, we will find a set of minimal vocabulary words sufficient for such augmentation.

2. We will also employ social network analysis techniques in the word embedding representation space to analyze the semantic relationship among words and observe how this changes through semantic augmentation techniques. Furthermore, we will analyze what changes are positive for DPD tasks and what changes are negative.

3. Since our initial experiment results are inconclusive (Chapter 4), we will investigate extensively to see if this augmentation also works for contextual representations such as sentence embeddings.

4. Finally, through these analyses we would like to come to a conclusion on what criteria should be fulfilled for better augmentation. For example, what

needs to be done to learn an embedding representation that leverages social media domain vocabulary and semantic knowledge transferred from depression specific embedding space.

ZSL Modelling of Text Based Clinical Depression Symptoms

We proposed ZSL modelling, which helps prevent being too dependent on depression symptoms labelled samples for DSD task (Chapter 4). However, it was not clear to us what specific label descriptors or text representation for each labels are useful for creating a better ZSL model. As an attempt to gain more clarity on this, we sketched a framework for ZSL based explainable DSD model.

We would like to extensively explore the following areas for finalizing such framework. Since we used average embedding representations to represent tweets and depression label descriptors and cosine similarity to find association between them, it is possible to apply our ZSL modelling in other text domains, not just tweets. Thus we see the following future directions in the area of overall ZSL modelling:

1. We would like to extend our STEP algorithm (Chapter 4) so that it would work on a set of multiple sentences instead of a short a single sentence based tweet. A simpler approach could be to use STEP on every sentence in an excerpt and create a dictionary of depression symptoms and relevant explanations of phrases. However, this may miss on the global semantics of the text excerpt and instead focus only on constituent sentence semantics. Here, we can either use a dependency parser and develop a STEP mechanism on this parsing or we can summarize the excerpt with the help of STEP as an explanation.

2. We would like to evaluate our explanation score in various edge cases to ensure whether this scoring rewards those explanations that are also useful for humans. We would like to employ human annotators for such evaluation.

3. We would like to integrate this mechanism to make our user-level clinical

depression modelling more transparent and hence useful for practicing clinicians.

Harvesting Relevant Depression Symptoms Samples

We harvested depression symptoms samples from large Depressive Tweets Repository (DTR) to ensure both the symptoms distribution of depression related language and the relevant linguistic clues are present (Chapter 5). However, due to the huge imbalance in the sample space of depression symptom labels, it is not possible to learn a DSD model that is equally better at detecting every symptom. One would argue that learning such a model may not be needed, as those samples are rare in depressed users' profile. SSL based data harvesting helped us create a DSD model with good accuracy. We would like to explore if this model can also be exploited for longer social media text. Our initial observation shows some promise while using a short text DPD model to identify depression in longer depression forum texts [37]. We consider the following options to develop an even better model in the future:

1. We will generate synthetic samples by using state-of-the-art language models for various depression symptom labels and observe their contribution during user-level clinical depression detection.

2. We would like to employ clinicians in our SSL model to make it a human-in-the-loop based Active Learning (AL) system. This way, we can collect more samples that carry important clinical indicators for depression.

3. We can develop a cost sensitive model that learns to put more emphasis on the symptoms which are important to detect user-level clinical depression for a particular social media domain.

Design and Development of TUD Model

We developed a TUD model through the integration of all the clinical sub-components stated earlier. Our TUD model leveraged state-of-the-art sequential or temporal model, i.e., BiLSTM-Attention, which is capable of learning long term dependency

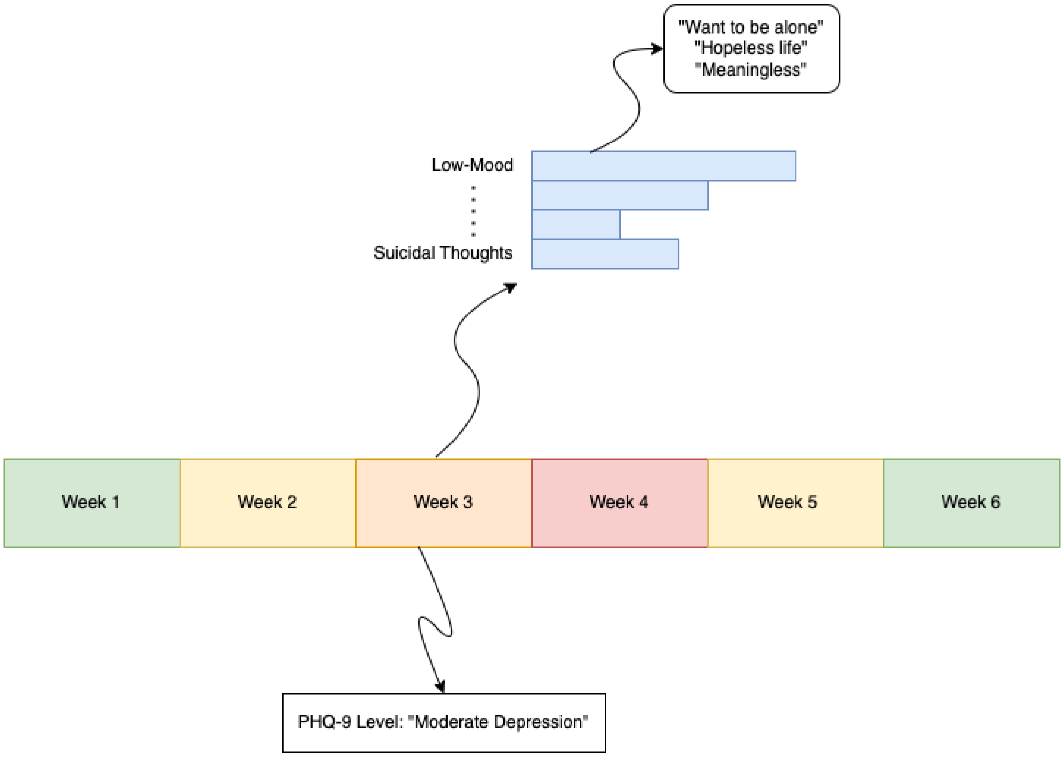


Figure 7.1: A tool for clinical depression monitoring through social media text.

among the depressive episodes. Moreover, our model can also learn important temporal episodes contributing to the detection of user-level clinical depression through Attention scores. In our work, we considered the clinical criteria for user-level depression detection; however, it is also possible to consider other textual features, such as basic and depression specific emotions and their fluctuations over time. We proposed a model that can detect basic and depression specific features through tweets [35]; however, we did not explore its potential in user-level clinical depression modelling. Since we proposed all the building blocks in this research to develop a transparent and clinically relevant depression detection model, we would like to use these building blocks to explore further in the following areas:

1. We will develop a tool (Figure 7.1) that can be used by the clinicians to monitor the depression levels over temporal depressive episodes in a user's social media timeline. Further, clinicians will be able to zoom into those episodes to see, what particular symptoms are expressed and relevant phrases that convey those symptoms.

2. We will investigate whether Attention scores are a better representation of depression levels in each episode.

3. We will investigate whether we can use the current model, provided that, we have temporal depression based feature representations in other text domain, e.g., day-to-day conversations. We believe it is possible; however we need to summarize each text conversation unit into a small text excerpt and then see the efficacy of the TUD model there.

4. We will find out how TUD model performs to forecast depression, i.e., how many days before an actual depression diagnosis we can predict a user's depression. Along these lines, we have already made some progress. For example, we identified few users with disclosures indicating their exact date of depression. However, this sample size is very small, and we need to collect more samples to conduct such experiments more extensively.

5. We will develop a system for verifying genuineness of an user for the self-disclosure of their depression based on their timeline and profile information, etc.

6. We will integrate our extracted features and compare their efficacy with other features related to other modalities, e.g. Emotion patterns, Voice, ECG etc, and find ways to combine them efficiently to build a more complete depression detection system.

7. We will try to find an optimal threshold for clinical depression scoring that does not hamper classification accuracy.

8. We will ensure that our model follows all the guidelines for ethically aligned design (Appendix A.2).

References

[1] A. Abboute, Y. Boudjeriou, G. Entringer, J. Azé, S. Bringay, and P. Poncelet, “Mining twitter for suicide prevention,” in *International Conference on Applications of Natural Language to Data Bases/Information Systems*, Springer, 2014, pp. 250–253.

[2] R. Agrawal, R. Srikant, *et al.*, “Fast algorithms for mining association rules,” in *Proc. 20th int. conf. very large data bases, VLDB*, Citeseer, vol. 1215, 1994, pp. 487–499.

[3] T. Al Hanai, M. M. Ghassemi, and J. R. Glass, “Detecting depression with audio/text sequence modeling of interviews.,” in *Interspeech*, 2018, pp. 1716–1720.

[4] J. Alonso, M. Codony, V. Kovess, *et al.*, “Population level of unmet need for mental healthcare in europe,” *The British journal of psychiatry*, vol. 190, no. 4, pp. 299–306, 2007.

[5] S. Atakishiyev, H. Babiker, N. Farruque, *et al.*, “A multi-component framework for the analysis and design of explainable artificial intelligence,” *arXiv preprint arXiv:2005.01908*, 2020.

[6] J. L. Baddeley, G. R. Daniel, and J. W. Pennebaker, “How henry hellyer’s use of language foretold his suicide.,” *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, vol. 32, no. 5, p. 288, 2011.

[7] D. Bahdanau, K. Cho, and Y. Bengio, “Neural machine translation by jointly learning to align and translate,” *arXiv preprint arXiv:1409.0473*, 2014.

[8] A. Benton, G. Coppersmith, and M. Dredze, “Ethical research protocols for social media health research,” in *Proceedings of the first ACL workshop on ethics in natural language processing*, 2017, pp. 94–102.

[9] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel, “Understanding the origins of bias in word embeddings,” in *International conference on machine learning*, PMLR, 2019, pp. 803–811.

[10] S. L. Burcusa and W. G. Iacono, “Risk for recurrence in depression,” *Clinical psychology review*, vol. 27, no. 8, pp. 959–985, 2007.

[11] S. Byun, A. Y. Kim, E. H. Jang, *et al.*, "Detection of major depressive disorder from linear and nonlinear heart rate variability features during mental task protocol," *Computers in biology and medicine*, vol. 112, p. 103 381, 2019.

[12] F. Celli, F. Pianesi, D. Stillwell, and M. Kosinski, "Workshop on computational personality recognition: Shared task," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 7, 2013, pp. 2–5.

[13] D. Cer, Y. Yang, S.-y. Kong, *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[14] X. Chen, M. D. Sykora, T. W. Jackson, and S. Elayan, "What about mood swings: Identifying depression on twitter with temporal measures of emotions," in *Companion of the The Web Conference 2018 on The Web Conference 2018*, International World Wide Web Conferences Steering Committee, 2018, pp. 1653–1660.

[15] P. G. F. Cheng, R. M. Ramos, J. Á. Bitsch, *et al.*, "Psychologist in a pocket: Lexicon development and content validation of a mobile-based app for depression screening," *JMIR mHealth and uHealth*, vol. 4, no. 3, 2016.

[16] L. A. Clark, B. Cuthbert, R. Lewis-Fernández, W. E. Narrow, and G. M. Reed, "Three approaches to understanding and classifying mental disorder: Icd-11, dsm-5, and the national institute of mental health's research domain criteria (rdoc)," *Psychological Science in the Public Interest*, vol. 18, no. 2, pp. 72–145, 2017.

[17] B. L. Cook, A. M. Progovac, P. Chen, B. Mullin, S. Hou, and E. Baca-Garcia, "Novel use of natural language processing (nlp) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid," *Computational and mathematical methods in medicine*, vol. 2016, 2016.

[18] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," *ACL 2014*, vol. 51, 2014.

[19] G. Coppersmith, M. Dredze, C. Harman, and K. Hollingshead, "From adhd to sad: Analyzing the language of mental health on twitter through self-reported diagnoses," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 1–10.

[20] 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*, 2015, pp. 31–39.

[21] G. Coppersmith, C. Harman, and M. Dredze, "Measuring post traumatic stress disorder in twitter," in *Eighth international AAAI conference on weblogs and social media*, 2014.

[22] G. Coppersmith, K. Ngo, R. Leary, and A. Wood, "Exploratory analysis of social media prior to a suicide attempt," in *Proceedings of the third workshop on computational linguistics and clinical psychology*, 2016, pp. 106–117.

[23] M. A. Covington, C. He, C. Brown, *et al.*, "Schizophrenia and the structure of language: The linguist's view," *Schizophrenia research*, vol. 77, no. 1, pp. 85–98, 2005.

[24] Z. Cui, R. Ke, Z. Pu, and Y. Wang, "Deep bidirectional and unidirectional lstm recurrent neural network for network-wide traffic speed prediction," *arXiv preprint arXiv:1801.02143*, 2018.

[25] M. De Choudhury, "Role of social media in tackling challenges in mental health," in *Proceedings of the 2nd international workshop on Socially-aware multimedia*, ACM, 2013a, pp. 49–52.

[26] M. De Choudhury and S. De, "Mental health discourse on reddit: Self-disclosure, social support, and anonymity.," in *ICWSM*, 2014.

[27] M. De Choudhury, M. Gamon, S. Counts, and E. Horvitz, "Predicting depression via social media.," in *ICWSM*, 2013b, p. 2.

[28] *Depression*. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/depression`.

[29] *Depression assessment instruments*. [Online]. Available: `https://www.apa.org/depression-guideline/assessment`.

[30] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[31] T. G. Dietterich, "Approximate statistical tests for comparing supervised classification learning algorithms," *Neural computation*, vol. 10, no. 7, pp. 1895–1923, 1998.

[32] P. S. Dodds, K. D. Harris, I. M. Kloumann, C. A. Bliss, and C. M. Danforth, "Temporal patterns of happiness and information in a global social network: Hedonometrics and twitter," *PloS one*, vol. 6, no. 12, e26752, 2011.

[33] M. Douriez, H. Doraiswamy, J. Freire, and C. T. Silva, "Anonymizing nyc taxi data: Does it matter?" In *2016 IEEE international conference on data science and advanced analytics (DSAA)*, IEEE, 2016, pp. 140–148.

[34] *DSM-5 FAQ*, `https://www.psychiatry.org/psychiatrists/practice/dsm/feedback-and-questions/frequently-asked-questions`, [Online; accessed 05-Dec-2019].

[35] N. Farruque, C. Huang, O. Zaiane, and R. Goebel, "Basic and depression specific emotion identification in tweets: Multi-label classification experiments," *arXiv preprint arXiv:2105.12364*, 2021.

[36] N. Farruque, O. Zaiane, and R. Goebel, “Augmenting semantic representation of depressive language: From forums to microblogs,” in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2019, pp. 359–375.

[37] N. Farruque, O. R. Zaıane, R. Goebel, and S. Sivapalan, “Deepblues@ lt-edi-acl2022: Depression level detection modelling through domain specific bert and short text depression classifiers,” in *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, 2022, pp. 167–171.

[38] S. de la Fuente Garcia, C. W. Ritchie, and S. Luz, “Artificial intelligence, speech, and language processing approaches to monitoring alzheimer’s disease: A systematic review,” *Journal of Alzheimer’s Disease*, vol. 78, no. 4, pp. 1547–1574, 2020.

[39] D. Gans, “Use of a preliminary test in comparing two sample means: Use of a preliminary test,” *Communications in Statistics-Simulation and Computation*, vol. 10, no. 2, pp. 163–174, 1981.

[40] A. Ghandeharioun, S. Fedor, L. Sangermano, *et al.*, “Objective assessment of depressive symptoms with machine learning and wearable sensors data,” in *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2017, pp. 325–332.

[41] C. H. E. Gilbert, “Vader: A parsimonious rule-based model for sentiment analysis of social media text,” in *Eighth International Conference on Weblogs and Social Media (ICWSM-14). Available at (20/04/16) http://comp. social. gatech. edu/papers/icwsm14. vader. hutto. pdf*, 2014.

[42] F. Godin, B. Vandersmissen, W. De Neve, and R. Van de Walle, “Multimedia lab @ acl wnut ner shared task: Named entity recognition for twitter microposts using distributed word representations,” in *Proceedings of the Workshop on Noisy User-generated Text*, 2015, pp. 146–153.

[43] K. Gowen, M. Deschaine, D. Gruttadara, and D. Markey, “Young adults with mental health conditions and social networking websites: Seeking tools to build community.,” *Psychiatric Rehabilitation Journal*, vol. 35, no. 3, p. 245, 2012.

[44] P. E. Greenberg, A.-A. Fournier, T. Sisitsky, C. T. Pike, and R. C. Kessler, “The economic burden of adults with major depressive disorder in the united states (2005 and 2010),” *The Journal of clinical psychiatry*, vol. 76, no. 2, pp. 155–162, 2015.

[45] K. Greff, R. K. Srivastava, J. Koutnık, B. R. Steunebrink, and J. Schmidhuber, “Lstm: A search space odyssey,” *IEEE transactions on neural networks and learning systems*, vol. 28, no. 10, pp. 2222–2232, 2016.

[46] J. F. Gunn and D. Lester, “Twitter postings and suicide: An analysis of the postings of a fatal suicide in the 24 hours prior to death,” *Suicidologi*, vol. 17, no. 3, 2012.

[47] K. Harrigian, C. Aguirre, and M. Dredze, "On the state of social media data for mental health research," *arXiv preprint arXiv:2011.05233*, 2020.

[48] M. G. Haselton, D. Nettle, P. W. Andrews, and D. M. Buss, "The handbook of evolutionary psychology," *The evolution of cognitive bias*, pp. 724–746, 2005.

[49] F. Holländare, G. Andersson, and I. Engström, "A comparison of psychometric properties between internet and paper versions of two depression instruments (bdi-ii and madrs-s) administered to clinic patients," *Journal of medical Internet research*, vol. 12, no. 5, e49, 2010.

[50] C. Homan, R. Johar, T. Liu, M. Lytle, V. Silenzio, and C. O. Alm, "Toward macro-insights for suicide prevention: Analyzing fine-grained distress at scale," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 107–117.

[51] A. Immonen, P. Pääkkönen, and E. Ovaska, "Evaluating the quality of social media data in big data architecture," *Ieee Access*, vol. 3, pp. 2028–2043, 2015.

[52] *Internet Encyclopedia of Philosophy*, `https://www.iep.utm.edu/ethics/`, [Online; accessed 05-Dec-2019].

[53] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, 2017, pp. 32–40.

[54] J. Jashinsky, S. H. Burton, C. L. Hanson, *et al.*, "Tracking suicide risk factors through twitter in the us.," *Crisis: The Journal of Crisis Intervention and Suicide Prevention*, vol. 35, no. 1, p. 51, 2014.

[55] S. Ji, T. Zhang, L. Ansari, J. Fu, P. Tiwari, and E. Cambria, "Mentalbert: Publicly available pretrained language models for mental healthcare," *arXiv preprint arXiv:2110.15621*, 2021.

[56] T. A. Ketter, "Diagnostic features, prevalence, and impact of bipolar disorder," *The Journal of clinical psychiatry*, vol. 71, no. 6, p. 27 652, 2010.

[57] S. Kiritchenko, X. Zhu, and S. M. Mohammad, "Sentiment analysis of short informal texts," *Journal of Artificial Intelligence Research*, vol. 50, pp. 723–762, 2014.

[58] R. Kiros, Y. Zhu, R. R. Salakhutdinov, *et al.*, "Skip-thought vectors," *Advances in neural information processing systems*, vol. 28, 2015.

[59] H. Kour and M. K. Gupta, "An hybrid deep learning approach for depression prediction from user tweets using feature-rich cnn and bi-directional lstm," *Multimedia Tools and Applications*, pp. 1–37, 2022.

[60] A. D. 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, no. 24, pp. 8788–8790, 2014.

[61] P. Kuppens, L. B. Sheeber, M. B. Yap, S. Whittle, J. G. Simmons, and N. B. Allen, “Emotional inertia prospectively predicts the onset of depressive disorder in adolescence.,” *Emotion*, vol. 12, no. 2, p. 283, 2012.

[62] O. Levy and Y. Goldberg, “Neural word embedding as implicit matrix factorization,” *Advances in neural information processing systems*, vol. 27, 2014.

[63] M. Lewis, Y. Liu, N. Goyal, *et al.*, “Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension,” *arXiv preprint arXiv:1910.13461*, 2019.

[64] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, “Temporal fusion transformers for interpretable multi-horizon time series forecasting,” *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

[65] G. Y. Lim, W. W. Tam, Y. Lu, C. S. Ho, M. W. Zhang, and R. C. Ho, “Prevalence of depression in the community from 30 countries between 1994 and 2014,” *Scientific reports*, vol. 8, no. 1, pp. 1–10, 2018.

[66] Y. Liu, M. Ott, N. Goyal, *et al.*, “Roberta: A robustly optimized bert pretraining approach,” *arXiv preprint arXiv:1907.11692*, 2019.

[67] D. E. Losada and F. Crestani, “A test collection for research on depression and language use,” in *International Conference of the Cross-Language Evaluation Forum for European Languages*, Springer, 2016, pp. 28–39.

[68] L. Lushi Chen, W. Magdy, H. Whalley, and M. Wolters, “Examining the role of mood patterns in predicting self-reported depressive symptoms,” *arXiv e-prints*, arXiv–2006, 2020.

[69] S. MacAvaney, B. Desmet, A. Cohan, *et al.*, “Rsdd-time: Temporal annotation of self-reported mental health diagnoses,” *arXiv preprint arXiv:1806.07916*, 2018.

[70] M. Marcus, M. T. Yasamy, M. van van Ommeren, D. Chisholm, and S. Saxena, *Depression: A global public health concern*, 2012.

[71] A. Markham and E. Buchanan, “Ethical decision-making and internet research: Version 2.0. recommendations from the aoir ethics working committee,” *Available online: aoir. org/reports/ethics2. pdf*, 2012.

[72] D. M. Maurer, T. J. Raymond, and B. N. Davis, “Depression: Screening and Diagnosis,” *Am Fam Physician*, vol. 98, no. 8, pp. 508–515, Oct. 2018.

[73] D. McClosky, E. Charniak, and M. Johnson, “Effective self-training for parsing,” in *Proceedings of the Human Language Technology Conference of the NAACL, Main Conference*, 2006, pp. 152–159.

[74] R. McKee, "Ethical issues in using social media for health and health care research," *Health policy*, vol. 110, no. 2-3, pp. 298–301, 2013.

[75] P. E. Meehl and A. Rosen, "Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores.," *Psychological bulletin*, vol. 52, no. 3, p. 194, 1955.

[76] N. C. C. for Mental Health (UK *et al.*, "The classification of depression and depression rating scales/questionnaires," in *Depression in Adults with a Chronic Physical Health Problem: Treatment and Management*, British Psychological Society, 2010.

[77] J. Mikal, S. Hurst, and M. Conway, "Ethical issues in using twitter for population-level depression monitoring: A qualitative study," *BMC medical ethics*, vol. 17, no. 1, pp. 1–11, 2016.

[78] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013c.

[79] D. N. Milne, G. Pink, B. Hachey, and R. A. Calvo, "Clpsych 2016 shared task: Triaging content in online peer-support forums," in *Proceedings of the Third Workshop on Computational Lingusitics and Clinical Psychology*, 2016, pp. 118–127.

[80] S. M. Mohammad and P. D. Turney, "Nrc emotion lexicon," *NRC Technical Report*, 2013.

[81] S. M. Mohammad and S. Kiritchenko, "Using hashtags to capture fine emotion categories from tweets," *Computational Intelligence*, vol. 31, no. 2, pp. 301–326, 2015, ISSN: 1467-8640. DOI: 10.1111/coin.12024. [Online]. Available: http://dx.doi.org/10.1111/coin.12024.

[82] *Montgomery Asberg Depression Rating Scale (MADRS) ,2013; Strokengine — strokengine.ca*, https://strokengine.ca/en/assessments/montgomery-asberg-depression-rating-scale-madrs/, [Accessed 22-Sep-2022].

[83] M. A. Moreno, L. A. Jelenchick, K. G. Egan, *et al.*, "Feeling bad on facebook: Depression disclosures by college students on a social networking site," *Depression and anxiety*, vol. 28, no. 6, pp. 447–455, 2011.

[84] M. Al-Mosaiwi and T. Johnstone, "In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation," *Clinical Psychological Science*, vol. 6, no. 4, pp. 529–542, 2018.

[85] M. Mousavian, J. Chen, Z. Traylor, and S. Greening, "Depression detection from smri and rs-fmri images using machine learning," *Journal of Intelligent Information Systems*, vol. 57, no. 2, pp. 395–418, 2021.

[86] D. Mowery, C. Bryan, and M. Conway, "Feature studies to inform the classification of depressive symptoms from twitter data for population health," *arXiv preprint arXiv:1701.08229*, 2017.

[87] D. Mowery, H. Smith, T. Cheney, *et al.*, "Understanding depressive symptoms and psychosocial stressors on twitter: A corpus-based study," *Journal of medical Internet research*, vol. 19, no. 2, 2017.

[88] D. L. Mowery, Y. A. Park, C. Bryan, and M. Conway, "Towards automatically classifying depressive symptoms from twitter data for population health," in *Proceedings of the workshop on computational modeling of people's opinions, personality, and emotions in social media (PEOPLES)*, 2016, pp. 182–191.

[89] N. Mukund, S. Thakur, S. Abraham, *et al.*, "An information retrieval and recommendation system for astronomical observatories," *The Astrophysical Journal Supplement Series*, vol. 235, no. 1, p. 22, 2018.

[90] M. Nadeem, A. Bethke, and S. Reddy, "Stereoset: Measuring stereotypical bias in pretrained language models," *arXiv preprint arXiv:2004.09456*, 2020.

[91] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, IEEE, 2008, pp. 111–125.

[92] J. Naslund, K. Aschbrenner, L. Marsch, and S. Bartels, "The future of mental health care: Peer-to-peer support and social media," *Epidemiology and psychiatric sciences*, vol. 25, no. 2, pp. 113–122, 2016.

[93] J. A. Naslund, S. W. Grande, K. A. Aschbrenner, and G. Elwyn, "Naturally occurring peer support through social media: The experiences of individuals with severe mental illness using youtube," *PLOS one*, vol. 9, no. 10, e110171, 2014.

[94] Y. Neuman, Y. Cohen, D. Assaf, and G. Kedma, "Proactive screening for depression through metaphorical and automatic text analysis," *Artificial intelligence in medicine*, vol. 56, no. 1, pp. 19–25, 2012.

[95] T. Nguyen, D. Phung, B. Dao, S. Venkatesh, and M. Berk, "Affective and content analysis of online depression communities," *IEEE Transactions on Affective Computing*, vol. 5, no. 3, pp. 217–226, 2014.

[96] T. Nguyen, A. Yates, A. Zirikly, B. Desmet, and A. Cohan, "Improving the generalizability of depression detection by leveraging clinical questionnaires," *arXiv preprint arXiv:2204.10432*, 2022.

[97] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[98] H. Nissenbaum, "Privacy as contextual integrity," *Wash. L. Rev.*, vol. 79, p. 119, 2004.

[99] B. O'dea, S. Wan, P. J. Batterham, A. L. Calear, C. Paris, and H. Christensen, "Detecting suicidality on twitter," *Internet Interventions*, vol. 2, no. 2, pp. 183–188, 2015.

[100] G. S. O'Keeffe, K. Clarke-Pearson, *et al.*, "The impact of social media on children, adolescents, and families," *Pediatrics*, vol. 127, no. 4, pp. 800–804, 2011.

[101] D. O'Connor, "The apomediated world: Regulating research when social media has changed research," *Journal of Law, Medicine & Ethics*, vol. 41, no. 2, pp. 470–483, 2013.

[102] 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*, 2018, pp. 88–97.

[103] L. Osterhout, A. Kim, and G. R. Kuperberg, "The neurobiology of sentence comprehension.," 2012.

[104] F. A. Paniagua, "Icd-10 versus dsm-5 on cultural issues," *Sage open*, vol. 8, no. 1, p. 2 158 244 018 756 165, 2018.

[105] M. J. Patel, A. Khalaf, and H. J. Aizenstein, "Studying depression using imaging and machine learning methods," *NeuroImage: Clinical*, vol. 10, pp. 115–123, 2016.

[106] J. W. Pennebaker, R. L. Boyd, K. Jordan, and K. Blackburn, "The development and psychometric properties of liwc2015," Tech. Rep., 2015.

[107] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001.

[108] J. Pennebaker, M. Mehl, and K. Niederhoffer, "Psychological aspects of natural language use: Our words, our selves," *Annual Review of Psychology*, vol. 54, pp. 547–577, 2003, ISSN: 0066-4308.

[109] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[110] A. Pérez, J. Parapar, and Á. Barreiro, "Automatic depression score estimation with word embedding models," *Artificial Intelligence in Medicine*, p. 102 380, 2022.

[111] M. E. Peters, M. Neumann, M. Iyyer, *et al.*, "Deep contextualized word representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202. [Online]. Available: https://aclanthology.org/N18-1202.

[112] L. L. Pipino, Y. W. Lee, and R. Y. Wang, "Data quality assessment," *Communications of the ACM*, vol. 45, no. 4, pp. 211–218, 2002.

[113] P. C. Price, R. S. Jhangiani, and I.-C. A. Chiang, “Reliability and validity of measurement,” *Research methods in psychology*, 2015.

[114] *Psychometrics*, `https://www.oxfordbibliographies.com/view/document/obo-9780195389678/obo-9780195389678-0156.xml`, [Online; accessed 05-Dec-2019], 2015.

[115] T. Pyszczynski and J. Greenberg, “Depression, self-focused attention, and self-regulatory perseveration,” in *Coping with negative life events*, Springer, 1987, pp. 105–129.

[116] A. G. Reece and C. M. Danforth, “Instagram photos reveal predictive markers of depression,” *EPJ Data Science*, vol. 6, no. 1, p. 15, 2017.

[117] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, C. M. Danforth, and E. J. Langer, “Forecasting the onset and course of mental illness with twitter data,” *Scientific reports*, vol. 7, no. 1, p. 13 006, 2017.

[118] N. Reimers and I. Gurevych, “Sentence-bert: Sentence embeddings using siamese bert-networks,” *arXiv preprint arXiv:1908.10084*, 2019.

[119] *Reliability vs. Validity*, `https://www.scribbr.com/methodology/reliability-vs-validity/`, [Online; accessed 05-Dec-2019], 2019.

[120] P. Resnik, W. Armstrong, L. Claudino, T. Nguyen, V.-A. Nguyen, and J. Boyd-Graber, “Beyond lda: Exploring supervised topic modeling for depression-related language in twitter,” in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2015, pp. 99–107.

[121] P. Resnik, A. Garron, and R. Resnik, “Using topic modeling to improve prediction of neuroticism and depression,” in *Proceedings of the 2013 Conference on Empirical Methods in Natural*, Association for Computational Linguistics, 2013, pp. 1348–1353.

[122] A. Rinaldi, J. E. F. Tree, and S. Chaturvedi, “Predicting depression in screening interviews from latent categorization of interview prompts,” in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7–18.

[123] F. Ringeval, B. Schuller, M. Valstar, *et al.*, “Avec 2017: Real-life depression, and affect recognition workshop and challenge,” in *Proceedings of the 7th Annual Workshop on Audio/Visual Emotion Challenge*, ACM, 2017, pp. 3–9.

[124] S. Rude, E.-M. Gortner, and J. Pennebaker, “Language use of depressed and depression-vulnerable college students.,” *Cognition & Emotion*, vol. 18, no. 8, pp. 1121–1133, 2004, ISSN: 02699931.

[125] M. Sabshin, “Depression: Clinical, experimental and theoretical aspects.,” *Archives of General Psychiatry*, vol. 19, no. 6, pp. 766–767, 1968.

[126] S. Saha, D. Chant, J. Welham, and J. McGrath, “A systematic review of the prevalence of schizophrenia,” *PLoS medicine*, vol. 2, no. 5, e141, 2005.

[127] G. Schomerus, H. Matschinger, and M. C. Angermeyer, "The stigma of psychiatric treatment and help-seeking intentions for depression," *European archives of psychiatry and clinical neuroscience*, vol. 259, no. 5, pp. 298–306, 2009.

[128] H. A. Schwartz, J. Eichstaedt, M. L. Kern, *et al.*, "Towards assessing changes in degree of depression through facebook," in *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, 2014, pp. 118–125.

[129] *Screening Tests for Depression*, `https://emedicine.medscape.com/article/1859039-overview`, [Online; accessed 05-Dec-2019].

[130] E. M. Seabrook, M. L. Kern, B. D. Fulcher, and N. S. Rickard, "Predicting depression from language-based emotion dynamics: Longitudinal analysis of facebook and twitter status updates," *Journal of medical Internet research*, vol. 20, no. 5, 2018.

[131] A. G. Shahraki and O. R. Zaïane, "Lexical and learning-based emotion mining from text," in *International Conference on Computational Linguistics and Intelligent Text Processing (CICLing)*, 2017.

[132] K. Shahriari and M. Shahriari, "Ieee standard review—ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems," in *2017 IEEE Canada International Humanitarian Technology Conference (IHTC)*, IEEE, 2017, pp. 197–201.

[133] G. Shen, J. Jia, L. Nie, *et al.*, "Depression detection via harvesting social media: A multimodal dictionary learning solution.," in *IJCAI*, 2017, pp. 3838–3844.

[134] J. H. Shen and F. Rudzicz, "Detecting anxiety through reddit," in *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality*, 2017, pp. 58–65.

[135] S. Siami-Namini, N. Tavakoli, and A. S. Namin, "A comparison of arima and lstm in forecasting time series," in *2018 17th IEEE international conference on machine learning and applications (ICMLA)*, IEEE, 2018, pp. 1394–1401.

[136] *Skip-gram Embedding*, `https://towardsdatascience.com/skip-gram-nlp-context-words-prediction-algorithm-5bbf34f84e0c`, [Online; accessed 05-Dec-2021].

[137] D. Smirnova, P. Cumming, E. Sloeva, N. Kuvshinova, D. Romanov, and G. Nosachev, "Language patterns discriminate mild depression from normal sadness and euthymic state," *Frontiers in psychiatry*, vol. 9, p. 105, 2018.

[138] Y. Suhara, Y. Xu, and A. Pentland, "Deepmood: Forecasting depressed mood based on self-reported histories via recurrent neural networks," in *Proceedings of the 26th International Conference on World Wide Web*, International World Wide Web Conferences Steering Committee, 2017, pp. 715–724.

[139] *Suicide*. [Online]. Available: `https://www.who.int/news-room/fact-sheets/detail/suicide`.

[140] S. Sun, "Meta-analysis of cohen's kappa," *Health Services and Outcomes Research Methodology*, vol. 11, no. 3, pp. 145–163, 2011.

[141] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[142] L. Townsend and C. Wallace, "Social media research: A guide to ethics," *Aberdeen: University of Aberdeen*, 2016.

[143] M. Trotzek, S. Koitka, and C. M. Friedrich, "Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences," *IEEE Transactions on Knowledge and Data Engineering*, 2018.

[144] S. Tsugawa, Y. Mogi, Y. Kikuchi, *et al.*, "On estimating depressive tendencies of twitter users utilizing their tweet data," in *2013 IEEE Virtual Reality (VR)*, IEEE, 2013, pp. 1–4.

[145] *Twitter by the Numbers (2022): Stats, Demographics Fun Facts — omnicoreagency.com*, `https://www.omnicoreagency.com/twitter-statistics/`, [Accessed 12-Sep-2022].

[146] *Validity and reliability in quantitative studies — ebn.bmj.com*, `https://ebn.bmj.com/content/18/3/66`, [Accessed 12-Sep-2022].

[147] M. Valstar, B. Schuller, K. Smith, *et al.*, "Avec 2013: The continuous audio/visual emotion and depression recognition challenge," in *Proceedings of the 3rd ACM international workshop on Audio/visual emotion challenge*, ACM, 2013, pp. 3–10.

[148] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[149] M. J. Vioulès, B. Moulahi, J. Azé, and S. Bringay, "Detection of suicide-related posts in twitter data streams," *IBM Journal of Research and Development*, vol. 62, no. 1, pp. 7–1, 2018.

[150] P. S. Wang, M. Angermeyer, G. Borges, *et al.*, "Delay and failure in treatment seeking after first onset of mental disorders in the world health organization's world mental health survey initiative," *World psychiatry*, vol. 6, no. 3, p. 177, 2007.

[151] X. Wang, Y. Ren, and W. Zhang, “Depression disorder classification of fmri data using sparse low-rank functional brain network and graph-based features,” *Computational and mathematical methods in medicine*, vol. 2017, 2017.

[152] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. Schumm, “Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework,” *arXiv preprint arXiv:2011.06149*, 2020.

[153] A. Yadollahi, A. G. Shahraki, and O. R. Zaiane, “Current state of text sentiment analysis from opinion to emotion mining,” *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, pp. 1–33, 2017.

[154] A. Yates, A. Cohan, and N. Goharian, “Depression and self-harm risk assessment in online forums,” *arXiv preprint arXiv:1709.01848*, 2017.

[155] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, *et al.*, “Semi-supervised approach to monitoring clinical depressive symptoms in social media,” in *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2017*, ACM, 2017, pp. 1191–1198.

[156] W. Yin, J. Hay, and D. Roth, “Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach,” *arXiv preprint arXiv:1909.00161*, 2019.

[157] J. Zhao, T. Wang, M. Yatskar, R. Cotterell, V. Ordonez, and K.-W. Chang, “Gender bias in contextualized word embeddings,” *arXiv preprint arXiv:1904.03310*, 2019.

[158] H. Zogan, I. Razzak, X. Wang, S. Jameel, and G. Xu, “Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media,” *World Wide Web*, vol. 25, no. 1, pp. 281–304, 2022.

Appendix A

Background Material

Here we provide a concise overview on key concepts related to statistical analyses used in Psychometrics, ethics of extracting and summarizing online information, i.e., Social Media contents and clinical process for detecting Major Depressive Disorder (MDD). We also provide layman's details of some well known deep learning constructs we use in our research, such as Bidirectional Long Short Term Memory (BiLSTM), Self-attention, Word and Sentence Embedding, Bidirectional Encoder Representations from Transformers (BERT), Natural Language Inference (NLI) and Semi-supervised Learning (SSL).

A.1 Definition of Psychometrics

Psychometrics is the branch of Psychology which deals with measuring psychological constructs such as, personality, attitude, aptitude and intelligence [114]. ***Reliability*** and ***validity*** are the two main concepts used by the psychometric researchers and practitioners for demonstrating the effectiveness of different psychological tests/questionnaires/scales [113]. In a traditional machine learning sense, we can (almost) think of a depression questionnaire as a machine learned depression detection model and the data, i.e., the interviews from the patient as depression related data.

Reliability

Reliability is used to measure the consistency of a psychometric test. The main idea is to measure whether a particular test score is consistent while it is repeatedly used under conditions where test score supposed to be consistent [113]. For example, a questionnaire for screening depression is used for different depressed patients and the score is pretty consistently indicating depressive state of those patients, hence the depression questionnaire is reliable.

Types of Reliability: Psychologists consider mainly three types of reliability or consistency measures, such as (1) **Test-retest reliability** over time for a test (2) **Internal consistency** across (test) items and (3) **Inter-rater reliability** across different raters for a test [113]. **(1) Test-retest Reliability:** This reliability measurement finds out whether a particular scale is consistent when it is repeated over time . For example, a personality questionnaire score should be consistent overtime for a particular person if it is used by that person in coming weeks or months. **Test-retest correlation** is one way to measure whether a particular test's score is consistent over time. For this, ***Pearson's r*** or correlation score is widely used. Given a pair of random variable (X, Y), the Pearson's correlation factor can be calculated using the following equation: $\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$, where, covariance of X and Y, $cov(X, Y) = E[[X - \mu_X][Y - \mu_Y]]$. Here, μ_X, σ_X is the mean and standard deviation of X respectively, μ_Y, σ_Y is the mean and standard deviation of Y respectively and E is the expectation. The more the score the better. **(2) Internal consistency:** This reliability measure finds out whether the items in a test correlates each other, thus confirming they are identifying same underlying construct. For example, In a depression questionnaire, if the item (or question) scores are not correlated, then those items are not internally consistent. The most widely used measure of such reliability is called **split-half consistency**. In such a scenario, a depression questionnaire can be split half into set of items and the scores of these split halves are calculated. Later, **Pearson's r** is calculated from those sets to see if they are correlated. Another widely used measure is called **Cronbach's** α. **Cronbach's** α is used to find out the mean of all possible split halves among set of items or questions of a questionnaire. The higher the score the better. **(3) Inter-rater reliability:**

This refers to the consistency among the raters for their judgment. For example, to measure whether an assessment criteria for depression is consistent over different clinicians, it is used by different clinicians for the same patient. If the score is different for each clinicians then the test has low inter-rater reliability, may be due to the fact that the assessment criteria are too subjective. Cronbach's α is used for finding out inter-rater reliability when the judgments of clinicians are continuous values and **Cohen's** κ is used when it is categorical.

Validity

Validity refers to how accurately a psychometric test measures what it supposed to measure. For example, a depression questionnaire is not valid if its score is highly correlated to the patient's ability to memorize things. Although depression may cause memory problems, the said questionnaire could be valid for memory test, not for depression screening.

Types of Validity There are three types of validity, (1) **Construct**, (2) **Content** and (3) **Criterion** [113]. **(1) Construct Validity:** This refers to, how much a test adheres to the existing theory and knowledge. For example, a depression questionnaire can be assessed through measuring other traits related to depression as established by earlier research, such as consistent Anhedonia, sad mood and psychomotor retardation. **(2) Content Validity:** This refers to the extent to which the test covers all aspects of the concept being measured. For example, if Anhedonia, sad mood and psychomotor retardation are the integral part of depression identification, then lacking any of them can make the depression questionnaire invalid. **(3) Criterion Validity:** This measures the extent to which a test is correlated to the other valid measures of the same concept. For example, given work performance degrades with depression, if depression questionnaire score is negatively correlated with work performance, then the depression questionnaire can be regarded as criterion valid, here the criterion is work performance.

Relation Between Validity and Reliability

A reliable test can be thought of as valid, although it is not always true. For example, a depression questionnaire can be reliable because it measures anxiety pretty much consistently over different subjects, although it's not valid, because it measures anxiety instead of depression.

Ensuring Reliability

Reliability can be ensured in the following ways [119], **(1) Method for data collection should be consistent:** It is to be ensured that, the clinicians who use the questionnaire to detect depression, should look for same criteria of depression across different patients. **(2) Standardizing the condition of the research:** It is to be ensured that the depression (interview) data was collected from the patients under same condition and they were given same information about the questionnaire.

Ensuring Validity

Validity can be ensured in following ways [119]: **(1) Choosing appropriate methods of measurement:** The measurement techniques should be of high quality and supported by established research. For example, a depression questionnaire should contain clinically established symptoms of depression. **(2) Using appropriate sampling method for selecting the subject:** The population from where the subject is sampled should be properly defined (i.e., the age group, gender, ethnicity etc). Also, It has to be ensured that the sample is appropriate representative of the population and big enough to infer something statistically significant.

Basics of Accuracy Measures

To find out whether a particular psychometric test is reliable and valid, its accuracy should be calculated. Higher accuracy means the test is measuring a particular concept (e.g. depression) more accurately. We discuss accuracy measures in the light of the following confusion matrix, where, True Positive = TP, False Positive = FP, False Negative = FN and True Negative = TN.

		True diagnosis		
		Positive	Negative	Total
Screening test	Positive	TP	FP (Type-1 error)	$TP + FP$
	Negative	FN (Type-2 error)	TN	$FN + TN$
	Total	$TP + FN$	$FP + TN$	N

Table A.1: Confusion matrix.

From A.1, we can define all the positive samples, $Pos = TP + FN$ and all the negative samples, $Neg = FP + TN$ in a True diagnosis set. So, total number of samples, $N = Pos + Neg$. Hence simple accuracy can be measured as, $acc = \frac{TP+TN}{TP+TN+FP+FN}$. However, accuracy can be susceptible of data imbalance, i.e., blindly predicting majority class can result in high accuracy. To avoid this there are other measures, such as balanced accuracy, $b_{acc} = \frac{TPR+TNR}{2}$, where, True Positive Rate, $TPR = \frac{TP}{Pos}$ and True Negative Rate, $TNR = \frac{TN}{Neg}$. It is to be noted, that TPR is also called, sensitivity, recall and hit rate and TNR is also called, specificity or selectivity. In clinical field, clinicians are more interested about the sensitivity and specificity of a test. A test with higher sensitivity means it has high false positive score, i.e., it does not miss any true diagnosis, however, some or many of its detected positive case may not be actually positive case. On the other hand, a highly specific test detects positive cases with high accuracy, i.e., the detected positive case is indeed positive, however, it may miss some real positive cases. There are another measure widely used in computing science which is called, Precision or Positive Predictive Value (PPV), $Prec = \frac{TP}{TP+FP}$ and Recall, $Rec = \frac{TP}{TP+FN}$, and their harmonic mean, $F - score = \frac{(1+\beta^2)\times Prec\times Rec}{\beta^2\times Prec+Rec}$. When $\beta = 1$, it is called F1-score. Although, F1-score is widely used and is proof to data imbalance, it puts equal weights to precision and recall, which may not be ideal, e.g, if a clinician is more interested on recall, rather than precision, s/he can put more emphasis on it, by modifying the formula for F-score, such that, $\beta = 2$. In multi-label classification settings two widely used accuracy measures are (1) **Macro-F1:** this is an average F1 score for all the labels and (2) **Weighted-F1:** this is a measure which assigns more weight to the labels for which we have majority samples[1]. Macro-F1 provides

[1] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

an idea of a multi-label classifiers performance in all labels regardless of the number of samples under those labels, i.e., it does not take data imbalance into account. On the other hand Weighted-F1 can be thought as a weighted average of label wise F1-scores; it puts more weight to labels with majority samples. Weighted-F1 provides an idea of a classifiers performance when there is a data imbalance.

Statistical Significance Test:

A result is statistically significant if the probability of its occurrence just by chance is less than a certain threshold. This probability is called p-value and the threshold is often set to be ≤ 0.05 for a result to be deemed as statistically significant. To compare between two machine learning models, mostly used statistical significance test is two-tailed paired t-test for different fold in k-fold cross validation [31]. Provided enough data and less processing overhead, there are some suggestions of using non-parametric methods or 5×2 cross validation for evaluation of machine learning models. However, with small test set and deep learning models these suggestions are not so feasible to follow [31].

Bayes' Theorem and Base Rate:

According to Bayes' theorem, probability of an event can be estimated from the prior knowledge about the conditions related to the event. For example, given a depression questionnaire score is positive for a random person, what is the probability that the person is depressed (D)? given, the depression questionnaire is 99% sensitive (i.e., 99% true positive for depressed person) and 80% specific (i.e., 80% true negative for non-depressed ($\neg D$) person) and the **base rate** or the proportion of depressed people in the population from where that random person is sampled is 10%. According to Bayes' theorem,

$$\begin{aligned} P(D|+ve) &= \frac{P(+ve|D)P(D)}{P(+ve)} \\ &= \frac{P(+ve|D)P(D)}{P(+ve|D)P(D) + P(+ve|\neg D)P(\neg D)} \\ &= \frac{0.99 \times 0.1}{0.99 \times 0.1 + 0.2 \times 0.90} \\ &= 35.4\% \end{aligned} \tag{A.1}$$

From the above calculation, we can see that although the test is highly sensitive and specific, the probability of an user being depressed given the test is positive is low, because the **base rate** is low. It is to be noted that, we can adjust the threshold of the questionnaire for depression screening, so that we have acceptable sensitivity and specificity [75].

A.2 Definition of Ethics

Ethics (also ‘moral philosophy’), refers to the branch of philosophy, which deals with “systematizing, defending and recommending concepts of right and wrong behavior.” [52].

Categories of Ethical Theory

Modern day philosophers divide ethical theories into three areas, such as (1) **Meta-ethics**, (2) **Normative Ethics**, and (3) **Applied Ethics** [52]. **Meta-ethics:** this area is concerned about the questions regarding the source of our ethical principles and their meaning. It deals with the questions, such as “what is the meaning of moral judgement? (e.g, what do the words ‘right’ and ‘wrong’ mean?),” “what is the nature of moral judgments? (i.e., is it universal or relative?),” and “how may the moral judgments be supported (i.e., how can we know that something is right or wrong?).” **Normative Ethics:** this category is concerned about articulating behaviors that we should acquire, the duties we should follow and the consequences of our behaviour towards others. **Applied Ethics:** this category is concerned about what is permitted for a person in a specific context or situation. In the next sections, we will mainly discuss about Applied Ethics in terms of Internet Research, more specifically social media research. We will also discuss briefly what ethically aligned design means and the key principles of ethically aligned design in autonomous and intelligent systems (A/IS).

Ethics in Internet Research

Internet has become an integral part of our life and playing important role as a ground for research, which is commonly referred as Internet Research (IR). IR can be used as an umbrella term, encompassing all the related research that use “innumerable technologies, devices, capacities, uses, and social spaces” came into the existence because of the internet [71]. Since many ethical questions can arise in the context of IR, it’s important to provide the scope of IR as follows: (1) Utilizing internet for collecting data, for example, through online surveys, interviews and automated means for data scraping. (2) Studying people’s online behavior through access patterns and usage of different online environments including, websites, blogs and social media. (3) Engaging in the process of storage and analysis of data-sets curated from the Internet. (4) Studying software, code and internet technologies. (5) Studying the design and structure of the internet technologies. (6) Utilizing various forms of analysis, such as semiotic, content and textual analysis of internet facilitated images, text and writings etc and (7) Studying large scale production, usage and regulation of internet by the governments, industries, corporation and military forces.

Principles of research ethics and ethical treatment to persons have been laid out in number of policies and documents, including, *UN Declaration of Human Rights*, the *Nuremberg Code*, the *Declaration of Helsinki* and the *Belmont Report*. Although, these basic principles largely originated from bio-medical contexts, they can be generalized into IR. Also, these basic principles echo the same idea, i.e., “respect for persons, justice and beneficience”. The key guiding principles fundamental to ethical approach of IR are as follows: (1) The obligation of the researchers to protect a community/author/participant depends on the vulnerability of the same, i.e., the more the vulnerability, the greater the protection required. (2) One of the important reasons for adopting ethical approach in IR is the minimization of harms. It is to be noted that the definition of ‘harm’ depends on the context, thus, ethical decision making is best made through the application of practical judgement that pays attention to specific context. (3) Although the involvement of human subjects are not immediately apparent, ethical principles regarding them should be considered, because, all digital information at some point involves individual humans. (4) Rights of research subjects should be balanced with the benefits of research and researchers' right to conduct research. In many cases, the rights of subjects may surpass the benefits of the research. (5) Ethical principles should be obeyed in all the steps of research, starting from planning, research conduct, publication and dissemination and (6) Since ethical decision making is a deliberative process, it should be done in consultation with as many people and resources as possible, including, people familiar with the research and/or participating in the research, fellow researchers, research review boards, etc. There are major tensions in confirming ethical concepts in three areas in IR, such as (1) **Human subjects:** the definition of human subject in not very clear in the context of IR as it is in biomedical research. (2) **Public/Private:** The definition of privacy is ambiguous in the context of IR. People may operate in public spaces but have strong expectations for privacy. (3) **Data(Text)/Persons:** In Internet it is tough to define the concept of 'personhood', i.e., whether ones digital information represents themselves. There is no straight forward answer to this question. Thus, the answer should be determined by rigorous discussion and follow the concept of "contextual integrity" as defined by *"Nissenbaum"* [98]. In the next sections, we provide few scenarios of contextual ethical principles in terms of Social Media Research which is a subset of IR.

Ethical Framework for Social Media Research

An ethical framework has been proposed by Townsend et al. [142] and Benton et al. [8] for social media research that take into account the key areas of the same, such as (1) **Data Privacy:** What data is considered public or private in the context of social media? (2) **Informed Consent:** When to seek for informed consent and how? What components should be present in informed consent? (3) **Anonymity:** When data needs to be anonymized? and (4) **Risk of Harm:** Is it possible to use the data to harm the research participants, i.e., using an user's personal information to embarrass them or damage their reputation or even prosecute them? The framework stands upon three main components as follows: (1) **Getting Aware of Terms, Conditions and Legalities:** Researcher should be aware of the most recent legal terms and conditions of the social media platform before starting their research. Also, researchers should ensure that their research is compliant with the research organization they are the part of, funding bodies and any disciplinary guidance provided by relevant disciplinary bodies. (2) **Privacy and Risk:** To determine whether social media data is public or private boils down to the fact that whether or not the social media user would like to be observed by the stranger [142]. The points to consider here are as follows: (1) The data researchers would like to access, is it located in a closed group/private group such as Facebook or in a open forum like, Twitter? (1.a.) Also, is it from a password protected forum, like online forums which require users' to open a password protected account? (1.b.) Is there any gate keeper in the group to whom researchers can turn to for approval or advice? (1.c.) Is there any security setting for the users of the forums? (2) Is the data potentially sensitive? (2.a.) Would the users of the group expect to share their data with the visitors who have interests and issues similar to themselves? (3) Is the data coming from an age group who are not aware of the risk of harm of social media? If the data is coming from closed/private/password protected groups, then it has more ethical concerns than the same from public/open forums. If there is a gatekeeper/admin of a password protected forum, researchers should consult him to find the best way to use the user data of that forum. In such case, the admin can assist researchers to seek consent from the users. Also, the admin can provide the capability to the users so that they can opt-out from the research whenever they want. If researchers would like to ask questions to the forum users about their social media behavior, s/he should disclose their identity to the users participating in the research. Also, due to the sensitive nature of the data, if possible informed consent from the users whose data is used, should be taken. If the data is coming from young (such as children) or vulnerable group who cannot provide informed consent and do not have clear idea about the risk of harm stemmed from social media, should be eliminated from the research. Overall, if the benefit of a research outweighs the risk of harm, researchers may get an exemption from following these privacy policies. (3) **Reuse and Republication:** For publishing results of a research or share data for the sake of better reproducibility of the same, researchers should keep in mind about the sensitivity of the data. If the data is sensitive, researchers may seek informed consent (for the dissemination of the users' data to public) and/or try various ways of anonymization so that the users are untraceable. Moreover, it has been observed that, generally aggregated user data (as opposed to individual user data), and data published by organizations or public figures or hash-tagged social media posts can be used without the consent of its author because the former prevent from identifying individuals and the latter are intended for public views. Also, researchers should be aware about the privacy conditions (of the platform from which the data is being collected), whether the data is allowed to reuse and republish. There is another crucial ethical question about "whether a social media post still remains public if the user later deletes it?". In case of Twitter (and other social media which support this), one can share dehydrated Tweets (i.e only Tweet IDs), so that those can be downloaded later by the other parties, preventing the interested parties from downloading deleted Tweets.

Ethically Aligned Design for Autonomous and Intelligent Systems (A/IS)

IEEE has published a guidance for designing intelligent agents [132] (eg. tool for depression modelling), so that the intelligent agents align with overall well being of the human society and environment and reach beyond mere functional goals they suppose to achieve, which they name as Ethically Aligned Designs (EAD). The general principles of EAD are as follows: (1) **Human Rights:** A/IS should protect human rights which have international recognition. (2) **Well-being:** A/IS developers should focus on the betterment of human well-being as their main success criterion. (3) **Data Agency:** Better data accessibility and security should be facilitated to the users of A/IS. (4) **Effectiveness:** A/IS should be very clear about its functionality, whether its suitable for performing something its intended to do.(5) **Accountability:** The decisions of an A/IS system should be explainable and unambiguous, for example an user should be able to clearly understand the rationale behind a decision made by such system. (6) **Awareness of Misuse:** A/IS systems should be protected against any kind of misuse and risk related to it. (7) **Competence:** The safe and effective operations of A/IS systems should be ensured both by the A/IS creators through properly specifying it and users properly following them and (8) **Transparency:** A particular decision made by an A/IS system should be easily found.

A.3 Definition of Major Depressive Disorder or Depression

Depression also known as Major Depressive Disorder (MDD) is a common and serious mood disorder. Depression can be characterized by persistent feelings of sadness and hopelessness and/or losing interest in day-to-day activities which were once enjoyable. Apart from emotional problems, depression can lead to various physical problems that can reduce a person's ability to function at work and at home [29].

Clinical Process of Depression Diagnosis

The clinical process of depression diagnosis starts with depression screening tools. If a patient is screened positive for depression, s/he is interviewed by clinicians to find out if s/he meets the criteria for depression diagnosis according to a widely known Diagnostic and Statistical Manual for Mental Disorders (DSM), version 5 [72].

Depression Screening Tools

There are several screening tools that are used for depression screening [129], such as **Hamilton Depression Rating Scale (HDRS):** HDRS is an interview scale, developed in 1960 to measure severity of depression in an inpatient population. It's a 21 item scale with 5 points for each item indicating the severity of depression symptoms. Items 18-21 are used for further qualifying depression. Scores 0-7 are considered normal while score greater than 20 indicate moderate to severe depression. This scale contains comparatively large numbers of somatic symptoms and few cognitive or affective symptoms. It has sensitivity of 86.4% and specifity of 92.2% , **Beck Depression Inventory (BDI):** BDI is a self-rating scale. It contains

21 items of emotional, behavioral and somatic symptoms and takes 5-10 minutes to administer. The items are scored within 0-3 indicating severity of symptoms. Scores, 10-18 indicates mild depression, 19-29 indicates moderate and greater than 30 indicates severe depression. One study has found that, BDI has 97% sensitivity and 99% specificity for identifying depression. **Patient Health Questionnaire (PHQ):** The Patient Health Questionnaire, PHQ is a self-administered tool of 2 (for PHQ-2) and 9 (for PHQ-9) items. PHQ-2 is a screening tool that assesses the frequency of depressive mood and Anhedonia for past two weeks and has scoring between 0 (not at all) to 3 (nearly every day). A PHQ-2 score greater than 3 has 83% sensitivity and 92% specificity for major depression. The cut-point of PHQ-9 is greater or equal to 10, which has sensitivity and specificity of 88%. PHQ-9 scores, 5, 10, 15 and 20 are representative of mild, moderate, moderately severe and severe depression. **Major Depression Inventory (MDI):** MDI is a self rating scale used for diagnosis or measurement of depression according to both DSM-4 criteria of major depression and ICD-10 moderate to severe depression. The symptoms should be present nearly everyday for last two weeks. It's a 10 point scale with each item can have values between (0-5). This scale has more emphasis on depressed mood and lack of interest. The cutoff score is 26. The sensitivity of MDI is between 86% and 92% and specificity is between 82% and 86%. **Center for Epidemiological Studies Depression Scale (CES-D):** This was published in 1977 as a screening tool for depression in general population. This scale has 20 items with 16 negatively worded and 4 positively worded. This instrument measures affective and somatic aspects of depression. Each question/item receives score between 0 to 3 and possible range of score is between 0 to 60 with 22 as a cut-off point. **Montgomery Asberg Depression Rating Scale (MADRS):** This rating scale is used by clinicians to rate the severity of depression among the depression diagnosed patients. It is designed to be sensitive to the change as a result of antidepressant therapy. This scale was developed in 1979 by Montgomery and Asberg. It has 10 items related to symptoms of depression. Each item has a severity scale from 0 to 6 and can be added to form a over-all score from 0 to 60. At a cut-off of > 6, the MADRS had a sensitivity of 0.90 and a specificity of 0.66. At a cut-off of

> 12, a sensitivity of 0.70 and a specificity of 0.86 was found. All cut-offs lower than 9 yielded sensitivities > 0.80 and specificities > 0.60. Of these, a cut-off of > 8 had the highest overall agreement (0.74), kappa (0.40), and positive predictive value (0.41). The AUC for the MADRS was excellent (AUC = 0.91) [82]. **Zung Self-Rating Depression Scale (SDS):** It's a self-administered depression scale and is a 20 item scale published in 1965. Half of the 20 items are positively worded and half are negatively worded. The items has scores between 1-4. Scores greater than 50 indicate mild depression, greater than 60 indicate moderate and 70 indicate severe depression. **Geriatric Depression Scale (GDS):** Originally developed for screening depression in geriatric population. It's a 30 item scale but later modified to 15 items. The questions are of "yes" or "no" nature for making it easy for older population. There is also a five point scale which is better received by older population. This 5-item scale has sensitivity of 94% and specificity of 81%, **Cornell Scale for Depression in Dementia (CSDD):** This scale is specially designed for the Dementia patients. Since, Dementia patients can provide unreliable answers to the questionnaire, it requires additional information from patient informant, i.e. someone who knows the patient. It's a 19 points scale where each item has scores between 0 (for absent) to 2(for severe). A total score of 10 indicates probable major depression, greater than 18 indicates definite major depression. According to a recent study, a score more than 6 has sensitivity of 93% and specificity of 97%. Most of the depression screening tools have been made keeping in mind the major symptoms of depression as specified in DSM-4 and more recently DSM-5.

Diagnostic and Statistical Manual for Mental Disorders (DSM) criteria for Depression Diagnosis

DSM is the handbook used by the clinicians around the world as the authoritative guide for diagnosing mental disorders. DSM contains description, symptoms and other criteria for diagnosing mental disorder. It provides a common platform for the clinicians around the world to communicate with their patients and establish reliable and consistent diagnosis of mental disorders [34]. According to recent version of DSM (i.e., DSM-5), to diagnose someone as depressed, one should have five or more symptoms related to depression during the past two weeks, most of the day or nearly everyday, and represent a change from previous functioning as indicated by either subjective report or observation made by others. At-least one symptom have to be either: (A) depressed mood or (B) loss of interest or pleasure. Symptoms due to general medical conditions or mood in-congruent delusions/hallucinations should not be included. The full list of MDD symptoms are as follows: (1): Depressed mood. For children and adolescents it can be irritable mood. (2) Markedly diminished interest or pleasure. (3) Significant weight loss. (4) Insomnia / Hypersomnia. (5) Psycho-motor agitation. (6) Fatigue. (7) Feeling of worthlessness. (8) Diminished ability to think or concentrate or indecisiveness and (9) Recurrent thoughts of death, suicidal ideation without a specific plan or a suicide attempt. Also, the symptoms (1) should not meet criteria for mixed (Bipolar) episodes (2) should cause clinically significant distress or impairment in social, occupational or other type of functioning and (3) the symptoms are not induced by substance abuse or medication of any general medical condition and (4) the symptoms are not better accounted for by bereavement. The medical professional/ clinician or health care provider thus conducts various physical test and takes interview of the patient to rule out whether the depression symptoms are caused by any other underlying medical/physiological/drug related/psychological reasons [72].

DSM vs. ICD on MDD Diagnosis

DSM and International Classification of Diseases (ICD) both contain the depression symptomatology, helpful for diagnosing depression. ICD is developed by an international body of clinicians appointed by WHO, where DSM is developed by American Psychiatric Association. ICD is widely used in European countries, while DSM is used mostly in North and South America and some Asian countries [16]. Although DSM uses the term Major Depressive Disorder (MDD), ICD does not and it refers MDD as Depressive Disorder. Moreover, ICD contains only guidelines for depression diagnosis and relies more on clinical judgment for depression diagnosis unlike DSM, which contains criteria for the depression diagnosis. DSM also includes sub-types of MDD which ICD does not have. However, both of these manuals acknowledge the fact that for diagnosing depression there is a basic set of symptomatology that needs to be confirmed in a patient. Both of these manuals recently added different cultural variables in MDD symptomatology [104], since, people from some cultures may mask their symptoms of depression and express them through somatic terms, such as “feeling of imbalance” or “ailment in their heart”.

A.4 Semi-supervised Learning (SSL)

SSL is a branch of machine learning, where a learned model is used to label samples from large unlabelled samples. Later, newly labelled samples along with original training samples are used to re-train the initial learned model and improve it on a iterative basis. There are few variations of SSL, among them two main methods are (1) self-training: In this method usually most confident predictions of a model are used to augment the existing training set and (2) co-training: In this method, two or more models each having multiple view of the data are used to label unlabelled samples and use those to augment existing training samples [73].

A.5 Deep Learning Models

Deep learning is a branch of machine learning which performs superior in many supervised learning tasks, including language related tasks, such as, text classification (e.g., sentiment detection) and sequence to sequence modelling (e.g., machine translation). Here we discuss the basic deep learning models which we use to develop our clinical depression model.

Bidirectional Long Short Term Memory (BiLSTM)

BiLSTM is a kind of Recurrent Neural Network (RNN). RNN is a variant of feed-forward neural network which is capable of handling sequence data of varying lengths and each node of this network represents each components of the sequence. It has a memory unit that can retain information from earlier portion/nodes of the input. However for very long sequence, due to vanishing gradient problem, RNN cannot retain early information of the sequence efficiently. Variants of RNN, such as LSTM [45] has been proposed to alleviate this problem. A BiLSTM is basically a bidirectional LSTM that consist of a forward LSTM and a backward LSTM as depicted in the Figure 3 [24]. LSTM consists of gating mechanism that enables it to retain or forget portion of the input through a memory cell, C (Figure 2 [24]). There are three kinds of gating mechanisms in an LSTM, such as: (1) Forget (f) (2) Input (i) and (3) Output gate (o). Forget gate ensures whether to retain previous information through previous cell state C_{t-1} of a sequence through a sigmoid activation function, input gate then passes current input (x_t, h_{t-1}) and decides whether to pass it or not through a sigmoid activation function and the output of this sigmoid function is multiplied by current input through tanh activation function to candidate cell state to $\tilde{C}_t$. Finally, through the output gate, output hidden state h_t is formed with the help of current input (x_t, h_{t-1}) and final cell state C_t. These two values are passed to the next LSTM node.

Self-attention

Self-attention is a main mechanism of transformer models [148]. The core idea of Self-attention is to learn the contribution of important components of a sequence while optimizing a particular learning objective. It is done through finding out each components of a sequence and how they interact with others including itself. To calculate this interaction, three vectors, such as Query (Q), Key (K) and Value (V) are calculated for each component. Later, a dot product of Q and K for each component is calculated; this dot product is scaled by dividing it with $\sqrt{d_k}$, where d_k is the dimension of the key, softmaxed and later point wise multiplied by V. Later, the output is calculated by adding weighted Vs for each component for that Q. This process is repeated for other components which results in component vectors with their importance for the learning task. Equation A.2 provides the details of scaled-dot-product Attention. Also, Figure 2 [148] provides details of Self-attention and multi-head Attention mechanism. Multi-head Attention allows for attending different representation subspaces at different positions of the input.

$$Self - attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \tag{A.2}$$

BERT

BERT is a pre-trained language representation model with stack of 12 transformer-encoder blocks with 12 multi-headed Self-attention layers. An illustration of BERT model is provided in Figure 3 [30] (extreme left). BERT is pre-trained on Book Corpus and English Wikipedia text and through Masked Language Model and Next Sentence Prediction tasks. BERT based models are found out to be outperforming many traditional machine learning models in various downstream Natural Language Processing (NLP) tasks. There are different variations of BERT, among them one of the most well known BERT model is **A Robustly Optimized BERT Pretraining Approach (RoBERTa)** proposed by Liu et al. [66]. They showed that BERT's performance can be improve even more through careful fine-tuning during its training, i.e., learning on large batches and longer sequences as well as removing next sentence prediction objective.

Natural Language Inference (NLI)

Natural Language Inference (NLI) is a branch of NLP field, where, a classifier predicts whether a possible hypothesis for a given premise entails it or contradicts it and their probabilities. For example, provided the following tweet as premise: "This is a wonderful world", an user's sentiment "satisfied" is entailed. This entailment and contradiction can be expressed as probability scores for the label "satisfied". **BART** [63] is a pre-trained model which is found out to be performing superior in NLI task based on Zero-Shot Learning setting proposed by Yin et al. [156]. BART is a denoising autoencoder which uses a transformer-based neural machine learning architecture, where it has a bidirectional encoder like BERT and left-to-right decoder like GPT, pre-trained on two tasks: (1) corrupting text with arbitrary noise function and (2) denoising the corrupted text to its original form.

Sentence Embedding

Sentence or contextual embedding model takes input a text and provides a fixed length vector that represent that text. Universal Sentence Encoder (USE) is one well known sentence embedding model [13].

There are two kinds of encoders used by USE, one is called Deep Averaging Network (DAN) encoder, which uses averaged bi-gram and word embedding of a sentence. Another is transformer encoder, which uses the context aware word representations of a sentence. Later, these word representations are converted to a fixed length sentence encoding vector by computing the element-wise sum of these representations. Further, both encoders send their respective representations to a feed-forward neural network, which is then multi-task trained on several downstream NLP tasks, such as: "a SkipThought [58] like task for the unsupervised learning from arbitrary running text; a conversational input-response task for the inclusion of parsed conversational data; and classification tasks for training on supervised data" [13]. Both of these encoders take a lower cased sentence further tokenized using PTB tokenizer and then produce a 512 dimensional sentence embedding. The data sources for pre-training these encoders are Wikipedia, web news, web question-answer pages and discussion forums.

It has been found out to be as good as other contemporary sentence embedding models such as: Embeddings from Language Model (ELMo) [111], which uses a bi-directional LSTM to compute contextualized character-based word representations. USE and ELMo are two prominent sentence embedding models before BERT came into existence. Later, BERT and its variation like SBERT [118] which is further fine-tuned in sentence similarity task based on triplet and siamese network loss objective has been found out to be providing state-of-the-art in many NLP downstream tasks.

Word Embedding

Word embedding is a technology which is based on the same underlying principle as word co-occurrence matrix for determining different semantic relationships among words. Instead of calculating a raw word-co-occurrence matrix, the word embedding is a elaboration that leverages a neural network which takes context words as inputs and the word for which context words were determined are used as output or prediction, also called, Contextual Bag of Words or CBOW method. There is a related method, that tries to predict the context words provided a particular word, and this method is called the skip-gram method. Skip-gram has been found to be very effective for such (Figure 2 [89]). In both of these methods, inputs and outputs are a one-hot vector representation of size equal to the total number of elements in the vocabulary. Mathematically, through this process, the neural network learns a hidden layer which is a matrix of dimension $V \times N$, where V is size of the vocabulary and N is the dimension of the embedding (Figure 2 [89]).

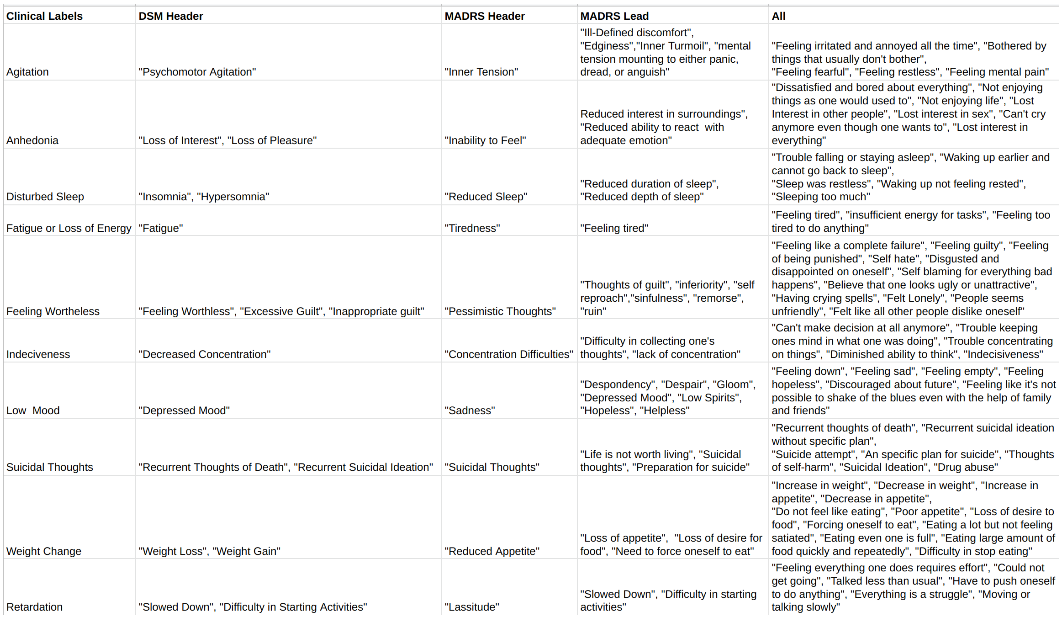
This embedding representation can be thought of as a compressed representation that represents the semantics of the word. The semantics of the word can be guessed through the analysis of its distance with other vectors. It has been shown in the earlier research, e.g., Levy et al. [62], that the hidden layer matrix mentioned earlier can be approximated through Singular Value Decomposition (SVD) of the PMI matrix, which looks like word co-occurrence matrix C, but, populated with point wise mutual information (PMI) for word and context word pairs. Glove [109] is an algorithm which uses similar ideas for creating word embedding representation as described by Levy et al. The loss function of the neural network for learning skip-gram word embedding and its description is provided as follows in Equation A.3 (adapted from [136]).

$$\mathcal{L} = -log \prod_{c=1}^{C} P(w_{c,j}|w_a) = -log \prod_{c=1}^{C} \frac{exp(O_{c,j*})}{\sum_{j=1}^{V} exp(O_{c,j})} \tag{A.3}$$

where, $w_{c,j}$ is the j-th word predicted on the c-th context position; w_a is the only input word; and $O_{c,j}$ is the j-th value in the O vector when predicting the word for the c-th context position.

A.6 Label Descriptors

Depression symptom labels and their corresponding descriptors, i.e., DSM Headers, MADRS Headers and Leads and All are provided in Figure A.1.



Clinical Labels	DSM Header	MADRS Header	MADRS Lead	All
Agitation	"Psychomotor Agitation"	"Inner Tension"	"Ill-Defined discomfort", "Edginess","Inner Turmoil", "mental tension mounting to either panic, dread, or anguish"	"Feeling irritated and annoyed all the time", "Bothered by things that usually don't bother", "Feeling fearful", "Feeling restless", "Feeling mental pain"
Anhedonia	"Loss of Interest", "Loss of Pleasure"	"Inability to Feel"	Reduced interest in surroundings", "Reduced ability to react with adequate emotion"	"Dissatisfied and bored about everything", "Not enjoying things as one would used to", "Not enjoying life", "Lost Interest in other people", "Lost interest in sex", "Can't cry anymore even though one wants to", "Lost interest in everything"
Disturbed Sleep	"Insomnia", "Hypersomnia"	"Reduced Sleep"	"Reduced duration of sleep", "Reduced depth of sleep"	"Trouble falling or staying asleep", "Waking up earlier and cannot go back to sleep", "Sleep was restless", "Waking up not feeling rested", "Sleeping too much"
Fatigue or Loss of Energy	"Fatigue"	"Tiredness"	"Feeling tired"	"Feeling tired", "insufficient energy for tasks", "Feeling too tired to do anything"
Feeling Wortheless	"Feeling Worthless", "Excessive Guilt", "Inappropriate guilt"	"Pessimistic Thoughts"	"Thoughts of guilt", "inferiority", "self reproach","sinfulness", "remorse", "ruin"	"Feeling like a complete failure", "Feeling guilty", "Feeling of being punished", "Self hate", "Disgusted and disappointed on oneself", "Self blaming for everything bad happens", "Believe that one looks ugly or unattractive", "Having crying spells", "Felt Lonely", "People seems unfriendly", "Felt like all other people dislike oneself"
Indeciveness	"Decreased Concentration"	"Concentration Difficulties"	"Difficulty in collecting one's thoughts", "lack of concentration"	"Can't make decision at all anymore", "Trouble keeping ones mind in what one was doing", "Trouble concentrating on things", "Diminished ability to think", "Indecisiveness"
Low Mood	"Depressed Mood"	"Sadness"	"Despondency", "Despair", "Gloom", "Depressed Mood", "Low Spirits", "Hopeless", "Helpless"	"Feeling down", "Feeling sad", "Feeling empty", "Feeling hopeless", "Discouraged about future", "Feeling like it's not possible to shake of the blues even with the help of family and friends"
Suicidal Thoughts	"Recurrent Thoughts of Death", "Recurrent Suicidal Ideation"	"Suicidal Thoughts"	"Life is not worth living", "Suicidal thoughts", "Preparation for suicide"	"Recurrent thoughts of death", "Recurrent suicidal ideation without specific plan", "Suicide attempt", "An specific plan for suicide", "Thoughts of self-harm", "Suicidal Ideation", "Drug abuse"
Weight Change	"Weight Loss", "Weight Gain"	"Reduced Appetite"	"Loss of appetite", "Loss of desire for food", "Need to force oneself to eat"	"Increase in weight", "Decrease in weight", "Increase in appetite", "Decrease in appetite", "Do not feel like eating", "Poor appetite", "Loss of desire to food", "Forcing oneself to eat", "Eating a lot but not feeling satiated", "Eating even one is full", "Eating large amount of food quickly and repeatedly", "Difficulty in stop eating"
Retardation	"Slowed Down", "Difficulty in Starting Activities"	"Lassitude"	"Slowed Down", "Difficulty in starting activities"	"Feeling everything one does requires effort", "Could not get going", "Talked less than usual", "Have to push oneself to do anything", "Everything is a struggle", "Moving or talking slowly"

Figure A.1: Label descriptors.

A.7 Annotation Guideline

Social Media Data Annotation by Human

For this annotation task, an annotator has to label or classify a social media post (i.e., a tweet) in one or more of the following depression symptom categories which suit best for that social media post through a web tool:

1. Inability to feel pleasure or Anhedonia
2. Low mood
3. Change in sleep pattern
4. Fatigue or loss of energy
5. Weight change or change in appetite

6. Feelings of worthlessness or excessive inappropriate guilt
7. Diminished ability to think or concentrate or indecisiveness
8. Psychomotor Agitation or Inner Tension
9. Psychomotor Retardation
10. Suicidal Thoughts or Self-Harm
11. Evidence of Clinical Depression
12. No evidence of Clinical Depression
13. Gibberish

Detailed description of these categories with examples are as follows:

The following sections need to be very carefully read to better understand what each category means. We divide the description under each category into three parts: “Lead”, “Elaboration”, and “Example”. “Lead” contains the summary or gist of the symptomatology. “Elaboration” provide a broader description of the symptomatology accompanied by a few relevant “Examples”. These sections have been developed with careful considerations of criteria defined in the DSM-5, MADRS, BDI, CES-D and PHQ-9 depression rating scales.

Depression Symptom Labels

1. **Inability to feel pleasure or anhedonia**

   (a) **Lead:** Subjective experience of reduced interest in the surroundings or activities, that normally give pleasure.

   (b) **Elaboration:** Dissatisfied and bored about everything. Not enjoying things as one would used to. Not enjoying life. Lost Interest in other people. Lost interest in sex. Can’t cry anymore even though one wants to.

   (c) **Example:**

        i. I feel numb.
        ii. I am dead inside.
        iii. I don't give a damn to anything anymore.

2. **Diminished ability to think or concentrate or indecisiveness**

    (a) **Lead:** Difficulties in collecting one's thoughts mounting to incapacitating lack of concentration.

    (b) **Elaboration:** Can't make decisions at all anymore. Trouble keeping one's mind on what one was doing. Trouble concentrating on things.

    (c) **Example:**

        i. I can't make up my mind these days.

3. **Change in sleep pattern**

    (a) **Lead:** Reduced duration or depth of sleep, or increased duration of sleep compared to one's normal pattern when well.

    (b) **Elaboration:** Trouble Falling or Staying Asleep. Waking up earlier and cannot go back to sleep. Sleep was restless (wake up not feeling rested). Sleeping too much.

    (c) **Example:**

        i. It's 3 am, and I am still awake.
        ii. I sleep all day!

4. **Fatigue or loss of energy**

    (a) **Lead:** Any physical manifestation of tiredness.

    (b) **Elaboration:** Elaboration: Feeling tired. Insufficient energy for tasks. Feeling too tired to do anything.

    (c) **Example:**

        i. I feel tired all day.
        ii. I feel sleepy all day.

iii. I get exhausted very easily.

5. **Feelings of worthlessness or excessive inappropriate guilt**

   (a) **Lead:** Representing thoughts of guilt, inferiority, self-reproach, sinfulness, and self-depreciation.

   (b) **Elaboration:** Feeling like a complete failure, Feeling guilty, Feeling of being punished. Self-hate. Disgusted and Disappointed on oneself. Self blaming for everything bad happens. Believe that one looks ugly or unattractive. Having crying spells. Feeling lonely. People seems unfriendly. Felt like all other people dislike oneself.

   (c) **Example:**

      i. Leave me alone, I want to go somewhere where there is no one.

      ii. I am so alone ...

      iii. Everything bad happens, happens because of me.

6. **Low mood**

   (a) **Lead:** Despondency, Gloom, Despair, Depressed Mood, Low Spirits, Feeling of being beyond help without hope.

   (b) **Elaboration:** Feeling down. Feeling sad. Discouraged about future. Hopelessness. Feeling like it's not possible to shake of the blues even with the help of family and friends.

   (c) **Example:**

      i. Life will never get any better.

      ii. I don't know why but I feel so empty.

      iii. I am so lost.

      iv. There is no hope to get out of this bad situation.

7. **Psychomotor agitation or inner tension**

   (a) **Lead:** Ill defined discomfort, edginess, inner-turmoil, mental tension mounting to either panic, dread or anguish.

(b) **Elaboration:** Feeling irritated and annoyed all the time. Bothered by things that usually don't bother. Feeling fearful. Feeling Restless. Feeling Mental Pain.

(c) **Example:**

i. It's my life so I decide what to do next, mind your own business, don't bother!

ii. You have no idea how much pain you gave me!

8. **Psychomotor retardation or lassitude**

(a) **Lead:** Difficulty getting started or slowness initiating and performing everyday activities.

(b) **Elaboration:** Feeling everything one do requires effort. Could not get going. Talked less than usual. Have to push oneself to do anything. Everything is a struggle. Moving or talking slowly.

(c) **Example:**

i. I don't feel like moving from the bed.

9. **Suicidal thoughts or self-harm**

(a) **Lead:** Feeling of Life is not worth living, suicidal thoughts, preparation for suicide.

(b) **Elaboration:** Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without specific plan, or suicide attempt, or a specific plan for suicide. Thoughts of self-harm. Suicidal ideation. Drug abuse.

(c) **Example:**

i. I want to leave for the good.

ii. 0 days clean.

10. **Weight change or change in appetite**

(a) **Lead:** Loss or gain of appetite or weight than usual.

(b) **Elaboration:** Increase in weight. Decrease in weight. Increase in appetite. Decrease in appetite. Do not feel like eating. Poor appetite. Loss of desire to food, forcing oneself to eat. Eating a lot but not feeling satiated. Eating even if one is full. Eating in large amount of food quickly and repeatedly. Difficulty in stop eating.

(c) **Example:**

i. I think I am over eating these days!

ii. I don't feel like eating anything!

11. **Evidence of clinical depression**

(a) **Elaboration:** Any social media post which do not necessarily fit into any of the above symptoms, however still carry signs of depression or representing many symptoms at a time, so it's very hard to fit it in a few symptoms.

(b) **Example:**

i. I feel like I am drowning . . .

12. **No evidence of clinical depression**

(a) **Elaboration:** Political stance or personal opinion, inspirational statement or advice, unsubstantiated claim or fact.

(b) **Example:**

i. People who eat dark chocolate are less likely to be depressed.

13. **Gibberish**

(a) **Elaboration:** If you are not sure what a social media post means i.e., if a social media post does not make sense or it's gibberish, then annotate it as Gibberish.

A.8 Depression Level Mapping

In the following Table A.2, we provide the mapping between the depression levels and corresponding range of depression scores. We use more stratification than the conventional PHQ-9 scale to get clearer idea about depression level distributions across our datasets.

Depression Level	Depression Score Range
None	$0- < 4$
Minimal	$4- < 9$
Mild	$9- < 14$
Moderate	$14- < 19$
Moderately Severe	$19- < 27$
Severe	$>= 27$

Table A.2: Depression Level Mapping Reference

A.9 Apriori Rules

Here we provide the strong rules mined from DSD-Clinician-Tweets-Original-Train (Table A.3)

Rules(Strong-Label $\rightarrow$ Weak-Label)
$1 \rightarrow 2$
$1 \rightarrow 6$
$4 \rightarrow 3$
$4 \rightarrow 8$
$4 \rightarrow 10$
$7 \rightarrow 6$
$7 \rightarrow 8$
$9 \rightarrow 6$
$9 \rightarrow 8$
$9 \rightarrow 10$

Table A.3: Strong Rules; indices for each labels are from Section A.7

.

A.10 Mental-BERT Training Configuration for DPD and DSD

Here we report the training configuration for Mental-BERT based DPD and DSD (Table A.4)

Hyperparameters	Values(DPD)	Values(DSD)
#Epochs	20	10
#Batch	32	Same
Max. sequence length	30	Same
Learning Rate	2×10^{-5}	Same
#GPUs	1	Same
Loss function	Binary Cross Entropy (BCE) Loss	Same

Table A.4: DPD and DSD model training parameters.

For DSD we use Binary Cross Entropy (BCE) Loss on the output of last layer of our Mental-BERT model which is based on sigmoid activation functions for each node corresponding to each depression symptom label. For DPD, we use BCE loss on the softmaxed output for each binary label, i.e., depression vs control. We do not freeze any layers in our fine-tuning process because it turned out to be detrimental to the model accuracy.

A.11 Temporal User-level Clinical Depression Detection (TUD) Model Training Configuration

Here we report the training configuration for TUD model (Table A.5)

Hyperparameters	Values
#Epochs	10
#Batch	16
LSTM Hidden-Dimension	100
#LSTM Hidden Layer	1
Drop-out	0.1
Learning Rate	1×10^{-3}
#GPUs	1
Loss function	Binary Cross Entropy (BCE) Loss

Table A.5: TUD model hyperparameters.

Since TUD is a binary classification task, we use the same settings for loss function as DPD described earlier.