ISLAMIC UNIVERSITY OF TECHNOLOGY

Detection of Severity of Depression from Social Media Data using Attention-Based Approach

By

MD. Mohsinul Kabir (181041021)

A thesis submitted in partial fulfilment of the requirements
for the degree of M.Sc. in Computer Science and Engineering

Academic Year: 2018-2019

Department of Computer Science and Engineering
Islamic University of Technology.
A Subsidiary Organ of the Organization of Islamic Cooperation.
Dhaka, Bangladesh.

May 2022

Declaration of Authorship

I, Md. Mohsinul Kabir, declare that this thesis titled, 'Detection of Severity of Depression from Social Media Data using Attention-Based Approach' and the work presented in it is my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Any part of this thesis has not been submitted for any other degree or qualification at this University or any other institution.
- Where I have consulted the published work of others, this is always clearly attributed.

Submitted By:

(Signature of the Candidate)
Md. Mohsinul Kabir- 181041021
April 2022

Detection of Severity of Depression from Social Media Data using Attention-Based Approach

Approved By:

Dr. Md. Kamrul Hasan
Thesis Supervisor,
Professor,
Department of Computer Science and Engineering,
Islamic University of Technology.

Dr. Md. Abu Raihan Mostofa Kamal
Head of the Department and Professor,
Department of Computer Science and Engineering,
Islamic University of Technology.

Dr. Hasan Mahmud
Assistant Professor,
Department of Computer Science and Engineering,
Islamic University of Technology.

Dr. Md. Saddam Hossain Mukta
Associate Professor,
Department of Computer Science and Engineering,
United International University (UIU), Dhaka, Bangladesh.

Abstract

In clinical psychology, diagnostic of mental illness is mostly done by patient's self-reported experiences, behaviors reported by the patients themselves, their relatives and a mental status examination. This method can lead to a variety of biases, such as cognitive bias, in which patients hide their illness for fear of judgement. Popular social networks can serve as a tool for dealing with this problem. But mental health research in this domain has been hindered by a lack of standard typology, scarcity of adequate data and lack of a robust classification network. In this thesis, the clinical articulation of depression is leveraged to build a typology for social media texts for detecting the severity of depression. It emulates the standard clinical assessment procedure Diagnostic and Statistical Manual of Mental Disorders (DSM-5) and Patient Health Questionnaire (PHQ-9) to encompass subtle indications of depressive disorders from social media texts. The typology has been developed with the association of two expert psychologists. To examine the typology, a dataset is constructed by scraping posts from Twitter, followed by a standard annotation method to label each tweet as 'non-depressed' or 'depressed', while three severity levels are considered for 'depressed' tweets: (1) mild, (2) moderate, and (3) severe. To classify severity of depression in this dataset, two attention-based models, namely BERT and DistilBERT are pre-trained and fine-tuned and a strong baseline result is provided. The findings of this study ought to provide strong directions for further research in this domain.

Keyword - Social Media; Mental Health; Depression Severity; Typology; Attention

Acknowledgements

I would like to express my whole-hearted gratitude to Allah Subhanu Wata'ala for giving me strength to complete this study, and being with me when none was there beside me. I would also like to express my grateful appreciation to Dr. Md. Kamrul Hasan and Dr. Hasan Mahmud for their constant motivation and support throughout this study.

This work is partially supported by

Systems and Software Lab (SSL)
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)

Contents

List of Figures

List of Tables

Dedicated to my parents and siblings for their lifelong dedicated support to my education ...

Chapter 1

Introduction

Depression is a major mood disorder and a serious medical illness that can negatively affect the way a person feels, thinks or acts. It can cause feelings of sadness and a loss of interest in daily chores. Depression decreases an individual's ability to function and can even lead to detrimental behavior like suicide. Moreover, the COVID-19 pandemic exerted a devastating impact on people's mental health [3]. Although depression is a common mental disorder, many people become aware of their depression only after experiencing significant functional deterioration [4]. The depressed individuals may not be aware of the symptoms of depression and may not care taking treatment at proper time as well. Sometimes not paying proper attention to this mental condition can lead to disastrous outcomes, including self-injurious behavior and suicide [5]. Identifying early signs of depression can prevent depressive disorder's negative consequences to a great extent [6].

There is a lack of reliable laboratory tests for most forms of mental illness, and in most cases, mental illness is diagnosed based on the patient's self-reported experiences, behaviors reported by relatives, and a mental status examination. Over the last decade, computational research devoted to modeling mental health phenomena using non-clinical data has grown at an exponential rate [7]. Studies analyzing web data, such as social media platforms and peer-to-peer messaging services, have piqued the interest of the research community due to their scope and deep entanglement in contemporary culture. Such research has provided novel insights into population-level mental health [8, 9] and demonstrated promising avenues for incorporating data-driven analyses into the treatment of psychiatric disorders. Social media platforms and other online discussion forums have been particularly appealing to the research community because of the massive scale of

data. This massive data flow has resulted from increasing rates of internet access and people spontaneously sharing their suffering, pain, and struggle anonymously [10] on these platforms. Recognizing the early symptoms of depressive disorder through a person's language use can prevent many disastrous outcomes like self-harm, suicide, etc., and can even help deploy effective treatment in proper time.

1.1 Problem Statement

Popular social networks can serve as a tool for detecting early symptoms of depression in user's behavior. Text messages published in these networks contain a lot of hidden information about their authors. Many researchers used social media to identify depressed users by analyzing the differences in language use. For example, Kim et al. [11] proposed a framework to classify specific mental disorder including depression, anxiety, bipolar, borderline personality disorder, schizophrenia, and autism from *Reddit* which can help identify potential sufferers with mental illness based on their posts. To extract data from social media that contains common depression symptoms, researchers utilized many established clinical assessment procedures like the PHQ-9 questionnaire [12]. Detection of depressive symptoms and providing first-line therapy online is known as Internet-delivered Psychological Treatment (IDPT) in literature. This IDPT approach has the potential to overcome mental distress for a large population with fewer resources. However, the main challenge of mental health research with social media data is the scarcity of large, shareable, and annotated datasets based on proper curation [13]. To be more specific, Harrigian et al. [13] discovered only a few available unique datasets in their study, that also rely on some form of self-reporting of the individuals but admittedly fail to meet ideal ground truth standards.

Another limitation in this domain of research is the lack of contextual consideration while classifying the texts with depressive symptoms, which in turn results in lower user adherence and incapability of the model to remember longer sequences [14]. For example, according to the PHQ-9 questionnaire, the severity of depression can differ from mild to suicidal. Psychologists have found symptoms of depression in users' social media posts and it can be utilized to diagnose depression and prevent destructive conclusion. But traditional deep learning models often fail to consider the full context of texts and thus lose important information in longer sequences. Attention-based mechanisms resolve this problem and have been proven to be

especially effective for capturing the context of documents in common natural language processing tasks.

This study focuses on the development of an annotated mental health dataset with ideal ground truth standards based on clinical validation to foster mental health research and an attention-based mechanism to classify the severity of depression from social media data.

1.2 Motivation & Scopes

Language is a major component of mental health assessment and treatment, and as such, it is a useful lens for mental health analysis. The psychology literature has a long history of studying the impact of various mental health conditions on a person's language use. The benefits of these computational approaches to understanding mental health state could be profound—for example, for new data to supplement clinical care, assessing developing conditions, identifying risky behaviors, providing timely interventions, or reaching populations that are difficult to reach through traditional clinical approaches. In fact, approaches like this have been adopted by platforms such as Facebook for suicide prevention efforts [15]. Complementary interest has emerged in an evolving field known as "digital psychiatry," [16], which uses these predictive signals to improve mental health service outcomes. Moreover, the ongoing outbreak of the COVID-19 pandemic is likely to have devastating impacts on the mental health of millions of individuals as lockdown in the affected areas has reported in high rises in the incident rates of mood disorder, including acute stress disorder, post-traumatic stress disorder, generalized anxiety disorder, and overall sub-clinical mental health deterioration [17]. The scope of mental health deterioration during the COVID-19 pandemic and the comprehensive nature of diagnosing depressive disorders have provided an unprecedented need to infer the mental states of individuals from all-inclusive resources. Recent studies have revealed that valuable insights into the impact of the pandemic on population-level mental health can be inferred from posts or comments on social media [3]. Reviews and meta-analyses have looked at the expression of depression and anxiety on social media, subjective mood, well-being, and mental health in social media and other non-clinical texts, and the development of technology in general for mental and affective health. Nevertheless, recent research has noted a lack of grounded recommendations detailing and evaluating

current practices for building algorithms to predict mental health state in social media data.

When it comes to classification models, effective context understanding from the input representations is very crucial to the task of severity of mental illness detection from social media texts. From this point of view, recent transformer-based models are likely to outperform traditional deep learning based models such as LSTM, BiLSTM or unidirectional transformer based models such as OpenAI GPT. For example, Ahmed et al. [18] showed in their study that a bidirectional Long Short-Term Memory (LSTM) architecture with an attention mechanism can achieve better result in a binary classification problem for the prediction of symptoms of depression. The attention mechanism looks at an input sequence and decides at each step which other parts of the sequence are important. In long sentences, this mechanism encodes each position of words which can relate two distant words of both the inputs and outputs with respect to itself and it can be parallelized to accelerate the training. In this mechanism, weights of each encoder state are preserved and not thrown away in later states, thus it solves the problem of loss of relevant information in long sentences, unlike the traditional RNNs.

The following points best summarize the motivation of this work-

- Firstly, text messages published in social media networks contain a lot of hidden information about their authors and there is a growing body of literature addressing the role of social networks on detecting the mental health state of their users.
- Secondly, researchers and practitioners are interested in using behavioral and linguistic cues from social media data to assess depressive symptomatology, such as self-harm, stress, and the severity of mental illness, without the use of in-person, clinical assessment.
- Finally, a curated dataset, consisting of large amount samples with ideal ground truth standards and an efficient attention-based method can foster mental health research and create many research scopes in this domain.

1.3 Research Challenges

A persistent challenge for the researchers specific to the mental health space is the need to: (a) establish a typology for text contents on social media to detect

the severity of mental illness with clinical validation and robustness [19], and (b) reliably apply this typology to obtain a sufficient sample size of high-quality data. Prior research has explored opportunities to capture mental health states from social media data using regular expressions to identify self-reported diagnosis or clustering activity patterns. However, deliberately relying on self-labeled data or unsupervised clustering leads to oversimplification and lacks clinical efficacy [19]. Practical exertion of mental health research includes identifying risky behaviors and providing timely interventions such as suicide prevention efforts adopted by Facebook [15]. The availability of high-quality, large-scale, annotated datasets addressing the severity of mental illness is one of the key elements for advancement on this front. Unfortunately, there are very few available datasets for depression severity which also lacks strong ground truths based on clinical validation [20].

1.4 Research Contribution

Considering all the limitations of the existing literature, this study addresses the problems in typology for extracting depression related contents from social media and in existing frameworks to detect severity of depression. Hence, in the first part of the study, a new typology to extract social media texts with depression severities is proposed with an attempt to mitigate all the limitations of the existing typologies. In the second part of this study, a new framework based on attention-mechanism is proposed to detect depression symptoms in social media posts. The principle contributions of this thesis can be summarized as below:

- **A new typology to extract social media texts with depressive symptoms:** Leveraging one of the established clinical methodology to diagnose depression, PHQ-9 [21], and expert opinions from two clinical psychologists, a new typology is established in this study for social media contents (i.e., tweet text) built upon a psychological theory for detecting the severity of the mental condition of depressed individuals. The procedure used to assess the severity of depression in this study was based on a well-established clinical assessment method known as the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [22], and it is carried out under the supervision of two expert clinical psychologists. The approach utilized in this study can be adopted to generate high-quality mental health data from various platforms in future investigations.

- **A unique dataset of labeled tweets based on strong ground truths and clinical validation:** A new dataset has been proposed in this study along with the typology to help alleviate the scarcity of data in mental health research. The labeling typology of the dataset assigns a higher-level classification to each tweet, such as (1) Non-depressed, (2) Mildly Depressed, (3) Moderately Depressed, and (4) Severely Depressed. There is also an associated confidence score (between 0.5 and 1) for each label. Standard data collection and annotation method is carried out while constructing this dataset.

- **An attention-based framework to detect depression severities:** Effective contextual consideration is very crucial when it comes to detect subtle nuances in different different severities. The existing frameworks mostly works with handpicked features and do not preserve the context of texts in long sequences. In this study, the capabilities of modern transformers have been explored, which utilize the attention mechanism in two forms: self-attention and multi-head attention. Two modern transformer based models, namely BERT and DistilBERT, is used to propose a framework for detecting depression severities in social media texts.

1.5 Thesis Outline

In Chapter 1 the objective of the study has been discussed in a concise manner. Chapter 2 deals with the necessary background & literature review for this study. In Chapter 3, the proposed methodology, data collection and annotation procedure, and the quality control mechanism have been discussed in detail. The summary statistic of the constructed dataset for this study is described in Chapter 4. The baseline classification model for this dataset and evaluation metrics are presented in Chapter 5. Chapter 6 discusses the classification results, potential sources of bias in the data, and the necessary aspects to consider while conducting additional research in this domain. Chapter 7 draws a conclusion to the current study and discusses future directions. The final segment of this study contains all the references and credits used.

Chapter 2

Literature review

In recent years, social media platforms have grown in popularity and have become an integral part of people's lives. This close relationship between social media platforms and their users has made these platforms to reflect the users' personal life on many levels. There is a growing body of literature addressing the role of social networks in the structure of social relationships such as breakups, mental illness, sexual harassment, and suicide ideation.

In the first portion of this chapter, the role of social media in depression is discussed, as well as various data extraction techniques to extract data from social media with depression symptoms. Following that, the state-of-the-art classification techniques and features for this purpose are addressed. Finally, in the latter part of this chapter, an in-depth discussion of the attention mechanism is provided.

2.1 Social Media Data and Depression

The prevalence of internet and communication technologies, particularly online social networks, has rejuvenated how people interact and communicate electronically. Applications like Facebook, Twitter, and Instagram not only host written and multimedia content, but also allow users to express their feelings, emotions, and sentiments about a topic, subject, or issue online. On the one hand, this is great for social networking site users to honestly and openly contribute and respond to any topic online; on the other hand, it creates opportunities for people working in the health sector to gain insight into what might be happening at the mental state of someone who reacted to a topic in a specific manner. Social media platforms and other online discussion forums have been particularly appealing to

the research community because of the massive scale of data. This massive data flow has resulted from increased internet access and people spontaneously sharing their hardship, sadness, and struggle on these platforms anonymously.

The common practice of managing depressive disorder involves detecting depression symptoms through survey-based methods or through questionnaires. However, these studies suffer from underrepresentation, sampling biases, and incomplete information. Cognitive biases, which prevents a patient from expressing their feelings openly in front of a doctor, is yet another limitation of this approach [23]. In contrast, people share their real-time frustrations, sadness, experienced trauma through on social media feeds. This can be used as a valuable resource for learning about users' feelings, emotions, behaviors, and overall mental state. In recent years, much progress has been made in studying the depressive symptoms through social media posts.

2.2 Extracting Data with Depression Symptoms from Social Media

Computational linguistics techniques are very difficult to be opted as a complete substitute for in-person mental illness diagnosis, but the successful application of this domain in identifying the progress and level of depression of individuals in online therapy may provide clinicians with more insights, allowing them to apply interventions more effectively and efficiently. Studies analyzing web data, especially social media platforms, have piqued the interest of the research community due to their scope and deep entanglement in contemporary culture [24]. The following segments elaborately discuss the methods and strategies of extracting samples with depressive symptoms in various social media contents.

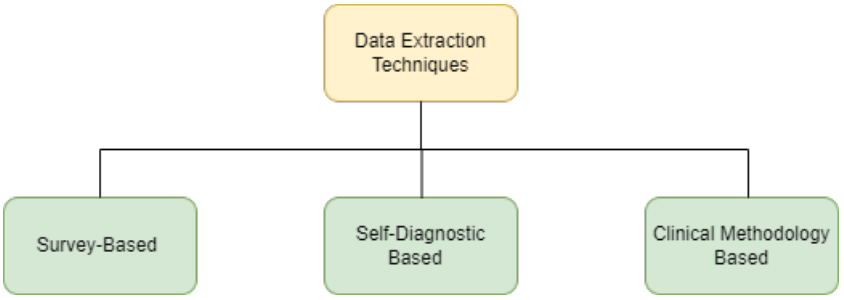


FIGURE 2.1: Data Extraction Techniques from Social Media

2.2.1 Survey Based Approach

Very few studies have investigated predicting the severity of depression based on users' language usage on web platforms. Choudhury et al. [8] proposed a metric named social media depression index (SMDI) using a probabilistic model to help characterize the levels of depression in the population level. This probabilistic model can predict whether or not a Twitter post contains symptoms of depression. To construct and train this model, they collected data using crowdsourcing technique and derived various linguistic and network features (e.g., number of followers) from tweets of individuals suffering from clinical depression, which was measured using the CES-D (Center for Epidemiologic Studies Depression Scale) screening test [25]. Schwartz et al. [26] attempted to predict and characterize the severity of depression based on people's Facebook language use. They gathered survey responses and Facebook posts from 28749 Facebook users and trained a classification model to predict depression symptoms using n-grams, linguistic behavior, and latent dirichlet allocation (LDA) topics. They tried to quantify the seasonal changes in depression symptoms based on social media posts and discovered that symptoms increase from summer to winter. These approaches had the potential to generate a large dataset with good quality data if they were developed in collaboration with expert psychologists and domain experts.

2.2.2 Self-Reported Diagnostic Based Approach

Coppersmith et al. [27] made a prominent contribution in mental health research domain by developing a procedure for extracting mental health data from social media. In their study, tweets were crawled from user profiles who publicly stated that they had been diagnosed with various mental illnesses on their Twitter feed. This method is known as 'self-reported diagnosis'. They mixed control samples from the general population (people who are not depressed) with the tweets of the self-reported diagnosed group. Control samples are samples that do not carry any depression related symptoms. They focused on the analysis of four mental illnesses: Post-Traumatic Stress Disorder (PTSD), Depression, Bipolar Disorder, and Seasonal Affective Disorder (SAD), and proposed this novel method to gather data for a range of mental illnesses quickly and cheaply. Numerous studies later followed this approach to detect relevant mental health data for various mental illnesses. For example, The Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task [28] collected self-reported data on Depression and

PTSD. They further annotated the data with human annotators to remove jokes, quotes, etc., from the collected data. The shared task participants had three binary classification tasks- identify depresison vs. control, identify PTSD vs. control, and indentify depression vs. PTSD. These datasets were used in a variety of studies to discover patterns in the language use of users suffering from various mental illnesses [9, 29, 30].

Following a similar approach, [31] collected tweets from self-reported depressed users and investigated the potential of non-temporal and temporal measures of emotions over time to identify depression symptoms from their tweets by detecting eight basic emotions (e.g. anger, fear, etc.). Additionally, classifiers were built to label Twitter users as either depressed or non-depressed (control) groups calculating the strength scores based on the intensity of each emotion and a time series analysis of each user. Among other social medias, Xianyun et al. [32] explored sleep complaints on Sina Weibo (a Chinese microblogging website) to discover users' diurnal activity patterns and gain insight into the mental health of insomniacs. Twitter data on mental health had also been collected, with specific Twitter campaigns being targeted. [33] prepared a dataset from the users who participated in the #BellLetsTalk 2015 campaign that was inaugurated to promote awareness about mental health issues. They collected public tweets from 25362 Canadian users and built a user-level classifier to detect at-risk users and a tweet-level classifier to predict symptoms of depression in tweets. From this campaign, they came across only 5% tweets that talk about depresison and 95% non-depressed tweets. While these methods can extract large volumes of data for a low cost, they do not ensure a sufficient sample of interest and have inevitably resulted in a low number of positive samples (mental-health related data) [33].

2.2.3 Clinical Methodology Based Approach

Several previous studies have investigated the use of clinical methodologies along with data mining tools to extract depression symptoms from diverse sources. Yazdavar et al. [12] created a lexicon of depression symptoms based on the nine disorders described in the clinically established Patient Health Questionnaire (PHQ-9) and utilized this to find symptoms of depression in tweets from users with self-reported depressive symptoms in their Twitter profile. They also developed a statistical model to categorize and monitor depressive symptoms for continuous temporal analysis of an individual's tweets. In a similar study, Mukhiya et al.

[34] proposed an open set of depression word embeddings that extracts depression symptoms from patient-authored text data based on PHQ-9 to deliver personalized intervention to people with symptoms of depression. Yadav et al. [35] utilized the nine symptom classes of the PHQ-9 questionnaire to manually annotate the tweets collected from 205 self-reported depression diagnosed users. Their proposed framework took into consideration the figurative language (metaphor, sarcasm etc) wired in the communication of depressive users on Twitter. Ahmed et al. [18] extracted depression symptoms in patient authored text in a similar fashion with PHQ-9 questionnaire but used an attention-based in-depth entropy active learning to annotate the unlabeled texts automatically. Their mechanism increased the trainable instances of mental health data using a semantic clustering mechanism with to reduce the data annotation task. Another mental health tool used by psychiatrists, namely the Diagnostic and Statistical Manual of Mental Disorders (DSM-5), has also been used to categorize mental disorders from social media content. Gaur et al. [36] developed an approach to map subreddits into DSM-5 categories. They created a lexicon from various subreddit posts by extracting n-grams and topics using LDA and mapped this lexicon with DSM-5 lexicon created by available medical knowledge bases (ICD-10[1], SNOMED-CT[2], DataMed[3]). Their approach attempted to connect a patient on social media platforms such as Reddit to appropriate mental health resources and to provide web-based intervention. Cavazos-Rehg et al. [37] investigated the most common themes of depression-related chatter on Twitter that corresponded to the DSM-5 symptoms for major depressive disorder. While these methods may have clinical validity, most studies that use them lack sufficient ground truth data due to the absence of a thorough annotation procedure.

Table 2.1 summarizes the dataset found in existing literature with related information such as, data domain, purpose and description of the approach, and target mental health symptom or condition. The availability of these datasets are unknown as availability of most of these datasets are not declared in the papers. Ethical considerations further complicate data acquisition, with the sensitive nature of mental health data requiring tremendous care when constructing, analyzing, and sharing datasets. This table is adapted from the survey of [2].

[1] https://bioportal.bioontology.org/ontologies/ICD10
[2] http://bioportal.bioontology.org/ontologies/SNOMEDCT
[3] https://datamed.org/

TABLE 2.1: Summary of Mental Health Dataset Including Data Domain, Purpose and Description of the Approach, and Target Mental Health Symptom or Condition (Table adapted from [2])

Reference	Data Domain	Purpose	Approach	Targetted Mental Illness
Nguyen et al. [38]	Social media: Live Journal	Understanding mental health content	Application of text-mining to better understand linguistic features and topics related to mental health discussed within online communities on the Live Journal platform.	Depression
Fatima et al. [39]	Social media: Live Journal	Detecting symptoms/ condition	Development of three ML models for classifying depressive posts, communities and the degree of depression from online social media (Live Journaling posts).	Depression
Gaur et al. [36]	Social media: Reddit	Detecting symptoms/ condition	Development of multi-class classification algorithm that analysis mental health subreddit posts and quantifies their relationship to DSM-5 categories.	Mental illness (generic)
Joshi et al. [40]	Social media: Twitter	Detecting symptoms/ condition	Development of a model to identify different types of mental health conditions from peoples' social media tweets.	Mental illness (generic)
Yazdavar et al. [12]	Social media: Twitter	Detecting symptoms/ condition	Development of a statistical model for monitoring different symptoms of depression by modeling user-generated content in social media tweets over time.	Depression
Chen et al. [31]	Social media: Twitter	Detecting symptoms/ condition	Development of a model that includes measures of eight basic emotions and temporal data as features in prediction self-reported diagnosis of depression on Twitter.	Depression
Ernala et al. [19]	Social media: Twitter + Facebook	Detecting symptoms/ condition	Empirical study to assess internal and external predictive validity of different social media-derived proxy diagnostic signals for schizophrenia.	Schizophrenia
Nobles et al. [41]	Messages (SMS)	Understanding/ predicting risks	Development of a model that identifies periods of suicidality. Report on collection + analysis of text messages of individuals with a history of suicidal behaviors.	Suicidal Behaviors
Pestian et al. [42]	Suicide notes	Understanding/ predicting risks	Development of a classifier for predicting suicide through natural language processing of written suicide notes.	Suicidal Behaviors
Adamou et al. [43]	Medical notes (from Health record)	Understanding/ predicting risks	Application of text-mining techniques of medical notes to improve accuracy of a predictive model of suicide risk within 3 or 6 months at point of referral to mental health services.	Suicidial Behaviors
Wilbourne et al. [44]	Messages (chat app)	Improving treatment	Use of ML tools to aid supporters of text-based, technology-enabled mental health intervention to assess the quality of their coaching in real-time.	Mental health (generic)
Saha and De Choudhury [45]	Social media: Reddit	Understanding mental health content	Development of a ML classifier for inferring expressions of stress from social media posts and time series analysis to examine temporal patterns (before/after) gun violence.	Stress
Kavuluru et al. [46]	Social media: Reddit	Understanding mental health content	Development of identifiers of "helpful" comments posted within the Reddit community: Suicide Watch (SW), using varied text-mining techniques.	Suicidal Behaviors

2.3 Classification Techniques Used in Literature

Both supervised and unsupervised methods have been explored as classification techniques of depressive symptoms in social media contents.

2.3.1 Feature Extraction Techniques

Various linguistic features have been used in classifying depression symptoms from social media texts. Some works have utilized other features in the training phase as well. Choudhury et al. [8] utilized various emotion and linguistic based features from twitter posts using Linguistic Inquiry and Word Count (LIWC). Linguistic Inquiry and Word Count (LIWC) is a text analysis method that calculates the percentage of words in a given text that fall into one or more of over 80 linguistic, psychological, and topical categories that indicate various social, cognitive, and affective processes. It has two central features- the processing component and the dictionaries. The processing feature is the program itself, which opens a series of text files—which can be essays, poems, blogs, novels, and so on—and then goes through each file word by word. Each word in a given text file is compared with the dictionary file. LIWC has been used extensively in mental health research for handpicking features from text data. For example, Coppersmith et al. [27] mixed control samples (non-depressed samples) with self-reported depression diagnosed tweets and conducted a LIWC analysis to measure deviations of mental disorder group from the control group. A similar kind of work was done in Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task [28].

Topic modeling is another technique used in finding prominent attributes in mental health related data. Topic modeling can automatically cluster word groupings and related expressions that best represent a set of corpus. This is particularly useful in cases of finding language patterns and discovering hidden themes in a set documents. Among the methods used to model topics in mental health research, Local Dirichlet Allocation (LDA) is very popular. In LDA, documents are represented as a collection of topics and a topic as a group of words. Those topics are found in a hidden layer, also known as a latent layer. LDA examines a document to identify a set of topics that are most likely to have produced that set of words. So, if a document contains certain words that are found in a topic, the document could be said to be about that topic. LDA consists of two parts, the words within a document (a known factor) and the probability of words belonging to a topic, which is what needs to be calculated. The algorithm tries to figure out how many words in a document belong to a specific topic. Schwartz et al. [26] used Latent Dirichlet Allocation (LDA) topics as features for predicting depression symptoms in Facebook posts. Gaur et al. [36] created a lexicon from various subreddit posts by extracting topics using LDA and mapped this lexicon with DSM-5 (Diagnostic and Statistical Manual of Mental Disorders)

lexicon created by available medical knowledge bases. Resnik et al. [47] conducted several topic modeling (supervised Latent Dirichlet Allocation (LDA), supervised anchor topic modeling, etc.) to differentiate the language usage of depressed and non-depressed individuals using the datasets of [27] and CLPSych Shared Task (2015). In literature, LDA topics are often used as features in classification models to detect depressive symptoms in social media text [26].

Among other handpicked features, BOW (Bag of Words) [33], bi-gram, n-gram [36], network features (e.g. number of friends, followers etc.) [8] have been used to classify mental illness.

2.3.2 Classification Models

A variety of machine learning techniques, such as classification, regression, association, and clustering, have been used to simplify high-dimensional datasets for common tasks like identifying correlations and pattern recognition to achieve more human-interpretable formats.

When it comes to detect depression from social media texts, the classification models used are either supervised or unsupervised in nature. In supervised learning, data is labeled and is used to train a model that can predict the label for new data. The dataset contains both the inputs and the desired outputs in this case. On the other hand, unsupervised learning clusters data using mathematical techniques to provide new insights. The dataset only contains inputs and no desired output labels in this case. Clustering methods respond to the presence or absence of commonalities in each piece of data to discover patterns and help structure the data. A vast majority of the papers used supervised learning, and most often described the application of one or more of these techniques: Support Vector Machines (SVM), Random Forest, Decision Trees, k-Nearest neighbors, supervised LDA, and Logistic Regression. These approaches are mostly supervised, or unsupervised way of classifying data. Choudhury et al. [8] proposed a probabilistic model consisted of an SVM classifier that can predict whether or not a twitter post contain symptoms of depression. Schwartz et al. [26] used regression modeling to predict depressive symptoms in facebook posts. In the Computational Linguistics and Clinical Psychology (CLPsych) 2015 shared task [28], three binary classification tasks were introduced- identify depresison vs control, identify PTSD vs control, indentify depression vs PTSD. An SVM classifier with linear

kernel achieved an average precision of 80% for all the three binary tasks. Jamil et al. [33] collected public tweets from 25362 Canadian users and built a user-level classifier to detect at-risk users and a tweet-level classifier to predict symptoms of depression in tweets. Both of these classifiers were based on Linear SVM.

Clustering is another approach that has been used extensively in mental health domain. Using this unsupervised technique, similar data points are grouped together to discover underlying patterns in multiple documents. Clustering techniques can discover hidden themes in documents and provide valuable understanding of the data in hand. For example, Low et al. [3] revealed valuable insights into the impact of the COVID-19 pandemic on population-level mental health inferring from posts and comments on *Reddit*. After careful observation, it was discovered that Supervised Latent Dirichlet Allocation (LDA) is the most widely used clustering algorithm in the literature. LDA is described elaborately in Section 2.3.1.

Ahmed et al. [18] introduced attention mechanism in the mental health research domain by proposing an attention-based in-depth entropy active learning model. Their proposed mechanism increased the trainable instances of mental health data using a semantic clustering algorithm and reduced the data annotation task. In this algorithm, semantic vectors based on semantic information derived from the context in which it appears are clustered. The resulting similarity metrics help to select the subset of unlabeled text by using semantic information. The proposed method separates unlabeled text and includes it in the next active learning mechanism cycle. This method updates model training by using the new training points. The cycle continues until it reaches an optimal solution, and it converts all the unlabeled text into the training set. As this research is strongly inspired by the attention mechanism, it will be broadly discussed in the following section.

2.4 Attention Mechanism

2.4.1 What is Attention?

Attention is a powerful concept originally introduced to improve the performance of the encoder-decoder architecture on neural network-based machine translation tasks. This mechanism is now used in various tasks like image captioning, speech processing, etc.

Google's Sequence-to-sequence (Seq2Seq) model is composed of encoder-decoder architecture, where the encoder processes the input sequence and preserve the processed information into a 'context' vector of fixed length. Encoder and decoders are basically RNN cells. After the encoder processes the input sequence, the decoder takes input of the 'context' vector and the last hidden state of the RNN-encoder cell and produces the output sequence. An overview of the Seq2Seq is provided in Figure 2.2.

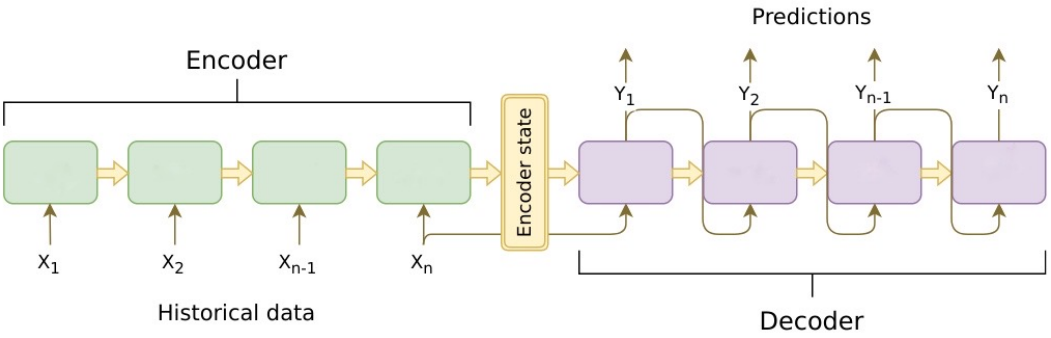


FIGURE 2.2: Seq2Seq Architecture

Seq2Seq models achieved a lot of success in tasks like machine translation, text summarization, and image captioning in the late 2016. But a critical disadvantage of this architecture was the fixed-length context vector, which is unable retain information in longer sequences. Often it forgets the earlier elements of the input sequence once it has processed the complete sequence. The attention mechanism was created to resolve this problem of long dependencies.

A solution to this problem was proposed by Bahdanau et al. [48] and Luong et al. [49], where they introduced and refined a technique called "Attention", which highly improved the quality of Seq2Seq's encoder-decoder architecture. Attention allows the model to focus on the relevant parts of the input sequence as needed. An attention model differs from a classic Seq2Seq model in two different ways-

- Instead of passing the last hidden state of the encoding state, the encoder passes all the hidden states to the decoder.
- Multiply each hidden state by a softmaxed score, thus amplifying certain hidden states with high scores, and drowning out certain hidden states with low scores.

An overview of the attention mechanism with the improved the Seq2Seq architecture is provided in Figure 2.3.

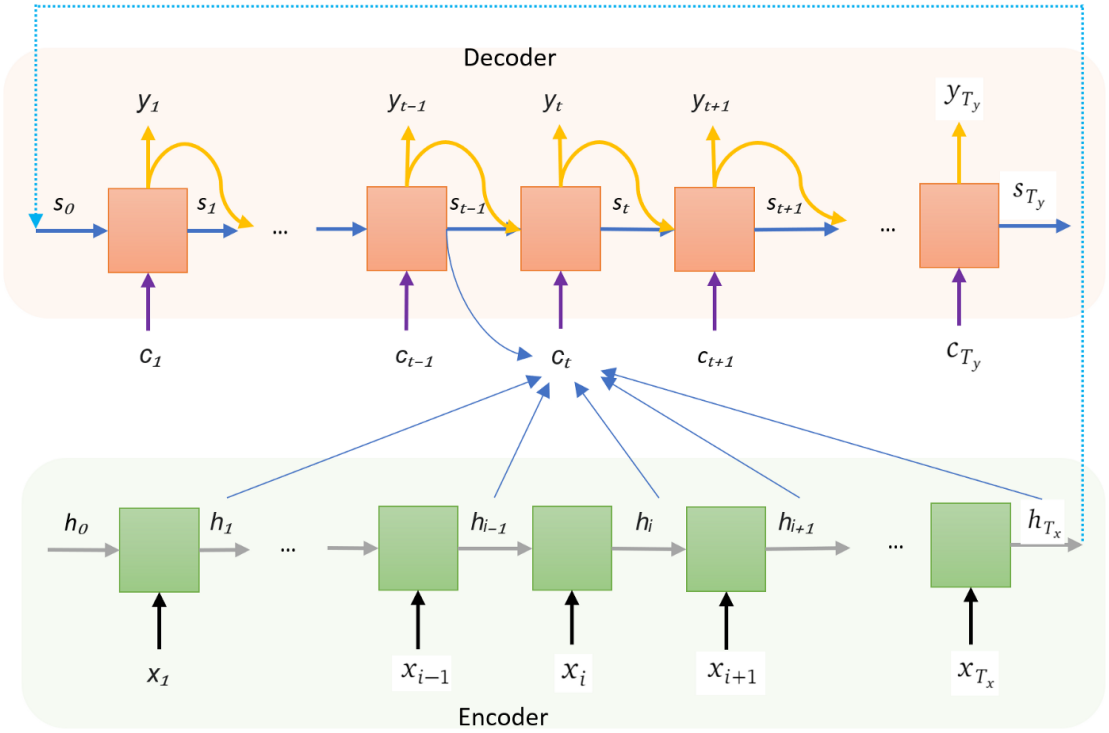


FIGURE 2.3: Improved Seq2Seq Architecture with Attention Mechanism

The context vector c_i for the output word y_i is generated using the weighted sum of the annotations:

$$c_i = \sum_{j=1}^{T_x} a_{ij} h_j \tag{2.1}$$

In the above equation, T_x referred to as window size, a_{ij} is the attention weight, and h_j is the hidden state. The attention weights are calculated by normalizing the output score of a feed-forward neural network described by the function that captures the alignment between input at j and output at i.

$$a_{ij} = \frac{exp(e_{ij})}{\sum_{k=1}^{T_x} exp(e_{ik})} \tag{2.2}$$

$$e_{ij} = a(s_{i-1}, h_j) \tag{2.3}$$

With this framework, the model can selectively focus on valuable parts of the input sequence and thus learn the relationship between them. This allows the model to deal with long input sentences more efficiently.

2.4.2 Utilization of Attention in Classification Tasks

After introducing attention mechanism, many models have been utilizing this mechanism in different variations. Several variations of the network include soft, hard, and global attention mechanism. In soft attention [48], the model uses the average of the hidden states and then builds the context vector. In hard attention [50], the context vector is computed by sampling the hidden states. In global attention [49], the model picks the attention point for each input batch, and thus helps the model converge quickly.

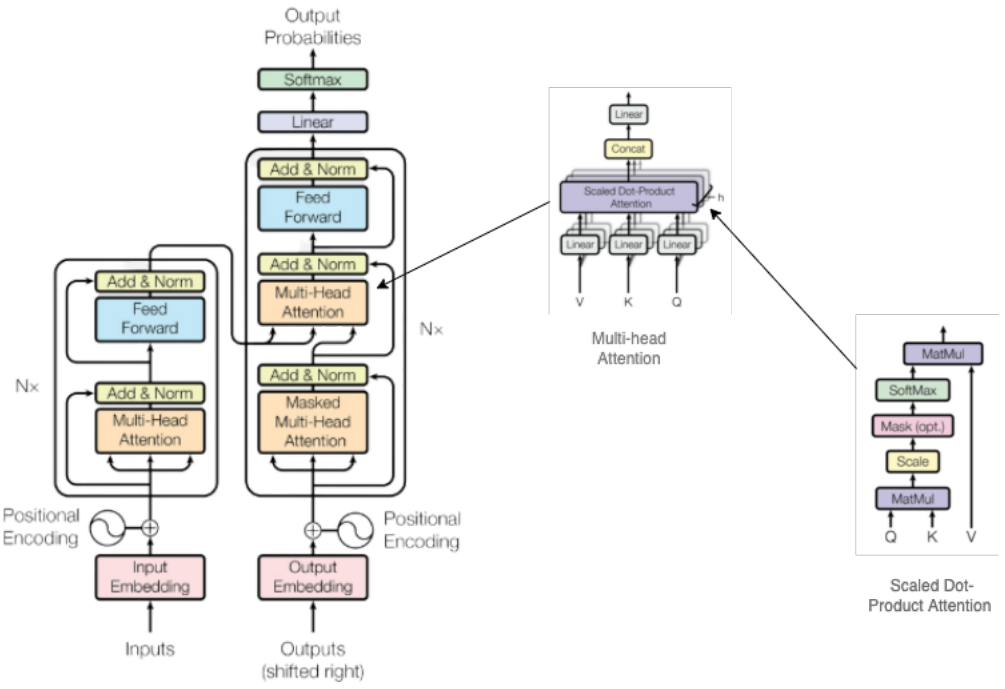


FIGURE 2.4: Transformer Architecture (Image adapted from [1])

The attention mechanism solves the problem of remembering longer sequences persisted in encoder-decoder model. However, models that utilized the attention mechanism took a lot of time to train. Vaswani et al. [1] provides a solution of this problem by introducing transformers in their famous paper 'Attention is All you Need'. Transformer is a model that uses attention to boost the speed by parallelization. Transformers relies solely on the use of self-attention, sometimes called intra-attention, where the representation of a sequence (or sentence) is computed

by relating different words in the same sequence. Vaswani et al. [1] proposed a scaled dot-product attention, and then build on it to propose multi-head attention. In this method, three vectors are used, namely the query (Q), keys (K) and values (V) that are used as inputs to the these attention mechanisms. They create different projections of the same input sentence. Therefore, the proposed attention mechanisms implement self-attention by capturing the relationships between the different elements (in this case, the words) of the same sentence. An overview of the transformer architecture is provided in Figure 2.4.

Later, this transformer architecture has been used in many language representation model like BERT [51], DistilBERT [52] which can be pre-trained and fine-tuned to create state-of-the-art models for a wide range of tasks. These tasks include question answering systems, sentiment analysis, language inference etc. The original transformer was designed for language translation, particularly from English to German. But, Vaswani et al. [1] showed that the architecture generalized well to other language tasks as well. Now-a-days, any language-related ML task is being completely dominated by some version of the transformer architecture as observed from related literature.

Chapter 3

Proposed Approach

The procedure used to assess the severity of depression in this study is based on a well-established clinical assessment method known as the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) [22], and it was carried out under the supervision of two expert clinical psychologists. Twitter was chosen as the data collection platform for ease of accessibility of data and avalaibility of APIs.

3.1 Measuring Severity of Depression

In this study, a user posting a tweet on social networking site Twitter is considered to be depressed if the tweet depicts behaviors portraying symptoms of depression. Such a tweet may not necessarily be complete, contain well-structured sentences, or even grammatically correct, making the task even more difficult.

According to the Diagnostic and Statistical Manual of Mental Disorders (DSM), clinical depression can be diagnosed considering the existence of a set of symptoms over a substantial amount of time [12]. Incorporating this idea, the Patient Health Questionnaire (PHQ-9) [21] provides a set of questionnaires, which is widely used to screen, diagnose and measure the severity of depression. Using this set of questionnaires, nine distinct symptoms related to different disorders, such as lack of interest, eating disorder, etc., can be extracted. (Table 3.1).

The frequency of these symptoms can help classify the severity of depression as none, mild, moderate, and severe conditions. This approach is called Clinical

TABLE 3.1: Sample Tweets, Seed Terms and Final Keywords List for Each Symptom of PHQ-9 Questionnaire

PHQ-9 Symptoms	**Sample Tweet**	**Seed Terms**	**Final Keyword List**
Lack of interest (S1)	Don't know whats happening with me. Can't feel like doing anything for days	disinterest	involved, occupied, pessimism, reversion, absorbed, lifelessness, bored, enthusiasm, engrossed, worried, apathy.
Feeling Down (S2)	I've never felt so hopeless in a while.	hopeless, depressed	dejected, dismayed, dispirited, demoralized, grimmed, misery, grim, downhearted, low-spirited, bleak, desperate, lost, frustrated.
Sleep Disorder (S3)	Ah! Sleepless nights are so painful. Tired of this already	awake, sleep	nap, restless, awake, whole night, bedtime.
Lack of Energy (S4)	Feeling very down this week...I know I have a s*** load of stuff going on, and I'm overtired and so drained	tired, energy	weary, fatigue, fag, fag out, overtire, overfatigued, burned-out, burnt-out, exhausted, dog-tired, washed-out, drained, whacked.
Eating Disorder (S5)	I was so depressed of my overweight and started starving to look skinny	appetite, overeating	aversion, distaste, loathing, malformed, bulimic, puffy, starve, fat
Low Self-estemm (S6)	So disgusted with myself	loser, failure	loser, relapse, downfall, ruined, flop, dead-duck, disappointment, achiever, misfire, underdog, falling-apart, disgusted
Concentration Problems (S7)	Can't even finish an article in one seat nowadays!	concentrate, focus	immersed, decentralize, deconcentrate, scattered, dispersed, unsettled, focus
Hyper/Lower Activity (S8)	So stressed out, I can't do anything	moving, immobile, restless	discontent, ungratified, unsatisfied, stand-still, refrained, immobile
Suicidal Thoughts (S9)	Gotta find a way to hide my cuts..	dead, hurt, suicide	trauma, harm, suffering, anguish, hemorrhage, penetrating-trauma, torment, agony, excruciate, damaged, gag, suffocate, self-destruction

Symptom Elicitation Process (CSEP) [53] . In this study, this was further extended using the mood scale provided by BipolarUK[1] to identify the characteristics related to different levels of depression. The following characteristics were then verified by the collaborator psychologists and used to detect the level of depression from the user tweets:

[1]https://www.bipolaruk.org/faqs/mood-scale

3.1.1 Non-depressed Tweets

A tweet can be labelled as a non-depressed tweet if it expresses a person's joy or delight, or makes a generalized statement about depression that does not reflect the person's own mental state, expresses casual tiredness or sadness (For example, sadness due to the defeat of their favorite sports team), or expresses temporary hopelessness.

3.1.2 Mildly Depressed Tweets

A tweet that expresses hopelessness or a feeling of disinterest that persists for a while can be labeled as a mildly depressed tweet. A mildly depressed tweet may contain symptoms of hopelessness, feelings of guilt or despair, difficulties concentrating at work, a loss of interest in activities, a sudden disinterest in socializing, a lack of motivation, insomnia, weight changes, daytime sleepiness and fatigue, appetite changes, and reckless behavior (such as, alcohol and drug abuse).

3.1.3 Moderately Depressed Tweets

Moderate depression has symptoms similar to mild depression. The differentiating factor is that the severity of symptoms hampers activities related to home and work. Tweets may contain symptoms of increased sensitivities, feeling of worthlessness, reduced productivity, problems with self-esteem, excessive worrying.

3.1.4 Severely Depressed Tweets

The symptoms of this category are more noticeable and life threatening. They contain delusions, feeling of near-unconsciousness or insensibility, hallucinations, suicidal thoughts, or behaviors.

The characteristics of different severities of depression with example tweets is summarized in Table 3.2.

3.2 Data Collection

For this study, seed terms are generated from the keywords extracted from each of the symptoms of PHQ-9 questionnaire by collaborating with two professional

TABLE 3.2: Characteristics of Severities of Depression with Example Tweets

Depression Levels	**Characteristics**	**Example Tweets**
Non-Depressed	A generalized statement about depression, casual tiredness or exhaustion due to hard work that is not persistent, etc.	When I first saw this fight I was thinking, "This is so unfair.He just fought his way up 4 floors in a single take, he must be exhausted..."
Mildly Depressed	Feelings of panic and anxiety, concentration difficulty, a sudden disinterest in socializing, feelings of guilt and despair	im getting tired of chasing you back, trying to fix us; at one point, i see like it's always me who apologize first
Moderately Depressed	Slow thinking, no appetite, excessive or no sleep, everything seems a struggle.	Everything I do now takes days to complete. Its my new reality. Every time I push myself I mess up my progress. Depression hittin hard lately, even in my dreams. https://t.co/35sg6aUd1g
Severely Depressed	Endless suicidal thoughts, no movement, everything seems bleak, feelings of hopelessness and guilt, impossible to do anything	Haven't felt this depressed since I last tried to hurt myself in 2009. Didn't think feeling low like this would return. Luckily I'm not at state where I would try to hurt myself but it's tough

psychologists. Seed terms generation from PHQ-9 to collect mental health data is a commonly used procedure employed in many previous studies [12, 18, 34]. After seed terms generation, they are then extended using WordNet [54]. It is a well-known lexical database developed by Princeton University that links words into semantic relations, including synonyms, hyponyms, meronyms, and antonyms. Each category of words is maintained according to their parts of speech, i.e. nouns, verbs, adjectives, and adverbs in the database and the synonyms are grouped into synsets. Words that are in the same synset are synonymous and interlinked using conceptual-semantic and lexical relations. There are several other methods used in different studies [12, 34] such as *Universal Sentence Encoding (USE), Global vector representation (Glove), Big Huge Thesaurus*, etc. In the evaluation showed by [34], WordNet performs significantly better in extracting symptoms from patient-authored text compared to other methods. For this study, the seed terms for each questionnaire of PHQ-9 were extended by WordNet, and the extended terms were handpicked afterward by the psychologist collaborators. After several rounds of filtration, a final lexicon list containing 88 depression-related keywords categorized into nine different clinical depression symptoms of PHQ-9 was prepared, which are likely to appear in the tweets of individuals suffering from different severities of depression. Table 3.1 illustrates samples of anonymized tweets, seed terms, final keywords list extended by WordNet and their associated symptoms in PHQ-9. Based on the final keyword list, a total of 344657 tweets were collected.

3.3 Data Annotation

From the collected samples, tweets that were posted in English are only preserved for annotation. Tweets with less than eight words are discarded as they might not contain enough context. Any tweets containing mentions (@) or hashtags (#), as well as retweets, are also discarded since they could violate the privacy of the users mentioned. Finally, 44100 tweets are randomly chosen from the remaining tweets for annotation.

3.3.1 Annotator Recruitment

The annotation job is done by recruiting participants who are fluent in English and had a previous experience of text assessment. The annotator pool consisted of 111 crowdworkers, and they were pre-screened for eligibility using two online sessions. Initially, 90 annotators were selected randomly for the annotation job after pre-screening. Each annotator received $20 for participating in the study. The task of the annotators was to label the tweets in four classes, i.e., non-depressed, mildly depressed, moderately depressed, and severely depressed tweets. The annotators were briefed through 2 long online sessions under the supervision of the collaborator psychologists about the classification and were also provided with a detailed document on the severity classes. Each annotator was given a datafile with only two columns: (1) tweet texts and (2) possible label suggestions (0: non-depressed, 1: mild, 2: moderate, 3: severe) and was asked to determine the tweet's possible class label.

The inherent subtlety and ambiguity of the attributes covered in this dataset makes the annotation process an unavoidably difficult process. Each annotator may have a unique perspective on the nuance of the context presented in tweets, as well as a unique perception of the severity of the depression. Annotators were asked to avoid personal bias while labeling the tweets and strictly follow the guidelines provided to them to classify the text. All tweets are annotated at least three times. The final label of each tweet is determined by majority voting of the labels provided by the three annotators. Tweets with different labels from all the three annotators are discarded because of too many disagreement. Final labels of the dataset are established with a confidence score to reflect the disagreement of the annotator because of reasonable difference of opinion.

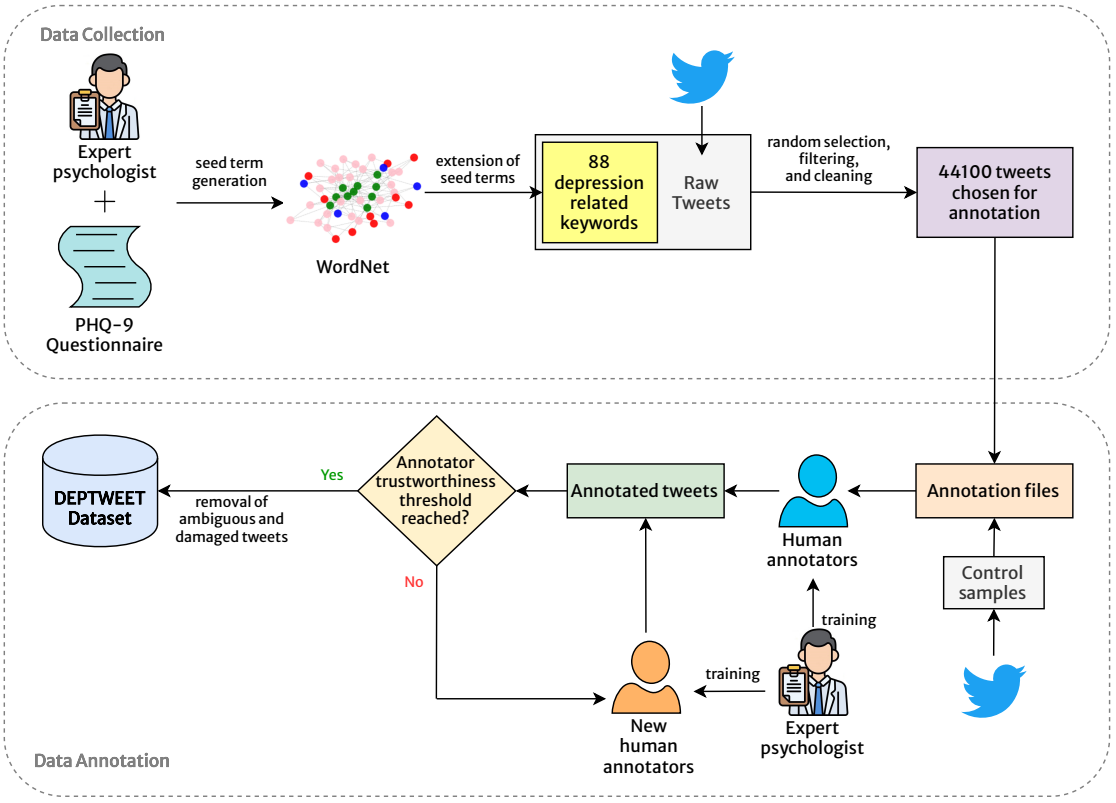


FIGURE 3.1: Overview of the Dataset Creation Process

3.3.2 Annotation Job Refinement

Though it is ensured that annotators' disagreement reflected a genuine difference of opinion, a means of quality control is required to prevent annotators' inattention, or misunderstanding of context. The quality control mechanism used by [55] is followed in this study. This mechanism aimed to reduce the number of 'bad' annotators, those who either did not correctly understand the task or annotated the datafiles too recklessly, without giving proper attention. As part of the quality control, a set of 'control samples' is collated with the actual data sample, for which the correct labels are manually established. Annotators encounters one control sample per batch of fifty tweets without knowing which of the tweets is the control sample. The running accuracy of these control samples is defined as annotator's 'trustworthiness score (T)'. The threshold trustworthiness score for this study is set to be at least 90%. If an annotator drops below this level, all of their annotations are discarded, and the annotator is removed from the annotator pool. Afterwards, another annotator from the pool is assigned to re-annotate those data samples.

A total of 900 control samples are added for quality control with the previously

TABLE 3.3: Metadata About the Datafiles Created for Annotation

Type	**Count**
Number of tweets collected	344657
Tweets chosen for annotation	44100
Total datafiles created	30
Data samples in each datafile	1470
Control samples per datafile	30
Total tweets per datafile	1500

chosen 44100 data samples. To generate datafiles for the annotators, the actual dataset containing 44100 samples are divided into 30 parts, each part containing (44100/30) = 1470 samples. For every 49 tweets in these 1470 samples, one unique control sample is added at a random position. The control samples are from the *non-depressed* category and are limited to only obvious and conclusive instances of attributes. Thus, one would fail on these control samples only if they had an incorrect comprehension of the attributes of the class labels or was too reckless while annotating. The tweet ID of the control samples are also tracked. Following this method, 30 datafiles are created containing (1470 data samples + 30 control samples) = 1500 tweets each. Each datafile only contains tweet text and the annotator label. All the other data columns are kept hidden from the annotators. The datafile creation procedure is summarized in Table 3.3. To annotate these datafiles, ninety annotators are divided into three groups, each with thirty annotators. Each datafile is given to three different annotators from three different groups. Before partitioning, the data samples were randomized so that no two data files contained identical tweets in the same order. Once the annotation process is finished, all the datafiles are merged and the control samples are removed from the dataset. An overview of the dataset creation process is provided in Figure A.

Chapter 4

Dataset Properties and Analysis

From the 44100 tweets considered for annotation, 1399 data samples are removed from the dataset because they were damaged (i.e., tweet text or tweet ID was changed) during the annotation process, and 2510 data samples are discarded due to annotator disagreement, as they received three different labels from three different annotators. The final dataset comprises a total of 40191 tweets along with their *tweet_id*, *replies_count*, *retweets_count*, *likes_count*, *target*, *label* and *confidence_score*. The label for each tweet is determined based on the aggregation of the labels provided by different annotators. If at least two of the three annotators agree on the label of a tweet, the matched annotation is accepted as the final label. Tweets that had three different annotations from three annotators, are discarded and saved in a separate datafile. The corresponding confidence score for each label is determined by an weighted average of the annotator's 'trustworthiness score'. Confidence Score for a particular label of a tweet sample can be written as:

$$\text{Confidence Score}(C) = \frac{\sum T_i}{T} \tag{4.1}$$

where T_i denotes trustworthiness of i^{th} annotator whose annotations match and T denotes sum of the trustworthiness score of all the annotators who annotate the tweet.

To demonstrate this process, consider a tweet sample annotated by three annotators A, B and C having trustworthiness scores $T_A = 0.90$, $T_B = 0.93$, and $T_C = 1.00$. If the annotated label of annotators A and B matches, then the

confidence score of the label will be $(T_A + T_B)/T$, where T is the sum of the trustworthiness score of the three annotators. In this case, the confidence score for the label of the particular tweet would be 0.647.

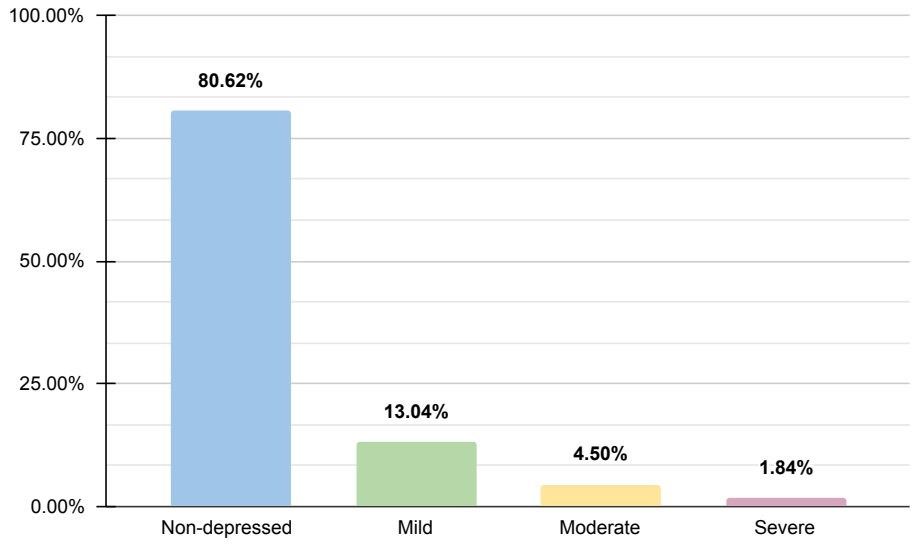


FIGURE 4.1: Percentage of Data Samples for Each Class

The proportion of classes shown in Figure 4.1 indicates that the *non-depressed* samples outnumber the other classes by a wide margin. Though all the data samples are scraped based on the keywords related to different severity levels of depression and the control samples were removed prior to the final preparation of the dataset, the number of data samples for different severities of depression is inevitably low. This class imbalance represents an important characteristic in the identification of various depressive disorders on social media. To discover this, manual analysis was done in two stages of this study: (i) while randomly choosing data samples for annotation, and (ii) during the initial iterations of the annotation job. The analysis indicated that the final class proportions roughly represent the percentage of similar attributes in similar live contexts.

Generally, the overall positive content shared in social media outnumbers the negative content. This is because people usually show their positive, friendly side over social media and tend to talk less about their struggles [56]. To mitigate this problem, previous studies depended on self-labeled data for collating large and balanced datasets on different mental disorders [3, 11]. However, depending only on self-labeled data to understand mental health from personal levels and measure the severity of the condition is not feasible without the intervention from

expert psychologists. But considering the lack of resources in the mental health sector, only relying on psychologists can be time-consuming and expensive. As a result, in this study, crowdsourcing supervised by psychologists was opted to obtain high-quality data on different depression severities.

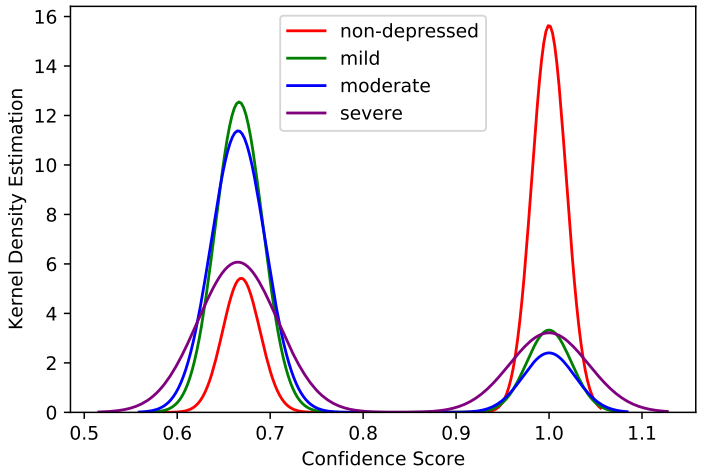


FIGURE 4.2: Kernel Density Estimation of Confidence Scores for Each Class

Despite the measures undertaken to ensure the quality of the dataset, the method of annotation warrants a certain level of noise in the dataset. This results in different yet rational interpretations of the same tweet. The kernel density estimation of the confidence scores portrayed in Figure 4.2 indicates that there is reasonable agreement among the annotators on deciding the class label of the *non-depressed* and *severe* classes. While these two classes lies on two different polarities of attributes, the subtle nuances of the *mild* and *moderate* classes allowed for rational disagreement among the annotators, which is evident from the high concentration of probability density for *mild* and *moderate* classes between 0.6 and 0.7 in Figure 4.2. This may be attributed not only to the lack of apprehension or awareness of the annotator, but also on the subjectivity of the topic at hand. It highlights the difficulty of using typical reliability metrics such as Inter-Rater Reliability (IRR), which calculates the level of agreement between two or more annotators. More sophisticated metrics like Fleiss' Kappa [57] can be applied in this scenario since the sample tweets were distributed randomly among the annotators and each annotator chose from one of the four mutually exclusive labels to indicate the severity of depression per tweet [58, 59]. However, Fleiss' Kappa assumes that the disagreement among the annotators on the same sample reduces the reliability of the dataset. Considering the subjective nature of the severity of

depression detected by different annotators, that might not be the case [60]. In spite of that, Fleiss' Kappa is calculated to get an understanding of the overall agreement of the annotators in this study. The value of Fleiss' Kappa ranges from -1 (indicating no observed agreement) to +1 (indicating a perfect agreement) [59]. Here, a value less than 0.20 indicates a poor agreement, 0.21 to 0.40 indicates a fair agreement, 0.41 to 0.60 indicates moderate agreement, 0.61 to 0.80 indicates substantial agreement and 0.81 to 1 indicates a near perfect agreement among the annotators.

TABLE 4.1: Fleiss' Kappa per Class

Class	**Fleiss' Kappa**
Non-depressed	0.44
Mild	0.27
Moderate	0.30
Severe	0.45
Overall	0.36

As reported in Table 4.1, the Fleiss' Kappa for the *non-depressed* and *severe* classes show a moderate agreement among the annotators. This can be explained considering the extreme nature of these two classes as they tend to be the polar opposite of each other. On the other hand, a fair agreement in *mild* and *moderate* classes highlight the intricate relationship among these two classes and the difficulty in identifying the subtle cues to differentiate them, even for the humans. However, despite the subjective nature of the severity of depression, an overall fair agreement provides indication of the quality of the annotation, and the dataset in general.

4.1 Visualizing the Dataset

To visualize and discover the hidden themes in the four classes of the dataset, some unsupervised topic modeling techniques are applied. Topic modeling identifies topics present in a text object and to derive hidden patterns exhibited by a text corpus. The following sections contain various techniques and visualizations to represent topics and themes of the classes to understand underlying pattern of the linguistic use.

4.1.1 Wordcloud

A word cloud is a simple visual representation object for text processing, which shows the most frequent word with bigger and bolder letters, and with different colors. The smaller the the size of the word the lesser it's important. figurename 4.7 represents the wordclouds for four classes. The wordclouds might look similar, but the difference of the words are in the context of their usages. *Non-depressed* class can contain terms like 'tired', 'exhausted', 'depressed', etc. but they might casually indicate to a temporary tiredness or exhaustion because of a work, not a permanent trait of the users' daily lives. The Non-depressed wordcloud containing 'work', 'love', 'today', etc. confirm this assumption. On the other hand, frequent words like 'suicide', 'hate', 'self-destruction', etc. in the wordcloud of the *severe* class provide an idea of the calamitous mental state of users potentially suffering from severe depression.

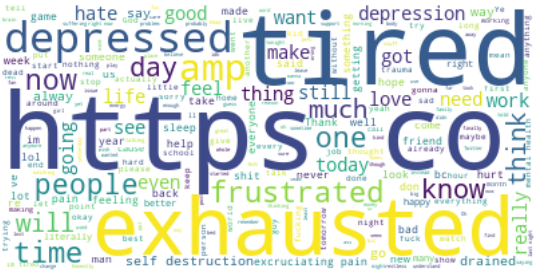


FIGURE 4.3: *Non-depressed* Wordcloud

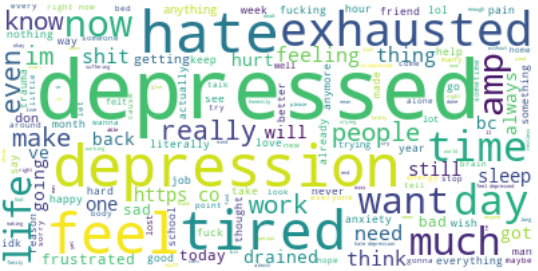


FIGURE 4.4: *Mild* Wordcloud

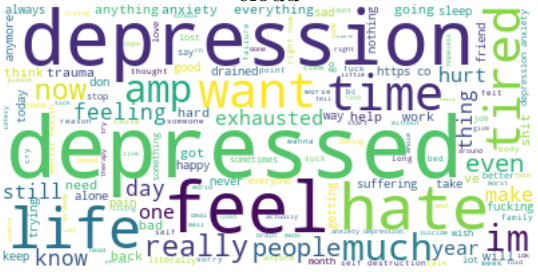


FIGURE 4.5: *Moderate* Wordcloud

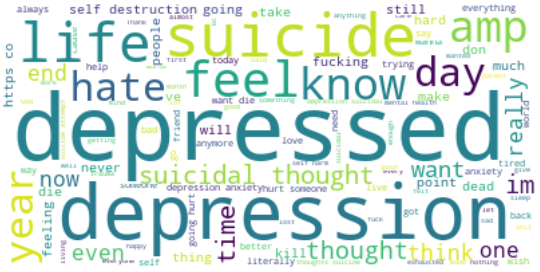


FIGURE 4.6: *Severe* Wordcloud

FIGURE 4.7: Wordclouds of Different Classes

4.1.2 Topic Modeling with Local Dirichlet Allocation

Latent Dirichlet Allocation (LDA) is a matrix factorization technique and one of the most popular topic modeling methods used by data analytics. LDA produces a generative probabilistic model that assumes each topic is a mixture over an

TABLE 4.2: Prevalent Topics of 4 Classes Discovered by LDA

Depression Levels	**Non-Depressed**	**Mildly Depressed**	**Moderately Depressed**	**Severely Depressed**
Topic 1	0.042*like+ 0.035*depress+ 0.033*http+ 0.025*feel+ 0.024*look+ 0.016*hate+ 0.015*say+ 0.015*love+ 0.010*frustrate+ 0.010*tire	0.035*tire+ 0.034*life+ 0.030*like+ 0.025*drain+ 0.023*mental+ 0.017*physic+ 0.016*feel+ 0.015*work+ 0.014*sick+ 0.013*know	0.035*know+ 0.031*go+ 0.030*feel+ 0.030*hurt+ 0.028*want+ 0.023*time+ 0.020*sick+ 0.020*suffer+ 0.019*fuck+ 0.018*month	0.137*feel+ 0.090*suicide+ 0.062*like+ 0.054*want+ 0.049*know+ 0.029*thing+ 0.027*thought+ 0.021*lose+ 0.020*live+ 0.020*anymore
Topic 2	0.034*depress+ 0.027*fuck+ 0.024*exhaust+ 0.022*http+ 0.020*mental+ 0.018*watch+ 0.018*need+ 0.018*health+ 0.012*care+ 0.011*life	0.058*exhaust+ 0.044*time+ 0.041*like+ 0.035*tire+ 0.030*know+ 0.019*work+ 0.017*feel+ 0.017*mental+ 0.014*hate+ 0.013*start	0.039*time+ 0.037*trauma+ 0.037*like+ 0.032*exhaust+ 0.023*sleep+ 0.022*anxieties+ 0.020*think+ 0.019*come+ 0.017*fuck+ 0.016*life	0.128*suicide+ 0.075*help+ 0.069*thought+ 0.048*time+ 0.037*life+ 0.034*heal+ 0.029*take+ 0.026*anxieties+ 0.024*know+ 0.021*day
Topic 3	0.027*tire+ 0.022*depress+ 0.019*http+ 0.016*frustrate+ 0.016*go+ 0.013*get+ 0.012*work+ 0.010*live+ 0.010*exhaust+ 0.009*time	0.049*exhaust+ 0.043*go+ 0.021*want+ 0.020*sleep+ 0.018*thing+ 0.017*work+ 0.016*tire+ 0.014*feel+ 0.013*little+ 0.013*life	0.038*tire+ 0.037*anxieties+ 0.034*feel+ 0.034*year+ 0.032*like+ 0.027*hate+ 0.027*trauma+ 0.020*life+ 0.019*exhaust+ 0.018*pain	0.145*suicide+ 0.059*life+ 0.048*tri+ 0.041*go+ 0.034*attempt+ 0.030*http+ 0.023*anxieties+ 0.023*year+ 0.022*hate+ 0.021*thing
Topic 4	0.151*tire+ 0.028*sleep+ 0.026*feel+ 0.025*exhaust+ 0.024*today+ 0.022*work+ 0.016*good+ 0.015*drain+ 0.014*http+ 0.014*go	0.053*fuck+ 0.050*feel+ 0.035*exhaust+ 0.032*hate+ 0.024*shit+ 0.023*tire+ 0.021*like+ 0.020*hurt+ 0.017*year+ 0.016*go	0.031*hate+ 0.031*anxieties+ 0.029*self+ 0.023*thing+ 0.018*time+ 0.018*fuck+ 0.018*shit+ 0.017*feel+ 0.017*think+ 0.016*break	0.069*suicide+ 0.057*hurt+ 0.052*want+ 0.051*thought+ 0.050*go+ 0.040*love+ 0.034*know+ 0.026*like+ 0.026*worst+ 0.026*think
Topic 5	0.060*depress+ 0.030*peopl+ 0.020*suffer+ 0.019*suicide+ 0.019*like+ 0.014*help+ 0.013*year+ 0.013*thing+ 0.012*think+ 0.012*trauma	0.042*like+ 0.042*hurt+ 0.034*tire+ 0.029*want+ 0.027*shit+ 0.027*feel+ 0.022*today+ 0.019*know+ 0.014*get+ 0.012*see	0.059*feel+ 0.030*like+ 0.030*people+ 0.024*today+ 0.023*anxieties+ 0.017*go+ 0.017*want+ 0.016*tell+ 0.015*suffer+ 0.014*know	0.092*suicide+ 0.092*self+ 0.078*like+ 0.058*destruct+ 0.041*hate+ 0.035*feel+ 0.034*fuck+ 0.031*think+ 0.030*sleep+ 0.027*harm

underlying set of words, and each document is a mixture of over a set of topic probabilities.

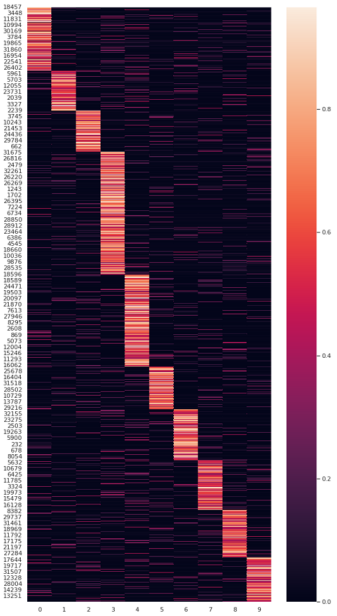


FIGURE 4.8: *Non-depressed* Class Topic Distribution

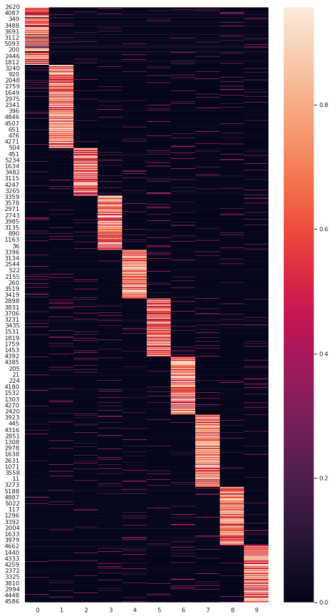


FIGURE 4.9: *Mild* Class Topic Distribution

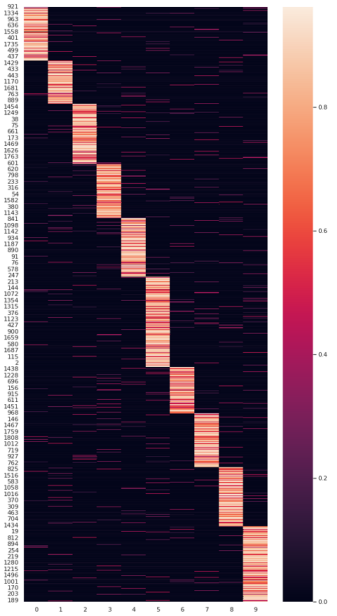


FIGURE 4.10: *Moderate* Class Topic Distribution

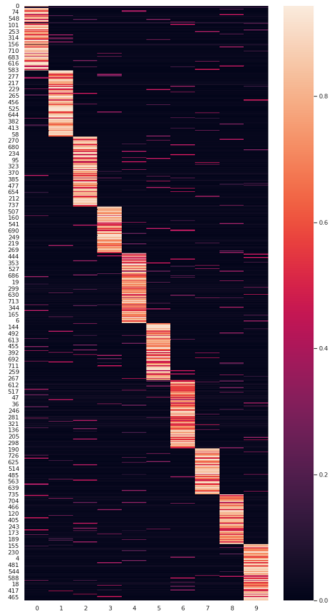


FIGURE 4.11: *Severe* Class Topic Distribution

FIGURE 4.12: Topic Distribution over Documents for all Classes

Table 4.2 shows top 10 words in the top 5 topics of all the classes. Along with the words, the probability distribution of the words over that topic is provided. This provides great insight into the dataset classes. The top topic of the *Non-depressed* class contains words like 'depress','frustrate', etc. which might look similar to the

other classes. But term like 'http' says this might be just a song/entertainment source provided by the users on their feed with captions containing depressive words. Topics in other classes, specially in 'Severe' and 'Moderate' classes lay out valuable insights about the lives of people suffering from depression. The prevalent terms in 'Severely Depressed' class, such as 'suicide', 'thought', 'live', 'anymore', etc., display the untold sufferings and mental struggles of a terminally depressed individual. The topic distribution over the documents of all the classes can be found in Figure 4.12.

Figure 4.12 reveals the prevalence of LDA topics in the data samples. For examples, topic 4 is prevalent among all the documents in *Non-depressed* class, while on the other hand, *severe* class has mostly uniform distribution of topics over its documents.

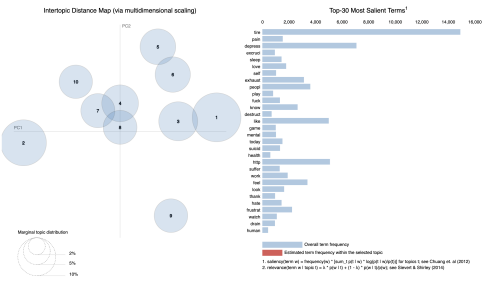


FIGURE 4.13: *Non-depressed* Salient terms and Intertopic distance Map

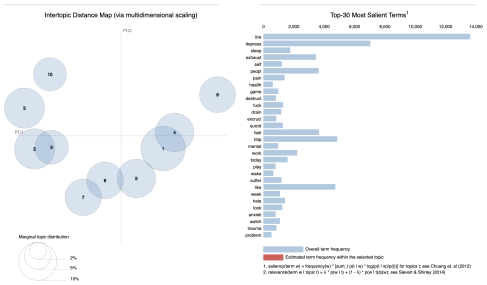


FIGURE 4.14: *Mild*Salient terms and Intertopic distance Map

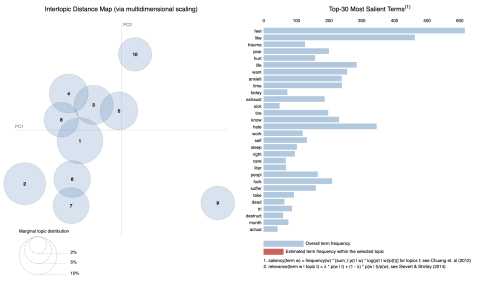


FIGURE 4.15: *Moderate* Salient terms and Intertopic distance Map

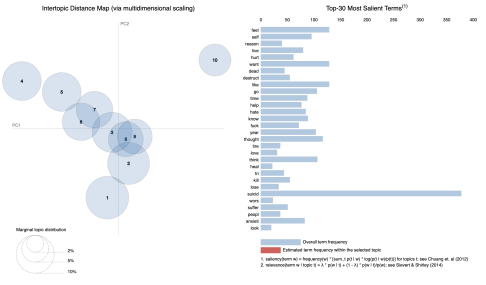


FIGURE 4.16: *Severe* Salient terms and Intertopic distance Map

FIGURE 4.17: Most Salient Terms and Inter-topic distance Map of all Classes

Finally, Figure 4.17 exhibits most salient terms and the intertopic distance map of all the classes. It is worth noticing that the topics in *Moderate* and *Severe* are classes are very well adjacent and connected, whereas the topics in *Non-depressed* class are rather sparse and have lesser inter-connection.

Chapter 5

Experimental Design

After constructing the dataset, it is very crucial to examine if the dataset classes are clearly distinguishable to machine learning frameworks. To examine this characteristic, two attention-based model have been used and a strong baseline result has been provided. Evaluation metrics also play a significant role in examining the quality of the dataset. In this chapter, the election process of baseline models, preprocessing techniques and evaluation metrics for the baseline have been discussed elaborately.

5.1 Baseline Models Selection

Baseline performance on the constructed dataset is evaluated using BERT [51] and DistilBERT [52]. Both models are pretrained using a large amount of unlabeled data in an unsupervised manner where the data source is a concatenation of English Wikipedia and Toronto Book Corpus [61]. BERT-based models are chosen in this study for the following reasons:

- BERT-based models can learn each word's context from the words that appear before and after it. Since effective context understanding from the input representations is very crucial to the task of severity detection from tweets, these models are likely to outperform traditional deep learning based models such as LSTM, BiLSTM or unidirectional transformer based models such as OpenAI GPT [62] where each token is capable of managing only the preceding tokens in the transformer's self-attention layers.

- Previous studies have shown that fine-tuning BERT-based models yield impressive performance in various downstream tasks such as text categorization, question-answering, etc., since these models are pre-trained on a large amount of unlabeled data via leveraging self-supervised learning. BERT-based models have demonstrated impeccable performance in the domain of categorizing social media posts or comments, for example, sentiment analysis of social media posts [63], political social media message categorization [64], rumor identification from tweets [65]. These models mitigate different limitations of previous state-of-the-art language models like ELMO [66] by adopting the transformer encoder instead of the recurrent neural network architecture.

- Implementing a system that can detect the severity of depression from social media texts on devices with limited computational power may be difficult due to the high parameter count of BERT (Base: 110 million), which increases the computational power and time requirements in both the training and inference phases. DistilBERT, on the other hand, alleviates these high requirements by achieving performance comparable to BERT with nearly 40% fewer parameters and 60% less inference time [52], allowing such systems to be implemented on edge devices.

Both BERT and DistilBERT relies on Auto Encoding (AE) language modeling during pre-training since the aim is to understand natural language representations. Although general transformer architecture proposed by [1] utilizes an encoder and a decoder network, BERT and DistilBERT, as pre-training models, only use the encoder to interpret the content of input sequences.

5.1.1 Fine-tuning Classifiers

Fine-tuning the pre-trained model weights in a task specific manner with respect to the tweet texts and their annotated labels is necessary to improve the classification performance considering that they are pre-trained using data from various sources. Below, the fine-tuning procedure for input representation is demonstrated, followed by the training parameters of the experiment.

Input Representation

Before being fed into the pre-trained models for embedding, each tweet text are converted into an acceptable format. A single vector representing the entire input

sentence is required to be passed to a classifier in order to complete the classification operation. BERT-based models use WordPiece tokenizer [67], which works by splitting the input sequence into full forms or word pieces. In case of full form, a word is represented by one token string, whereas, for word pieces, a word is represented by multiple token strings. Using word pieces helps the models to identify related words as they share similar token strings, which is crucial for context understanding. Some special token strings are generated during tokenization to indicate the task type, beginning of input sequence, mask, etc., e.g.,

- '[SEP]' refers to the end of one input sequence and the beginning of another.
- '[CLS]' refers to the classification task.
- '[PAD]' is used to indicate the necessary padding.
- '[UNK]' stands for unknown token.

Classifiers used in this study require the input sequences to be of the same length, i.e., each tweet text should have an equal number of tokens after converting them to token strings. Since a maximum token length of 128 is used, if a comment contains less than 128 tokens, extra '[PAD]' tokens are added at the end of the token sequence. Both BERT and DistilBERT are pre-trained with 30K token vocabularies. So some new input data might appear while fine-tuning, which was not present in the pre-trained vocabulary. In that case, the new input substring is replaced by '[UNK]' token. Subsequently, the final input vector for the models is prepared by converting the token strings to integer token IDs.

Hyper-parameters Selection

Fine-tuning and evaluating the classifiers required the proposed dataset to be splitted into three sets - train, validation, and test. Randomly selected 60% tweets from each class are placed into the train set, and the rest of the tweets are equally distributed among the validation and test sets. Base-uncased[1] versions of the pre-trained models are implemented for fine-tuning with a total of 768 hidden output states. Categorical Cross-Entropy loss function with AdamW optimizer [68] is used that utilizes a fixed weight decay unlike common implementations of Adam optimizer [69]. Considering that the learning rate was set to 3×10^{-5} and 20% of the steps are designated as warm-up steps, the training phase would use the first 20% of the steps to raise the learning rate from 0 to 3×10^{-5}. Here, steps

[1] https://huggingface.co/bert-base-uncased

denote the total number of times when the model weights get updated during the fine-tuning phase.

Both of these models are fine-tuned in a supervised manner for 10 epochs with a training batch size of 16 on the proposed dataset to predict the severity of depression from tweets and achieved a good performance on all four classes. Figure 5.1 depicts the process of predicting the severity of depression using the fine-tuned classifiers from a sample tweet.

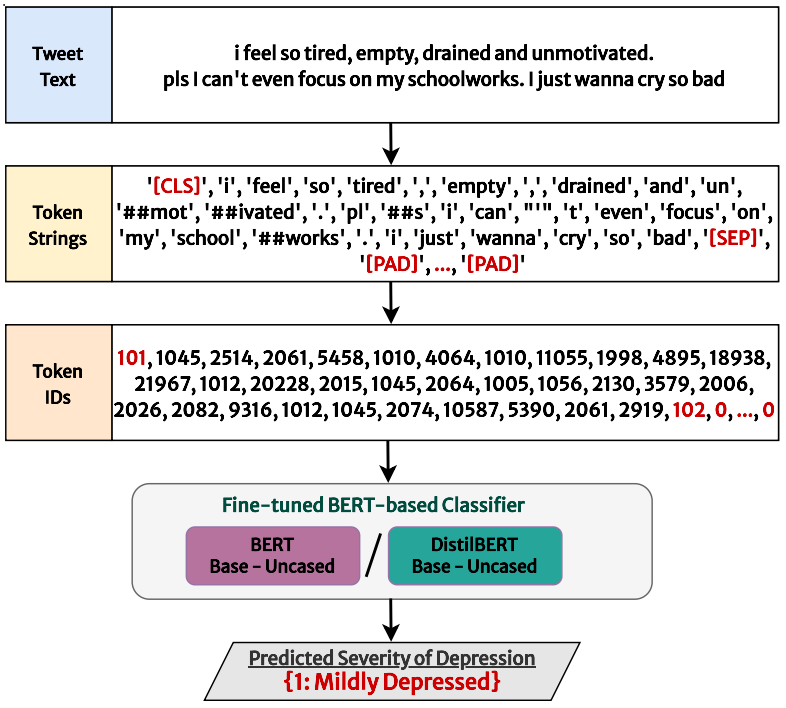


FIGURE 5.1: Severity of Depression Prediction from a Sample Tweet

5.2 Evaluation Metrics

Evaluation metrics play a crucial role in quantifying the performance of a predictive classifier [70]. Since the choice of metrics depends on the characteristic of the dataset, this can often lead to misleading conclusion regarding the experiment. For example, while evaluating an experiment on a highly imbalanced dataset, evaluation metrics such as accuracy, precision, or recall may lead to a conclusion that is practically useless. With imbalanced datasets, it is possible to reach very high accuracy without predicting any useful prediction since the majority predictions are from the densely populated classes [71].

Other widely used evaluation metrics like precision, recall etc. have their own limitations. Precision is about exactness of classification task and relies only on true positive and false positive, it is possible to get a precision score of 1.0 by only one true positive prediction. On the other hand, recall is about completeness and depends solely on true positive and false negative. As a result, predicting all the samples as positive will give a recall of 1.0, whereas precision will be very low.

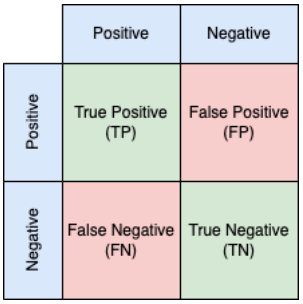


FIGURE 5.2: General Confusion Matrix

To tackle this issue, the Receiver Operating Characteristic (ROC) curve and area under the ROC curve (AUC-ROC) are used as evaluation measures in this work, such that models are evaluated based on how good they are at separating classes. ROC curve is a diagnostic diagram that calculates the False Positive Rate (FPR), and True Positive Rate (TPR) for a series of predictions made by the model at different thresholds to summarize the model's behavior which can be used to analyze the model's ability to discriminate classes. True Positive Rate (TPR) tells what proportion of the positive class get correctly classified by the classifier. False Positive Rate (FPR) tells what proportion of the negative class got incorrectly classified by the classifier. TPR and FPR are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{5.1}$$

$$FPR = \frac{FP}{TN + FP} \tag{5.2}$$

Definition of TP, FN, FP and TN can be derived from Figure 5.2.

The Receiver Operator Characteristic (ROC) curve is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. A model that with no discriminatory power between the classes

will be represented by a diagonal line between FPR 0 and TPR 0 (co-ordinate: 0,0) to FPR 1 and TPR 1 (co-ordinate: 1,1). Points below this line reflect models with less competence than none. A flawless model will be represented as a point in the plot's upper left corner.

Chapter 6

Results and Discussions

In this chapter, the baseline classification result on the dataset is presented. Later portion of the chapter contains discussion on some limitations of this study from the perspective of annotation, that caused some misclassification of samples in the models.

6.1 Classification Performance

According to the results shown in Table 6.1 and Figure 6.3, it can be observed that DistilBERT outperformed BERT in all classes. Since DistilBERT is pretrained under the supervision of its parent model, BERT through knowledge distillation, it is able to preserve 95% performance of the base uncased BERT [52] which is divergent to the experimental results shown in this study. The experiments were conducted in a computationally limited environment with a comparatively smaller batch size and fine-tuned only for 10 epochs.

TABLE 6.1: Performance Comparison of BERT and DistilBERT

Model	Class Name	ROC AUC Score
BERT	Non-depressed	0.763699
	Mild	0.740019
	Moderate	0.748115
	Severe	0.826488
DistilBERT	Non-depressed	0.788841
	Mild	0.747211
	Moderate	0.787959
	Severe	0.866003

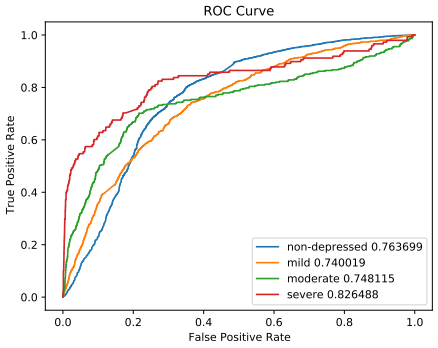


FIGURE 6.1: AUC-ROC for BERT

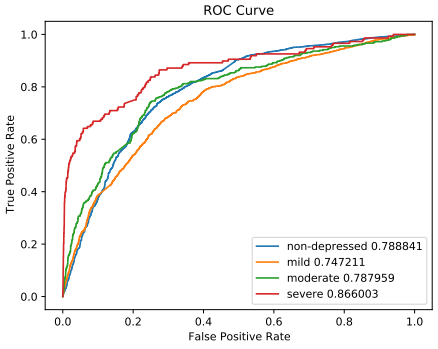


FIGURE 6.2: AUC-ROC for DistilBERT

FIGURE 6.3: Class-wise AUC-ROC curves

It is likely that, BERT will outperform DistilBERT if the models are fine-tuned for higher number of iterations with further hyper-parameter tuning.

TABLE 6.2: Model Predictions for the Terminal Classes

Tweet Text	**Ground Truth**	**Predicted Label**
I knew self destruction ain't the only way...	non-depressed	severe
Yes actually. I feel like it invalidates what queer people go through when they're depressed and attempt/want to attempt commit suicide.	non-depressed	severe
my stomach is killing me. my whole body hurts i'm so exhausted	non-depressed	severe
i inherited a thirst for self destruction and i'm scared of it	severe	non-depressed
Sorry I know what this feels like lost 23 of my best friends in combat... as well as suicide coming back home... depression does suck, but we can do this	severe	non-depressed
I don't like to brag. BUT, I don't think there's a soul on this earth that does self destruction like I do.	severe	non-depressed

As seen from Table **??**, the proposed dataset is mostly comprised of the samples from the *'non-depressed'* class, in which both models showed commendable performance in detecting classes with relatively smaller number of samples for other classes as well. From the confusion matrices in Figure 6.6, it can also be noticed that both the models performed better on the two terminal classes *'non-depressed'* and *'severe'* than the two closely related classes, *'mild'* and *'moderate'*. Upon careful observation, it was found that wrong predictions of the samples were mostly due to models failing to comprehend the contextual meaning of the comments properly and instead generalizing based on specific keywords to predict the final label. For example, as shown in Table 6.2, in few cases where the ground truth is *'non-depressed'* but the predicted label by the models is *'severe'* and vice-versa, most of these cases contain words related to suicide, depression, self-destruction, self-harm, etc. So, this enables the room for further improvement through error analysis.

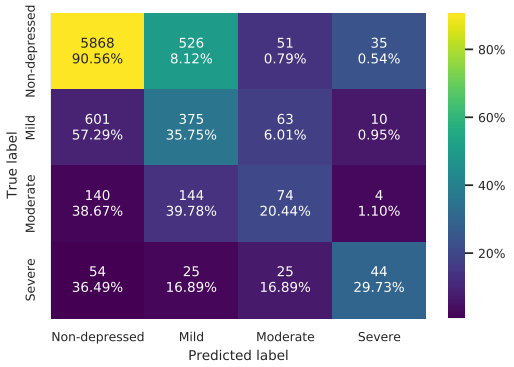


FIGURE 6.4: Confusion Matrix for BERT

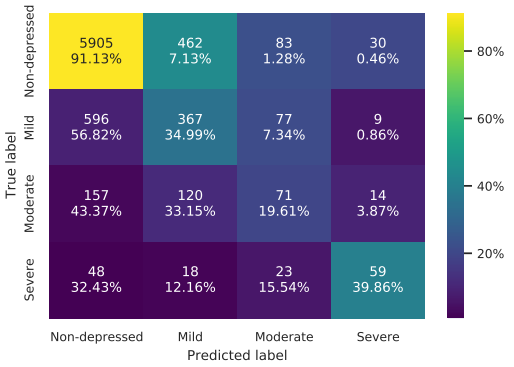


FIGURE 6.5: Confusion Matrix for DistilBERT

FIGURE 6.6: Confusion Matrix Obtained by Evaluating Test Set Using Fine-tuned Classifiers

For the proposed dataset, ROC curves using the test predictions from both of the classifiers is presented in Figure 6.3. These plots are summarized by calculating the area under the ROC curve (AUC-ROC) in Table 6.1. The better performance of DistilBERT over BERT is also distinguishable from the class-wise AUC-ROC curves in Figure 6.3.

6.2 Limitations

Figure 4.2 shows that *non-depressed* and *severe* classes are more condensed towards the complete agreement of the annotators. As these two classes lie on the two polarities and have distinguishable attributes, the annotators were likely to agree more on these two class labels while annotating. The main challenge was to differentiate between the other two classes, i.e., *moderate* and *severe* for their inherent subtleties and congruent attributes. With the tweet corpus being in English, and considering the subtle attributes of the different severities of depression, the dataset was likely to achieve higher annotation quality if the annotation was done by annotators with first-language proficiency in English. As the study requires a large pool of annotators and demands consistent supervision and interaction of the annotators with the collaborator psychologists, it limits the choice of recruiting only English-speaking annotators. This was attempted to be reduced by recruiting annotators with excellent abilities in English and pre-screening was done before the final pool of annotators were selected.

Another challenge that appeared in a similar context for the annotators was to avoid their individual bias while deciding the class labels. The source of the tweets

and their nuances in attributes complicated the annotation task and potentially introduced bias into the dataset. From the manual inspection of the scraped tweet samples, it was observed that the majority of the samples were from the North American region, while all the annotators were from South Asia. This can introduce a clear cultural and geographic bias in the annotation procedure. Though the tweets were presented in isolation to the annotators, without all the related information (i.e., tweet ID, retweets, location, etc.) and without the surrounding context of scraping the tweets, the collaborator psychologists speculated a bias in the annotation as there is a clear cultural and expressional difference between the users and annotators of the tweets. The annotators were reminded several times throughout the annotation process to avoid their personal bias and strictly follow the guidelines laid out by the psychologists, which included a document containing high-level descriptions of the attributes of the classes. This issue of systematic bias is common for large datasets, as addressed by [72], especially for complex multi-class tasks of this kind.

Chapter 7

Conclusion and Future Work

This work introduces a new typology for diagnosing depression severities from social media texts, an attention-based framework to detect depression severities, as well as a unique dataset of labeled tweets with a confidence score for each label. The dataset is constructed based on strong ground truths and clinical validation. The whole framework is expected to help alleviate the scarcity of mental health data to some extent. The description of the process and challenges in creating such a dataset may motivate researchers to collect similar corpora of this scale from other social media and discussion forums. Broader implications of this research may include personalizing and directing preventative and awareness messages by health professionals to the users in need.

The baseline classification result of the dataset was provided by fine tuning two modern pre-trained models, namely BERT and DistilBERT. It is worth noting that several features in the dataset, such as *replies_count* and *retweets_count*, were not used during training, and no pre-processing was performed on the data. Therefore, more accurate classification might be achieved on this dataset by: (1) including a pre-processing technique to clean the data before training, (2) increasing trainable instances by augmentation to eliminate the class imbalance of the dataset, (3) utilizing other features of the dataset during training, (4) fine-tuning more robust pre-trained models [73, 74, 75], etc. Because the data was collected during the post COVID-19 pandemic phase, careful examination of the dataset can provide valuable insight into the impact of the pandemic on people's mental health. Moreover, the DEPTWEET dataset can be expanded by annotating the remaining 2510 data samples for which a class label could not be determined due to annotators'

disagreement. Further work may also include refining the annotation task by including annotators from similar cultural and geographic contexts and exploring the unintended biases in the data and model.

Appendix A

Appendix

The PHQ-9 Questionnaire [21] and The Diagnostic and Statistical Manual of Mental Disorders-5 [22] have been used extensively in this research. These two documents have been added here for the readers. The mood scale provided by BipolarUK and the annotation manual created for annotators to annotate the datafiles are also attached in this section.

Patient Health Questionnaire (PHQ-9)

Name: _____ Date: _____

Over the last 2 weeks, how often have you been bothered by any of the following problems?	Not at all	Several days	More than half the days	Nearly every day
1. Little interest or pleasure in doing things	0	1	2	3
2. Feeling down, depressed, or hopeless	0	1	2	3
3. Trouble falling or staying asleep, or sleeping too much	0	1	2	3
4. Feeling tired or having little energy	0	1	2	3
5. Poor appetite or overeating	0	1	2	3
6. Feeling bad about yourself – or that you are a failure or have let yourself or your family down	0	1	2	3
7. Trouble concentrating on things, such as reading the newspaper or watching television	0	1	2	3
8. Moving or speaking so slowly that other people could have noticed? Or the opposite – being so fidgety or restless that you have been moving around a lot more than usual	0	1	2	3
9. Thoughts that you would be better off dead or of hurting yourself in some way	0	1	2	3

For office coding: Total Score _____ = _____ + _____ + _____

Total Score _____

If you checked off any problems, how difficult have these problems made it for you to do your work, take care of things at home, or get along with other people?

☐ Not difficult at all ☐ Somewhat difficult ☐ Very difficult ☐ Extremely difficult

PHQ-9 Questionnaire (Adapted from : [21])

How to Score the PHQ-9

Major depressive disorder (MDD) is suggested if:

- Of the 9 items, 5 or more are checked as at least 'more than half the days'
- Either item 1 or 2 is checked as at least 'more than half the days'

Other depressive syndrome is suggested if:

- Of the 9 items, between 2 to 4 are checked as at least 'more than half the days'
- Either item 1 or 2 is checked as at least 'more than half the days'

PHQ-9 scores can be used to plan and monitor treatment. To score the instrument, tally the numbers of all the checked responses under each heading (not at all=0, several days=1, more than half the days=2, and nearly every day=3). Add the numbers together to total the score on the bottom of the questionnaire. Interpret the score by using the guide listed below.

Guide for Interpreting PHQ-9 Scores		
Score	Depression Severity	Action
0 - 4	None-minimal	Patient may not need depression treatment.
5 - 9	Mild	Use clinical judgment about treatment, based on patient's duration of symptoms and functional impairment.
10 - 14	Moderate	Use clinical judgment about treatment, based on patient's duration of symptoms and functional impairment.
15 - 19	Moderately severe	Treat using antidepressants, psychotherapy or a combination of treatment.
20 - 27	Severe	Treat using antidepressants with or without psychotherapy.

Functional Health Assessment

The instrument also includes a functional health assessment. This asks the patient how emotional difficulties or problems impact work, life at home, or relationships with other people. Patient response of 'very difficult' or 'extremely difficult' suggest that the patient's functionality is impaired. After treatment begins, functional status and number score can be measured to assess patient improvement.

Note: Depression should not be diagnosed or excluded solely on the basis of a PHQ-9 score. A PHQ-9 score ≥ 10 has a sensitivity of 88% and a specificity of 88% for major depression.[1] Since the questionnaire relies on patient self-report, the practitioner should verify all responses. A definitive diagnosis is made taking into account how well the patient understood the questionnaire, as well as other relevant information from the patient.

Reference: Kroenke K, Spitzer RL, Williams JB. The PHQ-9: Validity of a brief depression severity measure. J Gen Intern Med. 2001;16(9):606-613.

PHQ-9 Questionnaire (Adapted from : [21])

Depressive Disorders

Depressive disorders include disruptive mood dysregulation disorder, major depressive disorder (including major depressive episode), persistent depressive disorder (dysthymia), premenstrual dysphoric disorder, substance/medication-induced depressive disorder, depressive disorder due to another medical condition, other specified depressive disorder, and unspecified depressive disorder. Unlike in DSM-IV, this chapter "Depressive Disorders" has been separated from the previous chapter "Bipolar and Related Disorders." The common feature of all of these disorders is the presence of sad, empty, or irritable mood, accompanied by somatic and cognitive changes that significantly affect the individual's capacity to function. What differs among them are issues of duration, timing, or presumed etiology.

In order to address concerns about the potential for the overdiagnosis of and treatment for bipolar disorder in children, a new diagnosis, disruptive mood dysregulation disorder, referring to the presentation of children with persistent irritability and frequent episodes of extreme behavioral dyscontrol, is added to the depressive disorders for children up to 12 years of age. Its placement in this chapter reflects the finding that children with this symptom pattern typically develop unipolar depressive disorders or anxiety disorders, rather than bipolar disorders, as they mature into adolescence and adulthood.

Major depressive disorder represents the classic condition in this group of disorders. It is characterized by discrete episodes of at least 2 weeks' duration (although most episodes last considerably longer) involving clear-cut changes in affect, cognition, and neurovegetative functions and inter-episode remissions. A diagnosis based on a single episode is possible, although the disorder is a recurrent one in the majority of cases. Careful consideration is given to the delineation of normal sadness and grief from a major depressive episode. Bereavement may induce great suffering, but it does not typically induce an episode of major depressive disorder. When they do occur together, the depressive symptoms and functional impairment tend to be more severe and the prognosis is worse compared with bereavement that is not accompanied by major depressive disorder. Bereavement-related depression tends to occur in persons with other vulnerabilities to depressive disorders, and recovery may be facilitated by antidepressant treatment.

A more chronic form of depression, persistent depressive disorder (dysthymia), can be diagnosed when the mood disturbance continues for at least 2 years in adults or 1 year in children. This diagnosis, new in DSM-5, includes both the DSM-IV diagnostic categories of chronic major depression and dysthymia.

After careful scientific review of the evidence, premenstrual dysphoric disorder has been moved from an appendix of DSM-IV ("Criteria Sets and Axes Provided for Further Study") to Section II of DSM-5. Almost 20 years of additional of research on this condition has confirmed a specific and treatment-responsive form of depressive disorder that begins sometime following ovulation and remits within a few days of menses and has a marked impact on functioning.

A large number of substances of abuse, some prescribed medications, and several medical conditions can be associated with depression-like phenomena. This fact is recognized in the diagnoses of substance/medication-induced depressive disorder and depressive disorder due to another medical condition.

DSM-5 for Depressive Disorder (Adapted from : [22])

should receive one of those diagnoses rather than disruptive mood dysregulation disorder. Children with disruptive mood dysregulation disorder may have symptoms that also meet criteria for an anxiety disorder and can receive both diagnoses, but children whose irritability is manifest only in the context of exacerbation of an anxiety disorder should receive the relevant anxiety disorder diagnosis rather than disruptive mood dysregulation disorder. In addition, children with autism spectrum disorders frequently present with temper outbursts when, for example, their routines are disturbed. In that instance, the temper outbursts would be considered secondary to the autism spectrum disorder, and the child should not receive the diagnosis of disruptive mood dysregulation disorder.

Intermittent explosive disorder. Children with symptoms suggestive of intermittent explosive disorder present with instances of severe temper outbursts, much like children with disruptive mood dysregulation disorder. However, unlike disruptive mood dysregulation disorder, intermittent explosive disorder does not require persistent disruption in mood between outbursts. In addition, intermittent explosive disorder requires only 3 months of active symptoms, in contrast to the 12-month requirement for disruptive mood dysregulation disorder. Thus, these two diagnoses should not be made in the same child. For children with outbursts and intercurrent, persistent irritability, only the diagnosis of disruptive mood dysregulation disorder should be made.

Comorbidity

Rates of comorbidity in disruptive mood dysregulation disorder are extremely high. It is rare to find individuals whose symptoms meet criteria for disruptive mood dysregulation disorder alone. Comorbidity between disruptive mood dysregulation disorder and other DSM-defined syndromes appears higher than for many other pediatric mental illnesses; the strongest overlap is with oppositional defiant disorder. Not only is the overall rate of comorbidity high in disruptive mood dysregulation disorder, but also the range of comorbid illnesses appears particularly diverse. These children typically present to the clinic with a wide range of disruptive behavior, mood, anxiety, and even autism spectrum symptoms and diagnoses. However, children with disruptive mood dysregulation disorder should not have symptoms that meet criteria for bipolar disorder, as in that context, only the bipolar disorder diagnosis should be made. If children have symptoms that meet criteria for oppositional defiant disorder or intermittent explosive disorder *and* disruptive mood dysregulation disorder, only the diagnosis of disruptive mood dysregulation disorder should be assigned. Also, as noted earlier, the diagnosis of disruptive mood dysregulation disorder should not be assigned if the symptoms occur only in an anxiety-provoking context, when the routines of a child with autism spectrum disorder or obsessive-compulsive disorder are disturbed, or in the context of a major depressive episode.

Major Depressive Disorder

Diagnostic Criteria

A. Five (or more) of the following symptoms have been present during the same 2-week period and represent a change from previous functioning; at least one of the symptoms is either (1) depressed mood or (2) loss of interest or pleasure.

   Note: Do not include symptoms that are clearly attributable to another medical condition.

   1. Depressed mood most of the day, nearly every day, as indicated by either subjective report (e.g., feels sad, empty, hopeless) or observation made by others (e.g., appears tearful). (**Note:** In children and adolescents, can be irritable mood.)
   2. Markedly diminished interest or pleasure in all, or almost all, activities most of the day, nearly every day (as indicated by either subjective account or observation).

DSM-5 for Depressive Disorder (Adapted from : [22])

3. Significant weight loss when not dieting or weight gain (e.g., a change of more than 5% of body weight in a month), or decrease or increase in appetite nearly every day. (**Note:** In children, consider failure to make expected weight gain.)
4. Insomnia or hypersomnia nearly every day.
5. Psychomotor agitation or retardation nearly every day (observable by others, not merely subjective feelings of restlessness or being slowed down).
6. Fatigue or loss of energy nearly every day.
7. Feelings of worthlessness or excessive or inappropriate guilt (which may be delusional) nearly every day (not merely self-reproach or guilt about being sick).
8. Diminished ability to think or concentrate, or indecisiveness, nearly every day (either by subjective account or as observed by others).
9. Recurrent thoughts of death (not just fear of dying), recurrent suicidal ideation without a specific plan, or a suicide attempt or a specific plan for committing suicide.

B. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

C. The episode is not attributable to the physiological effects of a substance or to another medical condition.

Note: Criteria A–C represent a major depressive episode.

Note: Responses to a significant loss (e.g., bereavement, financial ruin, losses from a natural disaster, a serious medical illness or disability) may include the feelings of intense sadness, rumination about the loss, insomnia, poor appetite, and weight loss noted in Criterion A, which may resemble a depressive episode. Although such symptoms may be understandable or considered appropriate to the loss, the presence of a major depressive episode in addition to the normal response to a significant loss should also be carefully considered. This decision inevitably requires the exercise of clinical judgment based on the individual's history and the cultural norms for the expression of distress in the context of loss.[1]

D. The occurrence of the major depressive episode is not better explained by schizoaffective disorder, schizophrenia, schizophreniform disorder, delusional disorder, or other specified and unspecified schizophrenia spectrum and other psychotic disorders.

E. There has never been a manic episode or a hypomanic episode.
Note: This exclusion does not apply if all of the manic-like or hypomanic-like episodes are substance-induced or are attributable to the physiological effects of another medical condition.

---

[1] In distinguishing grief from a major depressive episode (MDE), it is useful to consider that in grief the predominant affect is feelings of emptiness and loss, while in MDE it is persistent depressed mood and the inability to anticipate happiness or pleasure. The dysphoria in grief is likely to decrease in intensity over days to weeks and occurs in waves, the so-called pangs of grief. These waves tend to be associated with thoughts or reminders of the deceased. The depressed mood of MDE is more persistent and not tied to specific thoughts or preoccupations. The pain of grief may be accompanied by positive emotions and humor that are uncharacteristic of the pervasive unhappiness and misery characteristic of MDE. The thought content associated with grief generally features a preoccupation with thoughts and memories of the deceased, rather than the self-critical or pessimistic ruminations seen in MDE. In grief, self-esteem is generally preserved, whereas in MDE feelings of worthlessness and self-loathing are common. If self-derogatory ideation is present in grief, it typically involves perceived failings vis-à-vis the deceased (e.g., not visiting frequently enough, not telling the deceased how much he or she was loved). If a bereaved individual thinks about death and dying, such thoughts are generally focused on the deceased and possibly about "joining" the deceased, whereas in MDE such thoughts are focused on ending one's own life because of feeling worthless, undeserving of life, or unable to cope with the pain of depression.

DSM-5 for Depressive Disorder (Adapted from : [22])

Coding and Recording Procedures
The diagnostic code for major depressive disorder is based on whether this is a single or recurrent episode, current severity, presence of psychotic features, and remission status. Current severity and psychotic features are only indicated if full criteria are currently met for a major depressive episode. Remission specifiers are only indicated if the full criteria are not currently met for a major depressive episode. Codes are as follows:

Severity/course specifier	Single episode	Recurrent episode*
Mild (p. 188)	296.21 (F32.0)	296.31 (F33.0)
Moderate (p. 188)	296.22 (F32.1)	296.32 (F33.1)
Severe (p. 188)	296.23 (F32.2)	296.33 (F33.2)
With psychotic features** (p. 186)	296.24 (F32.3)	296.34 (F33.3)
In partial remission (p. 188)	296.25 (F32.4)	296.35 (F33.41)
In full remission (p. 188)	296.26 (F32.5)	296.36 (F33.42)
Unspecified	296.20 (F32.9)	296.30 (F33.9)

*For an episode to be considered recurrent, there must be an interval of at least 2 consecutive months between separate episodes in which criteria are not met for a major depressive episode. The definitions of specifiers are found on the indicated pages.
**If psychotic features are present, code the "with psychotic features" specifier irrespective of episode severity.

In recording the name of a diagnosis, terms should be listed in the following order: major depressive disorder, single or recurrent episode, severity/psychotic/remission specifiers, followed by as many of the following specifiers without codes that apply to the current episode.

Specify:

With anxious distress (p. 184)
With mixed features (pp. 184–185)
With melancholic features (p. 185)
With atypical features (pp. 185–186)
With mood-congruent psychotic features (p. 186)
With mood-incongruent psychotic features (p. 186)
With catatonia (p. 186). **Coding note:** Use additional code 293.89 (F06.1).
With peripartum onset (pp. 186–187)
With seasonal pattern (recurrent episode only) (pp. 187–188)

Diagnostic Features

The criterion symptoms for major depressive disorder must be present nearly every day to be considered present, with the exception of weight change and suicidal ideation. Depressed mood must be present for most of the day, in addition to being present nearly every day. Often insomnia or fatigue is the presenting complaint, and failure to probe for accompanying depressive symptoms will result in underdiagnosis. Sadness may be denied at first but may be elicited through interview or inferred from facial expression and demeanor. With individuals who focus on a somatic complaint, clinicians should determine whether the distress from that complaint is associated with specific depressive symptoms. Fatigue and sleep disturbance are present in a high proportion of cases; psychomotor disturbances are much less common but are indicative of greater overall severity, as is the presence of delusional or near-delusional guilt.

DSM-5 for Depressive Disorder (Adapted from : [22])

The essential feature of a major depressive episode is a period of at least 2 weeks during which there is either depressed mood or the loss of interest or pleasure in nearly all activities (Criterion A). In children and adolescents, the mood may be irritable rather than sad. The individual must also experience at least four additional symptoms drawn from a list that includes changes in appetite or weight, sleep, and psychomotor activity; decreased energy; feelings of worthlessness or guilt; difficulty thinking, concentrating, or making decisions; or recurrent thoughts of death or suicidal ideation or suicide plans or attempts. To count toward a major depressive episode, a symptom must either be newly present or must have clearly worsened compared with the person's pre-episode status. The symptoms must persist for most of the day, nearly every day, for at least 2 consecutive weeks. The episode must be accompanied by clinically significant distress or impairment in social, occupational, or other important areas of functioning. For some individuals with milder episodes, functioning may appear to be normal but requires markedly increased effort.

The mood in a major depressive episode is often described by the person as depressed, sad, hopeless, discouraged, or "down in the dumps" (Criterion A1). In some cases, sadness may be denied at first but may subsequently be elicited by interview (e.g., by pointing out that the individual looks as if he or she is about to cry). In some individuals who complain of feeling "blah," having no feelings, or feeling anxious, the presence of a depressed mood can be inferred from the person's facial expression and demeanor. Some individuals emphasize somatic complaints (e.g., bodily aches and pains) rather than reporting feelings of sadness. Many individuals report or exhibit increased irritability (e.g., persistent anger, a tendency to respond to events with angry outbursts or blaming others, an exaggerated sense of frustration over minor matters). In children and adolescents, an irritable or cranky mood may develop rather than a sad or dejected mood. This presentation should be differentiated from a pattern of irritability when frustrated.

Loss of interest or pleasure is nearly always present, at least to some degree. Individuals may report feeling less interested in hobbies, "not caring anymore," or not feeling any enjoyment in activities that were previously considered pleasurable (Criterion A2). Family members often notice social withdrawal or neglect of pleasurable avocations (e.g., a formerly avid golfer no longer plays, a child who used to enjoy soccer finds excuses not to practice). In some individuals, there is a significant reduction from previous levels of sexual interest or desire.

Appetite change may involve either a reduction or increase. Some depressed individuals report that they have to force themselves to eat. Others may eat more and may crave specific foods (e.g., sweets or other carbohydrates). When appetite changes are severe (in either direction), there may be a significant loss or gain in weight, or, in children, a failure to make expected weight gains may be noted (Criterion A3).

Sleep disturbance may take the form of either difficulty sleeping or sleeping excessively (Criterion A4). When insomnia is present, it typically takes the form of middle insomnia (i.e., waking up during the night and then having difficulty returning to sleep) or terminal insomnia (i.e., waking too early and being unable to return to sleep). Initial insomnia (i.e., difficulty falling asleep) may also occur. Individuals who present with oversleeping (hypersomnia) may experience prolonged sleep episodes at night or increased daytime sleep. Sometimes the reason that the individual seeks treatment is for the disturbed sleep.

Psychomotor changes include agitation (e.g., the inability to sit still, pacing, handwringing; or pulling or rubbing of the skin, clothing, or other objects) or retardation (e.g., slowed speech, thinking, and body movements; increased pauses before answering; speech that is decreased in volume, inflection, amount, or variety of content, or muteness) (Criterion A5). The psychomotor agitation or retardation must be severe enough to be observable by others and not represent merely subjective feelings.

Decreased energy, tiredness, and fatigue are common (Criterion A6). A person may report sustained fatigue without physical exertion. Even the smallest tasks seem to require

DSM-5 for Depressive Disorder (Adapted from : [22])

substantial effort. The efficiency with which tasks are accomplished may be reduced. For example, an individual may complain that washing and dressing in the morning are exhausting and take twice as long as usual.

The sense of worthlessness or guilt associated with a major depressive episode may include unrealistic negative evaluations of one's worth or guilty preoccupations or ruminations over minor past failings (Criterion A7). Such individuals often misinterpret neutral or trivial day-to-day events as evidence of personal defects and have an exaggerated sense of responsibility for untoward events. The sense of worthlessness or guilt may be of delusional proportions (e.g., an individual who is convinced that he or she is personally responsible for world poverty). Blaming oneself for being sick and for failing to meet occupational or interpersonal responsibilities as a result of the depression is very common and, unless delusional, is not considered sufficient to meet this criterion.

Many individuals report impaired ability to think, concentrate, or make even minor decisions (Criterion A8). They may appear easily distracted or complain of memory difficulties. Those engaged in cognitively demanding pursuits are often unable to function. In children, a precipitous drop in grades may reflect poor concentration. In elderly individuals, memory difficulties may be the chief complaint and may be mistaken for early signs of a dementia ("pseudodementia"). When the major depressive episode is successfully treated, the memory problems often fully abate. However, in some individuals, particularly elderly persons, a major depressive episode may sometimes be the initial presentation of an irreversible dementia.

Thoughts of death, suicidal ideation, or suicide attempts (Criterion A9) are common. They may range from a passive wish not to awaken in the morning or a belief that others would be better off if the individual were dead, to transient but recurrent thoughts of committing suicide, to a specific suicide plan. More severely suicidal individuals may have put their affairs in order (e.g., updated wills, settled debts), acquired needed materials (e.g., a rope or a gun), and chosen a location and time to accomplish the suicide. Motivations for suicide may include a desire to give up in the face of perceived insurmountable obstacles, an intense wish to end what is perceived as an unending and excruciatingly painful emotional state, an inability to foresee any enjoyment in life, or the wish to not be a burden to others. The resolution of such thinking may be a more meaningful measure of diminished suicide risk than denial of further plans for suicide.

The evaluation of the symptoms of a major depressive episode is especially difficult when they occur in an individual who also has a general medical condition (e.g., cancer, stroke, myocardial infarction, diabetes, pregnancy). Some of the criterion signs and symptoms of a major depressive episode are identical to those of general medical conditions (e.g., weight loss with untreated diabetes; fatigue with cancer; hypersomnia early in pregnancy; insomnia later in pregnancy or the postpartum). Such symptoms count toward a major depressive diagnosis except when they are clearly and fully attributable to a general medical condition. Nonvegetative symptoms of dysphoria, anhedonia, guilt or worthlessness, impaired concentration or indecision, and suicidal thoughts should be assessed with particular care in such cases. Definitions of major depressive episodes that have been modified to include only these nonvegetative symptoms appear to identify nearly the same individuals as do the full criteria.

Associated Features Supporting Diagnosis

Major depressive disorder is associated with high mortality, much of which is accounted for by suicide; however, it is not the only cause. For example, depressed individuals admitted to nursing homes have a markedly increased likelihood of death in the first year. Individuals frequently present with tearfulness, irritability, brooding, obsessive rumination, anxiety, phobias, excessive worry over physical health, and complaints of pain (e.g., headaches; joint, abdominal, or other pains). In children, separation anxiety may occur.

DSM-5 for Depressive Disorder (Adapted from : [22])

Although an extensive literature exists describing neuroanatomical, neuroendocrinological, and neurophysiological correlates of major depressive disorder, no laboratory test has yielded results of sufficient sensitivity and specificity to be used as a diagnostic tool for this disorder. Until recently, hypothalamic-pituitary-adrenal axis hyperactivity had been the most extensively investigated abnormality associated with major depressive episodes, and it appears to be associated with melancholia, psychotic features, and risks for eventual suicide. Molecular studies have also implicated peripheral factors, including genetic variants in neurotrophic factors and pro-inflammatory cytokines. Additionally, functional magnetic resonance imaging studies provide evidence for functional abnormalities in specific neural systems supporting emotion processing, reward seeking, and emotion regulation in adults with major depression.

Prevalence

Twelve-month prevalence of major depressive disorder in the United States is approximately 7%, with marked differences by age group such that the prevalence in 18- to 29-year-old individuals is threefold higher than the prevalence in individuals age 60 years or older. Females experience 1.5- to 3-fold higher rates than males beginning in early adolescence.

Development and Course

Major depressive disorder may first appear at any age, but the likelihood of onset increases markedly with puberty. In the United States, incidence appears to peak in the 20s; however, first onset in late life is not uncommon.

The course of major depressive disorder is quite variable, such that some individuals rarely, if ever, experience remission (a period of 2 or more months with no symptoms, or only one or two symptoms to no more than a mild degree), while others experience many years with few or no symptoms between discrete episodes. It is important to distinguish individuals who present for treatment during an exacerbation of a chronic depressive illness from those whose symptoms developed recently. Chronicity of depressive symptoms substantially increases the likelihood of underlying personality, anxiety, and substance use disorders and decreases the likelihood that treatment will be followed by full symptom resolution. It is therefore useful to ask individuals presenting with depressive symptoms to identify the last period of at least 2 months during which they were entirely free of depressive symptoms.

Recovery typically begins within 3 months of onset for two in five individuals with major depression and within 1 year for four in five individuals. Recency of onset is a strong determinant of the likelihood of near-term recovery, and many individuals who have been depressed only for several months can be expected to recover spontaneously. Features associated with lower recovery rates, other than current episode duration, include psychotic features, prominent anxiety, personality disorders, and symptom severity.

The risk of recurrence becomes progessively lower over time as the duration of remission increases. The risk is higher in individuals whose preceding episode was severe, in younger individuals, and in individuals who have already experienced multiple episodes. The persistence of even mild depressive symptoms during remission is a powerful predictor of recurrence.

Many bipolar illnesses begin with one or more depressive episodes, and a substantial proportion of individuals who initially appear to have major depressive disorder will prove, in time, to instead have a bipolar disorder. This is more likely in individuals with onset of the illness in adolescence, those with psychotic features, and those with a family history of bipolar illness. The presence of a "with mixed features" specifier also increases the risk for future manic or hypomanic diagnosis. Major depressive disorder, particularly with psychotic features, may also transition into schizophrenia, a change that is much more frequent than the reverse.

DSM-5 for Depressive Disorder (Adapted from : [22])

Despite consistent differences between genders in prevalence rates for depressive disorders, there appear to be no clear differences by gender in phenomenology, course, or treatment response. Similarly, there are no clear effects of current age on the course or treatment response of major depressive disorder. Some symptom differences exist, though, such that hypersomnia and hyperphagia are more likely in younger individuals, and melancholic symptoms, particularly psychomotor disturbances, are more common in older individuals. The likelihood of suicide attempts lessens in middle and late life, although the risk of completed suicide does not. Depressions with earlier ages at onset are more familial and more likely to involve personality disturbances. The course of major depressive disorder within individuals does not generally change with aging. Mean times to recovery appear to be stable over long periods, and the likelihood of being in an episode does not generally increase or decrease with time.

Risk and Prognostic Factors

Temperamental. Neuroticism (negative affectivity) is a well-established risk factor for the onset of major depressive disorder, and high levels appear to render individuals more likely to develop depressive episodes in response to stressful life events.

Environmental. Adverse childhood experiences, particularly when there are multiple experiences of diverse types, constitute a set of potent risk factors for major depressive disorder. Stressful life events are well recognized as precipitants of major depressive episodes, but the presence or absence of adverse life events near the onset of episodes does not appear to provide a useful guide to prognosis or treatment selection.

Genetic and physiological. First-degree family members of individuals with major depressive disorder have a risk for major depressive disorder two- to fourfold higher than that of the general population. Relative risks appear to be higher for early-onset and recurrent forms. Heritability is approximately 40%, and the personality trait neuroticism accounts for a substantial portion of this genetic liability.

Course modifiers. Essentially all major nonmood disorders increase the risk of an individual developing depression. Major depressive episodes that develop against the background of another disorder often follow a more refractory course. Substance use, anxiety, and borderline personality disorders are among the most common of these, and the presenting depressive symptoms may obscure and delay their recognition. However, sustained clinical improvement in depressive symptoms may depend on the appropriate treatment of underlying illnesses. Chronic or disabling medical conditions also increase risks for major depressive episodes. Such prevalent illnesses as diabetes, morbid obesity, and cardiovascular disease are often complicated by depressive episodes, and these episodes are more likely to become chronic than are depressive episodes in medically healthy individuals.

Culture-Related Diagnostic Issues

Surveys of major depressive disorder across diverse cultures have shown sevenfold differences in 12-month prevalence rates but much more consistency in female-to-male ratio, mean ages at onset, and the degree to which presence of the disorder raises the likelihood of comorbid substance abuse. While these findings suggest substantial cultural differences in the expression of major depressive disorder, they do not permit simple linkages between particular cultures and the likelihood of specific symptoms. Rather, clinicians should be aware that in most countries the majority of cases of depression go unrecognized in primary care settings and that in many cultures, somatic symptoms are very likely to constitute the presenting complaint. Among the Criterion A symptoms, insomnia and loss of energy are the most uniformly reported.

DSM-5 for Depressive Disorder (Adapted from : [22])

Gender-Related Diagnostic Issues

Although the most reproducible finding in the epidemiology of major depressive disorder has been a higher prevalence in females, there are no clear differences between genders in symptoms, course, treatment response, or functional consequences. In women, the risk for suicide attempts is higher, and the risk for suicide completion is lower. The disparity in suicide rate by gender is not as great among those with depressive disorders as it is in the population as a whole.

Suicide Risk

The possibility of suicidal behavior exists at all times during major depressive episodes. The most consistently described risk factor is a past history of suicide attempts or threats, but it should be remembered that most completed suicides are not preceded by unsuccessful attempts. Other features associated with an increased risk for completed suicide include male sex, being single or living alone, and having prominent feelings of hopelessness. The presence of borderline personality disorder markedly increases risk for future suicide attempts.

Functional Consequences of Major Depressive Disorder

Many of the functional consequences of major depressive disorder derive from individual symptoms. Impairment can be very mild, such that many of those who interact with the affected individual are unaware of depressive symptoms. Impairment may, however, range to complete incapacity such that the depressed individual is unable to attend to basic self-care needs or is mute or catatonic. Among individuals seen in general medical settings, those with major depressive disorder have more pain and physical illness and greater decreases in physical, social, and role functioning.

Differential Diagnosis

Manic episodes with irritable mood or mixed episodes. Major depressive episodes with prominent irritable mood may be difficult to distinguish from manic episodes with irritable mood or from mixed episodes. This distinction requires a careful clinical evaluation of the presence of manic symptoms.

Mood disorder due to another medical condition. A major depressive episode is the appropriate diagnosis if the mood disturbance is not judged, based on individual history, physical examination, and laboratory findings, to be the direct pathophysiological consequence of a specific medical condition (e.g., multiple sclerosis, stroke, hypothyroidism).

Substance/medication-induced depressive or bipolar disorder. This disorder is distinguished from major depressive disorder by the fact that a substance (e.g., a drug of abuse, a medication, a toxin) appears to be etiologically related to the mood disturbance. For example, depressed mood that occurs only in the context of withdrawal from cocaine would be diagnosed as cocaine-induced depressive disorder.

Attention-deficit/hyperactivity disorder. Distractibility and low frustration tolerance can occur in both attention-deficit/ hyperactivity disorder and a major depressive episode; if the criteria are met for both, attention-deficit/hyperactivity disorder may be diagnosed in addition to the mood disorder. However, the clinician must be cautious not to overdiagnose a major depressive episode in children with attention-deficit/hyperactivity disorder whose disturbance in mood is characterized by irritability rather than by sadness or loss of interest.

DSM-5 for Depressive Disorder (Adapted from : [22])

Adjustment disorder with depressed mood. A major depressive episode that occurs in response to a psychosocial stressor is distinguished from adjustment disorder with depressed mood by the fact that the full criteria for a major depressive episode are not met in adjustment disorder.

Sadness. Finally, periods of sadness are inherent aspects of the human experience. These periods should not be diagnosed as a major depressive episode unless criteria are met for severity (i.e., five out of nine symptoms), duration (i.e., most of the day, nearly every day for at least 2 weeks), and clinically significant distress or impairment. The diagnosis other specified depressive disorder may be appropriate for presentations of depressed mood with clinically significant impairment that do not meet criteria for duration or severity.

Comorbidity

Other disorders with which major depressive disorder frequently co-occurs are substance-related disorders, panic disorder, obsessive-compulsive disorder, anorexia nervosa, bulimia nervosa, and borderline personality disorder.

Persistent Depressive Disorder (Dysthymia)

Diagnostic Criteria **300.4 (F34.1)**

This disorder represents a consolidation of DSM-IV-defined chronic major depressive disorder and dysthymic disorder.

A. Depressed mood for most of the day, for more days than not, as indicated by either subjective account or observation by others, for at least 2 years.

   Note: In children and adolescents, mood can be irritable and duration must be at least 1 year.

B. Presence, while depressed, of two (or more) of the following:

   1. Poor appetite or overeating.
   2. Insomnia or hypersomnia.
   3. Low energy or fatigue.
   4. Low self-esteem.
   5. Poor concentration or difficulty making decisions.
   6. Feelings of hopelessness.

C. During the 2-year period (1 year for children or adolescents) of the disturbance, the individual has never been without the symptoms in Criteria A and B for more than 2 months at a time.

D. Criteria for a major depressive disorder may be continuously present for 2 years.

E. There has never been a manic episode or a hypomanic episode, and criteria have never been met for cyclothymic disorder.

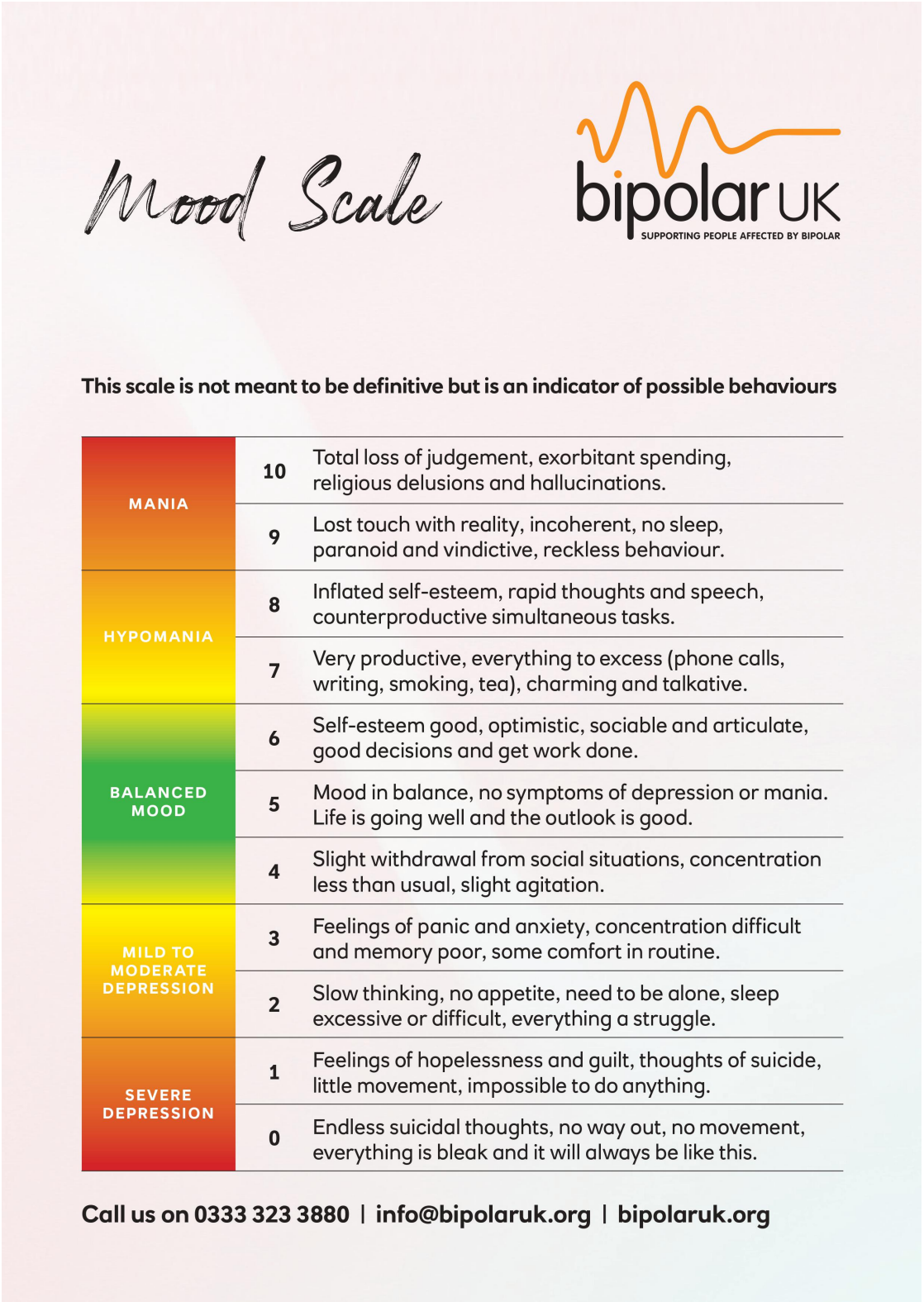
F. The disturbance is not better explained by a persistent schizoaffective disorder, schizophrenia, delusional disorder, or other specified or unspecified schizophrenia spectrum and other psychotic disorder.

G. The symptoms are not attributable to the physiological effects of a substance (e.g., a drug of abuse, a medication) or another medical condition (e.g. hypothyroidism).

H. The symptoms cause clinically significant distress or impairment in social, occupational, or other important areas of functioning.

Note: Because the criteria for a major depressive episode include four symptoms that are absent from the symptom list for persistent depressive disorder (dysthymia), a very limited

DSM-5 for Depressive Disorder (Adapted from : [22])



BipolarUK Mood Scale (Adapted from : [76])

Overview
Hello Annotator! By this point, you're already provided with a range of data that contains tweets from people all over the world. The datafile contains 3 columns- Tweet Text, Annotator Label and Possible Labels. The Annotator Label field is empty. In this job, you will be asked to read each tweet and express your opinion about the severity of depression in that tweet. You have to determine whether the tweet displays a range of characteristics that may contain symptoms of depression. These labels include Non-depressed (0), Mildly Depressed (1), Moderately Depressed (2), Severely Depressed (3). The data collected here will be used to help build tools to detect the severity of depression and help those in need.

Note

- Please bear in mind that we are not asking whether you agree or disagree with the substance of each tweet. Do your best to ignore your own opinion on the substantive idea or claim made in the comment when labeling the tweets.
- Please be sure to read the full text of the tweet before labeling it. Sometimes a part of a tweet might not display any symptom of depression, but the whole tweet might provide a completely different picture.

All of the comments you will see are real comments posted by users in online conversations. The data collected here will be used to help build tools that can detect the severity of depression online. To annotate this, the first thing you should know is what expresses depression and what doesn't.

It's common to feel down from time to time, but depression is a separate condition that should be treated with care. Aside from causing a general feeling of sadness, depression is known for causing feelings of hopelessness that don't seem to go away. Your annotation can have the following classes-

1. **Non-depressed Tweets**

   What are the characteristics of a non-depressed tweet?
   - Tweets that express a person's joy or delight
   - A generalized statement about depression or anxiety. So just containing the words 'depression', 'depressed', or 'exhausted' doesn't necessarily mean that it's a depressed person's tweet.
   - Tweets where a person talks about someone else's depression. As we aim to categorize depressed tweets on a personal level, this is an important point to address.
   - Tweets that express casual tiredness or exhaustion due to hard work. But if the tweet indicates that the exhaustion has been going on for quite some time, it might lead to a feeling of struggle and anxiety. Then understanding the context and reevaluation of the label is necessary.
   - Tweets that express sadness originating from a source of entertainment. For example, if someone talks about his/her sadness watching a football match or a tv series in a tweet, then it's just casual frustration, not depression.
   - Temporary hopelessness is not depression.

2. **Mildly Depressed Tweets**

   A tweet that expresses hopelessness or a feeling of disinterest that persists for quite some time and doesn't seem to go away, can be labeled as a mildly depressed tweet. A mildly depressed tweet may include:

   - hopelessness
   - feelings of guilt and despair
   - a loss of interest in activities you once enjoyed
   - difficulties concentrating at work
   - a lack of motivation
   - a sudden disinterest in socializing
   - daytime sleepiness and fatigue
   - insomnia

Annotation Manual for Annotators

- appetite changes
- weight changes
- reckless behavior, such as abuse of alcohol and drugs, or gambling

These feelings are not just casual but persistent and seem to come in the way of a person's normal flow of life.

3. **Moderately Depressed Tweets**

Moderate depression is the next level up from mild cases. Moderate and mild depression share similar symptoms. The greatest difference is that the symptoms of moderate depression are severe enough to cause problems at home and work. You may also find significant difficulties in the social life of these sorts of tweets. Additionally, moderate depression may include:

- problems with self-esteem
- reduced productivity
- feelings of worthlessness
- increased sensitivities
- excessive worrying

Hint: In most cases, moderately depressed tweets will include:
- ➔ the social anxiety of the person
- ➔ problems in completing day-to-do chores
- ➔ negative thinking about everyone including himself/herself

4. **Severely Depressed Tweets**

The symptoms in this category are more severe and noticeable. It may include:

- delusions
- feelings of near-unconsciousness or insensibility
- hallucinations
- suicidal thoughts or behaviors

Characteristics of Levels of Depression

Label	Depression Levels	Characteristics	Example Tweets
0	Non-Depressed	A generalized statement about depression, talking about someone's depression, casual tiredness or exhaustion due to hard work that is not persistent, etc.	1. If it is a suicide as announced by police in 15 mins nd as per bullyweed people the fake 'Depression' theory, why time of death is not mentioned in PM report ❓❓❓ SUSHANT DISHA DOUBLE MURDER 2. When I first saw this fight I was thinking, "This is so unfair. He just fought his way up 4 floors in a single take, he must be exhausted..." 3. I'm so exhausted, but still have so much work to do because i had hella meetings today and no free time to prep😭
1	Mildly Depressed	Feelings of panic and anxiety, concentration difficulty and memory poor, some comfort in routine, a sudden disinterest in socializing, a loss of interest in	1. Here and there, work was making me exhausted, and it was starting to be looong days. And I was spending too much money so I took a break for a little lol wish I could've taken you guys on break with me 🤣 how have you been how's everything

		activities you once enjoyed, feelings of guilt and despair	on here? Anything new? 2. im getting tired of chasing you back, trying to fix us; at one point, i see like it's always me who apologize first 3. At this point I'm only holding onto you to feel a little better because I'm tired, I just wanna cry because I can't take this anymore, I'm just lost and exhausted, I feel so..
2	Moderately Depressed	Slow thinking, no appetite, need to be alone, sleep excessive or difficult, everything is a struggle.	1. I've struggled a few years ago with a depression as well for years & it did set me back professionally, personally & physically. I was scientifically aware of the "normal" choice for a professional support as I do for any other part that hurts me in my health. 2. i just feel so drained and used by everyone 3. Everything I do now takes days to complete. Its my new reality. I've lost so much strength over the past 5 months :/ every time I push myself I mess up my progress. Depression hittin hard lately, even in my dreams. https://t.co/35sg6aUd1g
3	Severely Depressed	Endless suicidal thoughts, no way out, no movement, everything is bleak and it will always be like this. Feelings of hopelessness and guilt, thoughts of suicide, little movement, impossible to do anything	1. For I wanna share my truth with you. I've lived with high functioning depression all my life. I have anxiety, social anxiety. I tried to commit suicide when I was 16, 27, & 30 years old. But I survived. I also battle with intrusive thoughts every day. 2. i feel so depressed. i should give up and die. nothing matters anyway. i will never get the healthcare i desperately needs. whats the use? 3. Haven't felt this depressed since I last tried to hurt myself in 2009. Didn't think feeling low like this would return. Luckily I'm not at state where I would try to hurt myself but it's tough

Sources:

[1] Tolentino, Julio C, and Sergio L Schmidt. "DSM-5 Criteria and Depression Severity: Implications for Clinical Practice." Frontiers in psychiatry vol. 9 450. 2 Oct. 2018, doi:10.3389/fpsyt.2018.00450

Annotation Manual for Annotators

Bibliography

[1] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[2] A. Thieme, D. Belgrave, and G. Doherty, "Machine learning in mental health: A systematic review of the hci literature to support the development of effective and implementable ml systems," vol. 27, no. 5, aug 2020. [Online]. Available: https://doi.org/10.1145/3398069

[3] D. M. Low, L. Rumker, T. Talkar, J. B. Torous, G. A. Cecchi, and S. S. Ghosh, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study," *Journal of Medical Internet Research*, vol. 22, 2020.

[4] R. Kadison and T. F. DiGeronimo, "College of the overwhelmed: The campus mental health crisis and what to do about it," *Jossey-Bass*, 2004.

[5] S. R. Furr, J. S. Westefeld, G. N. McConnell, and J. Jenkins, "Suicide and depression among college students: A decade later," *Professional Psychology: Research and Practice*, vol. 32, pp. 97–100, 2001.

[6] A. L. Calear and H. Christensen, "Systematic review of school-based prevention and early intervention programs for depression," *Journal of adolescence*, vol. 33 3, pp. 429–38, 2010.

[7] S. Bucci, M. Schwannauer, and N. Berry, "The digital revolution and its impact on mental health care," *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 92, no. 2, pp. 277–297, 2019. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/papt.12222

[8] M. D. Choudhury, S. Counts, and E. Horvitz, "Social media as a measurement tool of depression in populations," in *WebSci*, 2013.

[9] S. Amir, G. A. Coppersmith, P. Carvalho, M. J. Silva, and B. C. Wallace, "Quantifying mental health from social media with neural user embeddings," *ArXiv*, vol. abs/1705.00335, 2017.

[10] N. Ofek, G. Katz, B. Shapira, and Y. Bar-Zev, "Sentiment Analysis in Transcribed Utterances," in *Advances in Knowledge Discovery and Data Mining*, T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, and H. Motoda, Eds. Cham: Springer International Publishing, 2015, pp. 27–38. [Online]. Available: https://rd.springer.com/chapter/10.1007/978-3-319-18032-8_3

[11] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific reports*, vol. 10, no. 1, pp. 1–6, 2020.

[12] A. H. Yazdavar, H. S. Al-Olimat, M. Ebrahimi, G. Bajaj, T. Banerjee, K. Thirunarayan, J. Pathak, and A. Sheth, "Semi-supervised approach to monitoring clinical depressive symptoms in social media," *Proceedings of the 2017 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, 2017.

[13] K. Harrigian, C. A. Aguirre, and M. Dredze, "On the state of social media data for mental health research," *ArXiv*, vol. abs/2011.05233, 2021.

[14] A. Konrad, V. Bellotti, N. Crenshaw, S. Tucker, L. Nelson, H. Du, P. Pirolli, and S. Whittaker, "Finding the adaptive sweet spot: Balancing compliance and achievement in automated stress reduction," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2015.

[15] J. Vincent, "Facebook is using AI to spot users with suicidal thoughts and send them help," Nov. 2017. [Online]. Available: https://www.theverge.com/2017/11/28/16709224/facebook-suicidal-thoughts-ai-help

[16] J. B. Torous, M. S. Keshavan, and T. G. Gutheil, "Promise and perils of digital psychiatry." *Asian journal of psychiatry*, vol. 10, pp. 120–2, 2014.

[17] S. Singh, D. Roy, K. Sinha, S. Parveen, G. Sharma, and G. Joshi, "Impact of COVID-19 and lockdown on mental health of children and adolescents: A narrative review with recommendations," *Psychiatry Research*, vol. 293, p. 113429, Nov. 2020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016517812031725X

[18] U. Ahmed, S. K. Mukhiya, G. Srivastava, Y. Lamo, and J. C. Lin, "Attention-based deep entropy active learning using lexical algorithm for mental health treatment," *Frontiers in Psychology*, vol. 12, 2021.

[19] S. K. Ernala, M. L. Birnbaum, K. A. Candan, A. F. Rizvi, W. A. Sterling, J. M. Kane, and M. De Choudhury, "Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: Association for Computing Machinery, 2019, p. 1–16. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3290605.3300364

[20] J. C. Tolentino and S. L. Schmidt, "DSM-5 Criteria and Depression Severity: Implications for Clinical Practice," *Frontiers in Psychiatry*, vol. 9, p. 450, Oct. 2018. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyt.2018.00450/full

[21] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1525-1497.2001.016009606.x

[22] G. Arbanas, "Diagnostic and Statistical Manual of Mental Disorders (DSM-5)," *Alcoholism and psychiatry research*, vol. 51, pp. 61–64, 2015. [Online]. Available: https://www.amberton.edu/media/Syllabi/Spring%202022/Graduate/CSL6798_E1.pdf

[23] M. G. Haselton, D. Nettle, and P. W. Andrews, "The evolution of cognitive bias," 2015.

[24] C. Fuchs, *Culture and Economy in the Age of Social Media*. New York: Routledge, 2015. [Online]. Available: https://www.taylorfrancis.com/books/mono/10.4324/9781315733517/culture-economy-age-social-media-christian-fuchs

[25] L. S. Radloff, "The ces-d scale: A self-report depression scale for research in the general population," *Applied psychological measurement*, vol. 1, no. 3, pp. 385–401, 1977.

[26] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, G. J. Park, M. Sap, D. Stillwell, M. Kosinski, and L. H. Ungar, "Towards assessing changes in degree of depression through facebook," in *CLPsych@ACL*, 2014.

[27] G. A. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," in *CLPsych@ACL*, 2014.

[28] G. A. Coppersmith, M. Dredze, C. Harman, K. Hollingshead, and M. Mitchell, "Clpsych 2015 shared task: Depression and ptsd on twitter," in *CLPsych@HLT-NAACL*, 2015.

[29] T. Pedersen, "Screening Twitter Users for Depression and PTSD with Lexical Decision Lists," in *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*. Denver, Colorado: Association for Computational Linguistics, Jun. 5 2015, pp. 46–53. [Online]. Available: https://aclanthology.org/W15-1206

[30] 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*. San Diego, CA, USA: Association for Computational Linguistics, Jun. 2016, pp. 106–117. [Online]. Available: https://aclanthology.org/W16-0311

[31] X. Chen, M. D. Sykora, T. W. Jackson, and S. Elayan, "What about mood swings: Identifying depression on twitter with temporal measures of emotions," *Companion Proceedings of the The Web Conference 2018*, 2018.

[32] X. Tian, G. Yu, and F. He, "An analysis of sleep complaints on Sina Weibo," *Computers in Human Behavior*, vol. 62, pp. 230–235, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563216302795

[33] Z. Jamil, D. Inkpen, P. Buddhitha, and K. White, "Monitoring tweets for depression to detect at-risk users," in *CLPsych@ACL*, 2017.

[34] S. K. Mukhiya, U. Ahmed, F. Rabbi, K. I. Pun, and Y. Lamo, "Adaptation of idpt system based on patient-authored text data using nlp," *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, pp. 226–232, 2020.

[35] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. A. Schumm, "Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework," in *COLING*, 2020.

[36] M. Gaur, U. Kursuncu, A. Alambo, A. Sheth, R. Daniulaityte, K. Thirunarayan, and J. Pathak, ""let me tell you about your mental

health!"": Contextualized classification of reddit posts to dsm-5 for web-based intervention," *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, 2018.

[37] P. A. Cavazos-Rehg, M. J. Krauss, S. Sowles, S. Connolly, C. Rosas, M. Bharadwaj, and L. J. Bierut, "A content analysis of depression-related tweets," *Computers in Human Behavior*, vol. 54, pp. 351–357, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563215300996

[38] T. Nguyen, B. O'Dea, M. E. Larsen, D. Q. Phung, S. Venkatesh, and H. Christensen, "Using linguistic and topic analysis to classify sub-groups of online depression communities," *Multimedia Tools and Applications*, vol. 76, pp. 10 653–10 676, 2015.

[39] I. Fatima, H. Mukhtar, H. F. Ahmad, and K. Rajpoot, "Analysis of user-generated content from online social communities to characterise and predict depression degree," *Journal of Information Science*, vol. 44, pp. 683 – 695, 2018.

[40] D. J. Joshi, M. Makhija, Y. Nabar, N. Nehete, and M. S. Patwardhan, "Mental health analysis using deep learning for feature extraction," *Proceedings of the ACM India Joint International Conference on Data Science and Management of Data*, 2018.

[41] A. L. Nobles, J. J. Glenn, K. Kowsari, B. A. Teachman, and L. E. Barnes, "Identification of imminent suicide risk among young adults using text messages," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 2018.

[42] J. P. Pestian, P. Matykiewicz, J. Grupp-Phelan, S. A. Lavanier, J. Combs, and R. A. Kowatch, "Using natural language processing to classify suicide notes," *AMIA ... Annual Symposium proceedings. AMIA Symposium*, p. 1091, 2008.

[43] M. Adamou, G. Antoniou, E. Greasidou, V. Lagani, P. Charonyktakis, and I. Tsamardinos, "Mining free-text medical notes for suicide risk assessment," *Proceedings of the 10th Hellenic Conference on Artificial Intelligence*, 2018.

[44] P. Wilbourne, G. Dexter, and D. Shoup, "Research driven: Sibly and the transformation of mental health and wellness," *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, 2018.

[45] K. Saha and M. D. Choudhury, "Modeling stress with social media around incidents of gun violence on college campuses," *Proceedings of the ACM on Human-Computer Interaction*, vol. 1, pp. 1 – 27, 2017.

[46] R. Kavuluru, M. Ramos-Morales, T. Holaday, A. G. Williams, L. Haye, and J. Cerel, "Classification of helpful comments on online suicide watch forums," *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2016.

[47] P. Resnik, W. Armstrong, L. M. B. Claudino, T. Nguyen, V.-A. Nguyen, and J. L. Boyd-Graber, "Beyond lda: Exploring supervised topic modeling for depression-related language in twitter," in *CLPsych@HLT-NAACL*, 2015.

[48] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[49] T. Luong, H. Pham, and C. D. Manning, "Effective approaches to attention-based neural machine translation," in *EMNLP*, 2015.

[50] K. Xu, J. Ba, R. Kiros, K. Cho, A. C. Courville, R. Salakhutdinov, R. S. Zemel, and Y. Bengio, "Show, attend and tell: Neural image caption generation with visual attention," in *ICML*, 2015.

[51] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[52] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[53] W. H. Organization, "Icd-10: The icd-10 classification of mental and behavioural disorders: diagnostic criteria for research," in *ICD-10: the ICD-10 classification of mental and behavioural disorders: diagnostic criteria for research*, 1993, pp. xiii–248.

[54] G. A. Miller, "WordNet: A Lexical Database for English," *Commun. ACM*, vol. 38, no. 11, p. 39–41, nov 1995. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/219717.219748

[55] I. Price, J. Gifford-Moore, J. Flemming, S. Musker, M. Roichman, G. Sylvain, N. Thain, L. Dixon, and J. Sorensen, "Six Attributes of Unhealthy Conversations," in *Proceedings of the Fourth Workshop on Online Abuse and*

Harms. Online: Association for Computational Linguistics, Nov. 2020, pp. 114–124. [Online]. Available: https://aclanthology.org/2020.alw-1.15

[56] A. Vermeulen, H. Vandebosch, and W. Heirman, "#Smiling, #venting, or both? Adolescents' social sharing of emotions on social media," *Computers in Human Behavior*, vol. 84, pp. 211–219, 2018. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0747563218300803

[57] J. L. Fleiss, B. Levin, and M. C. Paik, *Statistical Methods for Rates and Proportions*, 3rd ed., ser. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., 2013. [Online]. Available: https://onlinelibrary.wiley.com/doi/book/10.1002/0471445428

[58] K. L. Gwet, *Handbook of Inter-rater Reliability: The Definitive Guide to Measuring the Extent of Agreement Among Raters*. Advanced Analytics, LLC, 2014.

[59] Leard Statistics, "Fleiss' kappa in SPSS statistics," 2019. [Online]. Available: https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php

[60] J. O. Salminen, H. A. Al-Merekhi, P. Dey, and B. J. Jansen, "Inter-Rater Agreement for Social Computing Studies," in *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*. IEEE, 2018, pp. 80–87. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/8554744

[61] Y. Zhu, R. Kiros, R. S. Zemel, R. Salakhutdinov, R. Urtasun, A. Torralba, and S. Fidler, "Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books," in *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*. IEEE Computer Society, 2015, pp. 19–27. [Online]. Available: https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html

[62] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving Language Understanding by Generative Pre-Training," 2018. [Online]. Available: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf

[63] V. Moshkin, A. Konstantinov, and N. Yarushkina, "Application of the BERT Language Model for Sentiment Analysis of Social Network Posts," in *Artificial Intelligence*, S. O. Kuznetsov, A. I. Panov, and K. S. Yakovlev,

Eds. Cham: Springer International Publishing, 2020, pp. 274–283. [Online]. Available: https://rd.springer.com/chapter/10.1007/978-3-030-59535-7_20

[64] S. Gupta, S. Bolden, J. Kachhadia, A. Korsunska, and J. Stromer-Galley, "PoliBERT: Classifying political social media messages with BERT," in *Social, Cultural and Behavioral Modeling (SBP-BRIMS 2020) conference. Washington, DC*, 2020. [Online]. Available: https://news.illuminating.ischool.syr.edu/2020/11/24/polibert-classifying-political-social-media-messages-with-bert/

[65] R. Anggrainingsih, G. M. Hassan, and A. Datta, "BERT based classification system for detecting rumours on Twitter," *CoRR*, vol. abs/2109.02975, 2021. [Online]. Available: https://arxiv.org/abs/2109.02975

[66] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer, "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. [Online]. Available: https://aclanthology.org/N18-1202

[67] Y. Wu, M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, Łukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean, "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144

[68] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019. [Online]. Available: https://openreview.net/forum?id=Bkg6RiCqY7

[69] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[70] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, pp. 687–719, 2009. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218001409007326

[71] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, pp. 1–30, 2018. [Online]. Available: https://link.springer.com/article/10.1186/s40537-018-0151-6

[72] B. Vidgen, A. Harris, D. Nguyen, R. Tromble, S. Hale, and H. Margetts, "Challenges and frontiers in abusive content detection," in *Proceedings of the Third Workshop on Abusive Language Online*. Florence, Italy: Association for Computational Linguistics, Aug. 2019, pp. 80–93. [Online]. Available: https://aclanthology.org/W19-3509

[73] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[74] Z. Yang, Z. Dai, Y. Yang, J. G. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," in *NeurIPS*, 2019.

[75] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," *ArXiv*, vol. abs/2003.10555, 2020.

[76] "Bipolar UK Mood Scale." [Online]. Available: https://www.bipolaruk.org/faqs/mood-scale