Computers in Human Behavior 139 (2023) 107503

Contents lists available at ScienceDirect

Computers in Human Behavior

journal homepage: www.elsevier.com/locate/comphumbeh

Full length article

DEPTWEET: A typology for social media texts to detect depression severities

Mohsinul Kabir [a,*], Tasnim Ahmed [a], Md. Bakhtiar Hasan [a], Md Tahmid Rahman Laskar [b,c], Tarun Kumar Joarder [d], Hasan Mahmud [a], Kamrul Hasan [a]

[a] Department of Computer Science and Engineering, Islamic University of Technology, Board Bazar, Gazipur, 1704, Dhaka, Bangladesh
[b] Department of Computer Science, York University, 44700 Keele St, Toronto, M3J 1P3, Ontario, Canada
[c] Dialpad Canada Inc., Melville Street, Vancouver, 1100, British Columbia, Canada
[d] Department of Psychology, University of Rajshahi, Matihar, Rajshahi, 6205, Bangladesh

ARTICLE INFO

Keywords:
Social media
Mental health
Depression severity
Dataset

ABSTRACT

Mental health research through data-driven methods has been hindered by a lack of standard typology and scarcity of adequate data. In this study, we leverage the clinical articulation of depression 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 tweets. Along with the typology, we present a new dataset of 40191 tweets labeled by expert annotators. Each tweet is labeled as 'non-depressed' or 'depressed'. Moreover, three severity levels are considered for 'depressed' tweets: (1) mild, (2) moderate, and (3) severe. An associated confidence score is provided with each label to validate the quality of annotation. We examine the quality of the dataset via representing summary statistics while setting strong baseline results using attention-based models like BERT and DistilBERT. Finally, we extensively address the limitations of the study to provide directions for further research.

1. Introduction

Analyzing the presence of mood and psychological disorders through behavioral and linguistic cues from social media data remains a critical area of interdisciplinary research. In addition to these disorders, the last decade has seen exponentially increasing attempts to assess related symptomatology such as depressive disorders, self-harm, and severity of mental illness using non-clinical data (Bucci et al., 2019). Social media platforms and other online discussion forums have been particularly appealing to the research community for various research purposes (e.g., population-level mental health monitoring (Conway & O'Connor, 2016), personal traits detection (Marouf et al., 2020), cyberbullying spotting (Bozyiğit et al., 2021), etc.) 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 on these platforms (Ofek et al., 2015). Recognizing the early symptoms of depressive disorder through a person's language use can prevent many disastrous outcomes like self-harm, suicide, etc., and even help deploy effective treatment in proper time. Moreover, the 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 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 (Singh et al., 2020). 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 (Low et al., 2020).

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 (Ernala et al., 2019), 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 diagnoses or by using vectorization-driven methods to cluster activity patterns of users. However, deliberately relying on self-labeled data or unsupervised clustering leads to oversimplification and lacks clinical efficacy (Ernala et al., 2019). Practical exertion of mental health research includes identifying risky behaviors and providing timely interventions such as suicide prevention efforts adopted by

* Corresponding author.
E-mail address: mohsinulkabir@iut-dhaka.edu (M. Kabir).

https://doi.org/10.1016/j.chb.2022.107503
Received 15 February 2022; Received in revised form 8 August 2022; Accepted 25 September 2022
Available online 28 September 2022

Facebook (Vincent, 2017). 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 (Tolentino & Schmidt, 2018).

This study aims to contribute to this domain through (a) establishing a typology for social media contents (i.e., tweet text) built upon a psychological theory for detecting the severity of the mental condition of depressed individuals, (b) constructing a dataset named DEPTWEET[1] containing 40,191 tweets with corresponding crowdsourced labels and confidence scores. 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.

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) (Arbanas, 2015), and it was carried out under the supervision of two expert clinical psychologists. The DEPTWEET dataset contributes further high-quality data on attributes like none, mild, moderate, or severe depression, adding to existing datasets on these and related attributes (Ahmed, Mukhiya et al., 2021; Mukhiya et al., 2020), and provides the first dataset of this scale on depression severities to the best of our knowledge. The approach utilized in this study can be adapted to generate high-quality mental health data from various platforms in future investigations. Moreover, given that the data was collected in the latter half of 2021, topic modeling on this dataset can provide useful insight into the impact of the COVID-19 pandemic on individuals' mental health.

The remaining sections of the paper are structured as follows: Sections 2 and 3 outline the motivation and background of the DEPTWEET dataset. The data collection, quality control mechanisms, and summary statistics of the data are described in Section 4. The baseline classification model for this dataset and evaluation metrics are presented in Section 5. Section 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. Finally, Section 7 draws a conclusion to the current study and discusses future directions.

2. Related work

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 (Fuchs, 2015). Coppersmith et al. (2014) made a prominent contribution in this 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. They mixed control samples from the general population (people who are not depressed) with the tweets of the self-reported diagnosed group. Additionally, they conducted an Linguistic Inquiry Word Count (LIWC) analysis to measure deviations of each disorder group from the control group. 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 (Coppersmith et al., 2015) 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 depression vs. control, identify PTSD vs. control, and identify 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 (Amir et al., 2017; Coppersmith et al., 2016; Pedersen, 2015). In particular, Resnik et al. (2015) conducted several topic modeling such as supervised Latent Dirichlet Allocation (LDA), supervised anchor topic modeling, etc. to differentiate the language usage of depressed and non-depressed individuals using the previously mentioned datasets.

Following a similar approach, Chen et al. (2018) 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 media, Tian et al. (2016) explored sleep complaints on Sina Weibo 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. For instance, Jamil et al. (2017) 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 25,362 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 depression 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).

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. (2017) 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 profiles. 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. (2020) proposed an open set of depression word embeddings that extracts depression symptoms from patient-authored text data based on PHQ-9 to deliver a personalized intervention to people with symptoms of depression. Yadav et al. (2020) utilized the nine symptom classes of the PHQ-9 questionnaire to manually annotate the tweets collected from 205 users who self-reported to have been diagnosed with depression. Their proposed framework took into consideration the figurative language (metaphor, sarcasm, etc.) wired in the communication of depressive users on Twitter. Ahmed, Mukhiya et al. (2021) extracted depression symptoms in patient-authored text in a similar fashion using PHQ-9 questionnaire. They used an attention-based in-depth entropy active learning to annotate the unlabeled texts automatically. Their methodology increased the trainable instances of mental health data using a semantic clustering mechanism 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. (2018) developed an approach

[1] The DEPTWEET dataset is available at https://github.com/mohsinulkabir14/DEPTWEET

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,[2] SNOMED-CT,[3] DataMed[4]). 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. (2016) 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.

Very few studies have investigated predicting the severity of depression based on users' language usage on web platforms. De Choudhury et al. (2013) proposed a metric named Social Media Depression Index (SMDI) using a probabilistic model to help characterize the levels of depression at the population level. This probabilistic model is a Support Vector Machine (SVM) classifier that 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 Center for Epidemiologic Studies Depression Scale (CES-D) screening test (Radloff, 1977). Schwartz et al. (2014) attempted to predict and characterize the severity of depression based on people's Facebook language use. They gathered survey responses and Facebook posts from 28,749 Facebook users and trained a classification model to predict depression symptoms using n-grams, linguistic behavior, and 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 partnership with expert psychologists and domain experts.

While previous research has made significant progress towards the development of automatic depression assessment tools based on social media, some limitations have been identified through critical evaluation. Most previous works have relied on self-reported depressed user profiles when it comes to data extraction. While this is an inexpensive way to gather a massive scale of data, it does not guarantee enough samples with depressive symptoms without manual intervention. Also, this approach lacks enough clinical validation to extract depression symptoms. Studies that leveraged clinical assessment tools to extract data, such as the PHQ-9 or DSM-5, lacked supervision from domain experts and mostly annotated their data in an automated manner, such as using unsupervised topic modeling or clustering techniques. Moreover, only a few studies have investigated how to collect data on different depression severities with sufficient clinical efficacy. The existing datasets only concentrate on binary detection of whether a particular tweet manifests depression or not, the severity level of which is mostly ignored. This might lead to models, competent enough in detecting subtle cues of depression, turning a blind eye towards them. A dataset containing sufficient samples to train large models with strong ground truth labels depicting the severity of depression can go a long way to alleviate these issues.

3. Measuring severity of depression

In the current 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 be 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 (Yazdavar et al., 2017). Incorporating this idea, the Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001) 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 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 Symptom Elicitation Process (CSEP) (World Health Organization, 1993). In the current study, this was further extended using the mood scale provided by BipolarUK[5] 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:

3.1. Non-depressed tweets

A tweet can be labeled 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 that person's mental state, expresses casual tiredness or sadness (For example, sadness due to the defeat of their favorite sports team), or expresses temporary hopelessness. It can also convey any other emotion except for depression.

3.2. 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.

3.3. 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, and excessive worrying.

3.4. 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.

4. The DEPTWEET dataset

In this section, the complete methodology of constructing the DEPTWEET dataset and the summary statistics of the data are discussed extensively. TWINT[6] was used to collect tweets from Twitter for this study. The collected tweets went through a preliminary screening process before being distributed to the annotators. The annotation job was carefully observed and regulated in order to maintain the quality of the data. An overview of the data collection and annotation procedure is displayed in Fig. 1. Below, we first present how we collected the data. Then, the data annotation process is demonstrated in detail. Finally, we discuss the properties of the dataset.

[2] https://bioportal.bioontology.org/ontologies/ICD10
[3] http://bioportal.bioontology.org/ontologies/SNOMEDCT
[4] https://datamed.org/
[5] https://www.bipolaruk.org/faqs/mood-scale
[6] https://github.com/twintproject/twint

Table 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)	Am I depressed or am I just bored? Apathy and irony, postmodern anxiety	Disinterest	Involved, occupied, pessimism, reversion, absorbed, lifelessness, bored, enthusiasm, engrossed, worried, apathy.
Feeling down (S2)	High functioning depression, I cannot fester in my misery but i'm fuckin miserable	Hopeless, depressed	Dejected, dismayed, dispirited, demoralized, grimmed, misery, grim, downhearted, low-spirited, bleak, desperate, lost, frustrated.
Sleep disorder (S3)	Forcing myself up now so I am not awake when the power goes off much later, lol	Awake, sleep	Nap, restless, awake, whole night, bedtime.
Lack of energy (S4)	I am so exhausted and I still have work 9-5 and then red rocks day three	Tired, energy	Weary, fatigue, fag, fag out, overtire, overfatigued, burned-out, burnt-out, exhausted, dog-tired, washed-out, drained, whacked.
Eating disorder (S5)	Another saturday night where i'm too depressed to sleep after overeating....i am extremely bored of this life	Appetite, overeating	Aversion, distaste, loathing, malformed, bulimic, puffy, starve, fat
Low self-esteem (S6)	I got on the scale today and I am disgusted. Like utterly disgusted. Depression really beat my ass and had me slacking	Loser, failure	Loser, relapse, downfall, ruined, flop, dead-duck, disappointment, achiever, misfire, underdog, falling-apart, disgusted
Concentration problems (S7)	Whenever it gets close to my bday I always go through some type of cleansing/depression.. Scattered focus...	Concentrate, focus	Immersed, decentralize, deconcentrate, scattered, dispersed, unsettled, focus
Hyper/Lower activity (S8)	I spend hours of my day staring at screens, immobile. Why am I depressed???	Moving, immobile, restless	Discontent, ungratified, unsatisfied, stand-still, refrained, immobile
Suicidal thoughts (S9)	I know that I cannot undo The self-destruction, the damage I have done	Dead, hurt, suicide	Trauma, harm, suffering, anguish, hemorrhage, penetrating-trauma, torment, agony, excruciate, damaged, gag, suffocate, self-destruction

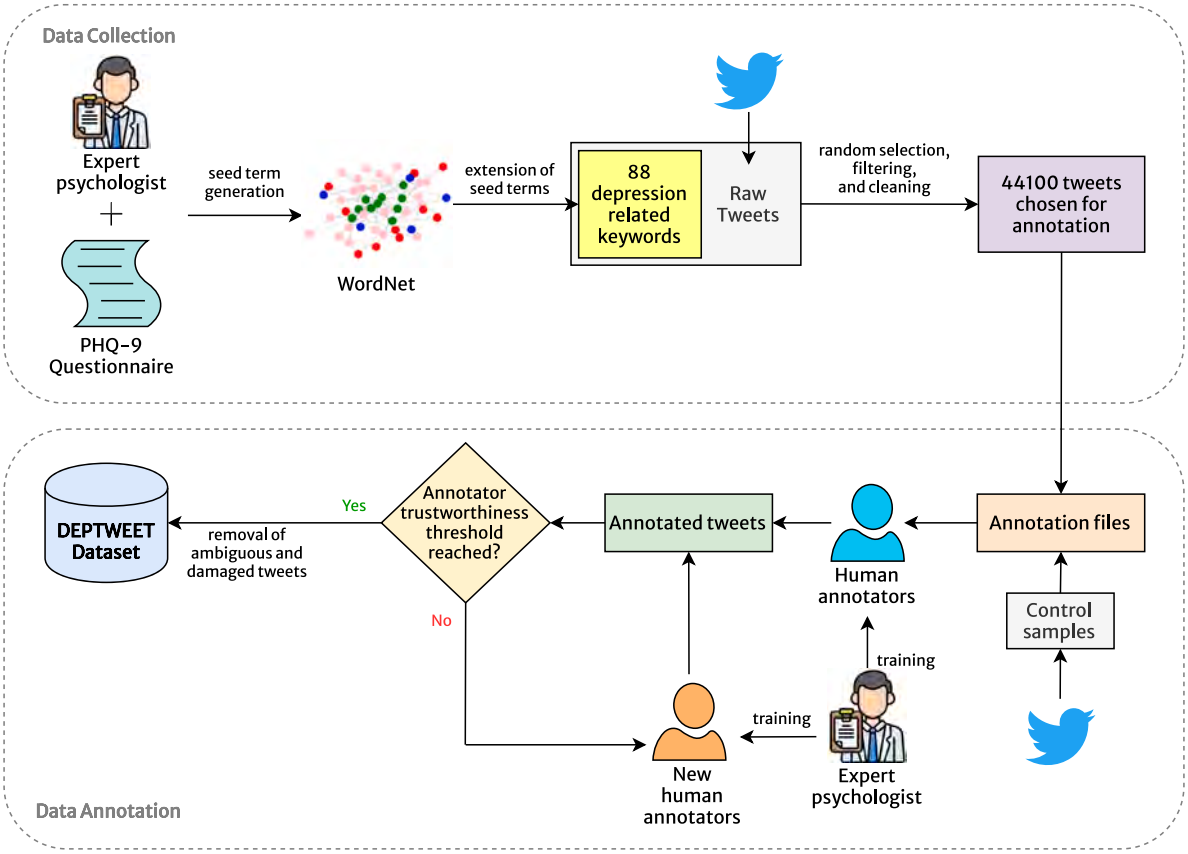


Fig. 1. Overview of the dataset creation process.

4.1. Data collection

Seed terms were generated from the keywords extracted from each of the symptoms of the PHQ-9 questionnaire by collaborating with two professional psychologists. This is a commonly used procedure employed in many previous studies (Ahmed, Mukhiya et al., 2021; Mukhiya et al., 2020; Yazdavar et al., 2017). After seed terms generation, they were then extended using WordNet (Miller, 1995). 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 (Mukhiya et al., 2020; Yazdavar et al., 2017) such as *Universal Sentence Encoding (USE)* (Cer et al., 2018), *Global vector representation (GloVe)* (Pennington et al., 2014), *Big Huge Thesaurus* (Watson, 2007), etc. In the evaluation shown by Mukhiya et al. (2020), WordNet performs significantly better in extracting symptoms from patient-authored text compared to other methods. For this reason, in the current study, the seed terms for each questionnaire of PHQ-9 were extended by WordNet, and the extended terms were handpicked afterwards 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 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 344,657 tweets were collected.

From the collected samples, tweets that were posted in English were only preserved for annotation. Tweets with less than eight words were discarded as they might not contain enough context. Any tweets containing mentions (@) or hashtags (#), as well as retweets, were also discarded since they could violate the privacy of the users mentioned. Finally, 44 100 tweets were randomly chosen from the remaining tweets for annotation.

4.2. Data annotation

Several data annotation techniques can be applied to determine the class label for the sample tweets. Since the number of classes is known beforehand, one intuitive approach can be creating vector representations of the tweets using Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2017), etc. and then using unsupervised clustering to find the optimal distribution of the samples into different clusters. However, such approaches lack human input who can understand the subtle nuances of tweets to identify different levels of severity, resulting in a poorly annotated dataset (Ernala et al., 2019). To ensure clinical accuracy, annotators, trained by expert psychologists, were employed to perform dataset annotation.

4.2.1. Annotator recruitment

The annotation job was done by recruiting participants who were 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 as one of the 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 procedure 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. Each tweet was annotated at least three times. The final label of a tweet was determined by majority voting of the labels provided by the three annotators. Tweets with different labels from all three annotators were discarded because of too many disagreements. Final labels of the dataset were established with a confidence score to reflect the disagreement of the annotator because of reasonable differences of opinion.

Table 2
Metadata about the datafiles created for annotation.

Type	Count
Number of tweets collected	344 657
Tweets chosen for annotation	44 100
Total datafiles created	30
Data samples in each datafile	1470
Control samples per datafile	30
Total tweets per datafile	1500

4.2.2. Annotation job refinement

Though it was ensured that annotators' disagreement reflected a genuine difference of opinion, a means of quality control was required to prevent annotators' inattention or misunderstanding of context. The quality control mechanism used by Price et al. (2020) was 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' was collated with the actual data sample, for which the correct labels were manually established. Annotators encountered one control sample per batch of fifty tweets without knowing which of the tweets was the control sample. The running accuracy of these control samples was defined as annotator's 'trustworthiness score (T)'. The threshold trustworthiness score for this study was set to be at least 90%. If an annotator dropped below this level, all of their annotations were discarded, and the annotator was removed from the annotator pool. Afterwards, another annotator from the pool was assigned to re-annotate those data samples.

A total of 900 control samples were added for quality control with the previously chosen $44,100$ data samples. To generate datafiles for the annotators, the actual dataset containing 44,100 samples was divided into 30 parts, each part containing $(44100/30) = 1470$ samples. For every 49 tweets in these 1470 samples, one unique control sample was added at a random position. The control samples were from the *non-depressed* category and were 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 were too reckless while annotating. The tweet ID of the control samples were also tracked. Following this method, 30 datafiles were created containing 1500 tweets (1470 data samples + 30 control samples) each. Each datafile consisted of two columns: one having tweet texts, and another empty column for annotator label. All the other data columns were kept hidden from the annotators. The metadata related to the datafile creation procedure is summarized in Table 2.

To annotate these datafiles, ninety annotators were divided into three groups, each with thirty annotators. Each datafile was 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 was finished, all the datafiles were merged and the control samples were removed from the dataset.

Table 3
Percentage of data samples for each class.

Class	Proportion
Non-depressed	80.62%
Mild	13.04%
Moderate	4.5%
Severe	1.84%

4.3. Dataset properties & analysis

Of the 44,100 tweets considered for annotation, 1,399 data samples were 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 were discarded due to annotator disagreement, as they received three different labels from three different annotators. The final dataset comprises a total of 40,191 tweets along with their *tweet_id*, *replies_count*, *retweets_count*, *likes_count*, *target*, *label* and *confidence_score*. The label for each tweet was determined based on the aggregation of the labels provided by different annotators. If at least two of the three annotators agreed on the label of a tweet, the matched annotation was accepted as the final label. Tweets that had three different annotations from three annotators, were discarded and saved in a separate datafile. Further annotation is required to achieve a class label for these samples and were left because of budget and time constraint. The corresponding confidence score for each label was 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:

$$\text{Confidence Score}(C) = \frac{\sum T_i}{T} \tag{1}$$

where T_i denotes trustworthiness of ith annotator whose annotations matched and T denotes sum of the trustworthiness score of all the annotators who annotated 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.

Manual analysis was performed in two stages of the study to gain insights into the dataset: (i) while randomly choosing data samples for annotation, and (ii) during the initial iterations of the annotation job. The proportion of classes shown in Table 3 indicates that the *non-depressed* samples outnumber the other classes by a wide margin. Though all the data samples were 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. 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 (Vermeulen et al., 2018). To mitigate this problem, previous studies depended on self-labeled data for collating large and balanced datasets on different mental disorders (Kim et al., 2020; Low et al., 2020). 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.

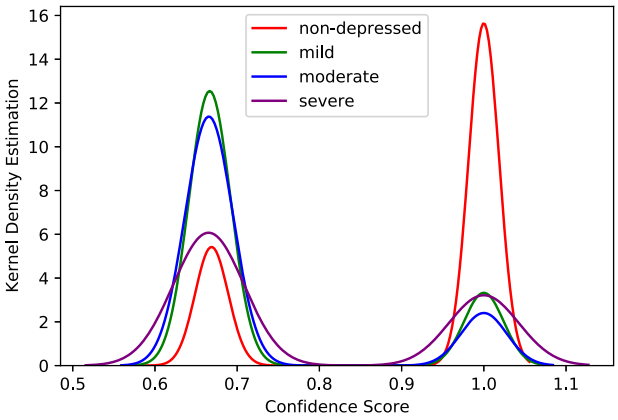


Fig. 2. Kernel density estimation of confidence scores for each class.

Table 4
Fleiss' Kappa per class.

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

Despite the measures undertaken to ensure the quality of the dataset, the method of annotation warrants a certain level of noise. This results in different yet rational interpretations of the same tweet. The kernel density estimation of the confidence scores portrayed in Fig. 2 indicates that there was reasonable agreement among the annotators on deciding the class label of the *non-depressed* and *severe* classes. While these two classes lie 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 Fig. 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 (Fleiss et al., 2013) 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 (Gwet, 2014; Leard Statistics, 2019). 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 (Salminen et al., 2018). In spite of that, Fleiss' Kappa was 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) (Leard Statistics, 2019). 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.

As reported in Table 4, 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.

5. Experimental design

The choice of baseline models and evaluation metrics for this study are discussed in this section.

5.1. Baseline models selection

One of the main challenges in language-related tasks comes from the use of homonyms and synonyms as well as different kinds of ambiguity in sentences such as lexical, semantic, and syntactic ambiguity. Another challenging task for a model is to extract context from various domain-specific languages. Empirical studies have shown that rule-based methods and traditional machine learning-based methods fail to overcome these complexities by understanding the inherent meaning of the sentences (González-Carvajal & Garrido-Merchán, 2020; Kansara & Sawant, 2020). Multilingualism is another challenge with classic machine learning techniques (González-Carvajal & Garrido-Merchán, 2020). Rules for a specific language can be formed, but the alphabet and the sentence structure can differ from one language to other, requiring the development of new rules. Most of these aforementioned shortcomings are alleviated by transformer (Vaswani et al., 2017) based architectures that use an attention mechanism to capture bi-directional context and are also capable of handling larger datasets than traditional machine learning-based architectures. Considering these issues, a series of baseline models were chosen to evaluate the proposed dataset, namely Support Vector Machine (SVM) (Cortes & Vapnik, 1995), Bidirectional LSTM (BiLSTM) (Schuster & Paliwal, 1997), BERT (Devlin et al., 2019), and DistilBERT (Sanh et al., 2020). Bidirectional LSTM (BiLSTM) was selected as it is a widely used recurrent neural network based on deep learning architecture, Support Vector Machine (SVM) as a classical machine learning model, and BERT and DistilBERT as two transformer-based models. While the word embedding of SVM and BiLSTM models rely on choice, both BERT and DistilBERT are pre-trained using a large amount of data from English Wikipedia[7] and Toronto Book Corpus (Zhu et al., 2015). The pre-training is generic enough to be fine-tuned for downstream tasks such as sequence classification, named entity recognition, natural language inference, etc.

Reasons for choosing these models can be summarized as follows:

- A diverse set of classifiers are chosen as baseline models to evaluate the validity of the dataset. SVM has already been used by De Choudhury et al. (2013) to create a probabilistic model to predict the severity of depression from tweets. BiLSTM is a sequence processing model that calculates the input sequence from the opposite direction to a forward hidden sequence and a backward hidden sequence. Due to its effective contextual understanding ability, BiLSTM has been frequently used as a baseline classifier (Moon et al., 2020).
- Previous studies have shown that fine-tuning BERT-based models (Devlin et al., 2019; Lewis et al., 2019; Liu et al., 2019; Sanh et al., 2020) yield impressive performance in various downstream tasks such as text categorization (Rogers et al., 2020; Wu et al., 2020; Yamada et al., 2020) question-answering (Garg et al., 2020; Laskar, Huang et al., 2020), summarization (Laskar et al., 2020a, 2020b, 2021; Liu & Lapata, 2019), sentiment analysis of social media posts (Ahmed, Kabir et al., 2021; Moshkin et al., 2020), etc., since these models are pre-trained on a large amount of unlabeled data via leveraging self-supervised learning.
- 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). According to research on pre-trained models such as MegatronLM (Shoeybi et al., 2019), bigger models with billions of parameters usually result in superior performance on downstream tasks. However, the overall performance boost comes at the price of higher computational power and memory needs for both training and inference, rendering them unsuitable for use on the edge devices, such as smartphones. To address this issue, Sanh et al. (2020) proposed DistilBERT, which has a similar architecture as BERT and is pre-trained on the same corpus. By removing token-type embeddings and the pooler from the BERT implementation, DistilBERT reduces the number of layers by a factor of two, because hidden size dimensions have less of an influence on computation efficiency than the number of levels. DistilBERT is pre-trained through knowledge distillation via the supervision of a larger model incorporating triple loss functions (Distillation Loss - L_{ce}, Masked Language Modeling Loss - L_{mlm}, and Cosine Embedding Loss - L_{cos}). DistillBERT maintains 97% of BERT performance on downstream tasks with 40% fewer parameters. Additionally, it reduces the inference time of BERT in downstream tasks by around 60%. The fundamental reason for this is a compression method called knowledge distillation, which enables a compact model to replicate the behavior of larger models as well as the components of triple loss.

Both BERT and DistilBERT rely on Auto Encoding (AE) language modeling during pre-training since the aim is to understand natural language representations. Although the general transformer architecture proposed by Vaswani et al. (2017) 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.

It is to be noted that, all of the baseline models that were chosen are data-driven approaches. As a result, these models are unable to extract semantic information from a context that is not explicitly in the data, unlike humans who can use their pre-existing knowledge to judge new contexts that they never encountered before (Cocarascu & Toni, 2018; d'Avila Garcez et al., 2019). One solution to this problem could be the use of symbolic approaches. Unfortunately, these approaches fall short due to scalability. Recent approaches combine symbolic and data-driven approaches to solve this problem (Cocarascu & Toni, 2018; Faghihi et al., 2021; Schockaert & Gutiérrez-Basulto, 2022). However, we limit ourselves to data-driven approaches to keep the baseline models simple.

5.2. Classifier configuration

The training procedure of the baseline classifiers is demonstrated below, followed by the training parameters of the experiment.

5.2.1. Support Vector Machine (SVM)

SVM tries to draw a hyperplane that best separates multi-dimensional data points in their potential classes and is ideal for binary classification (Cortes & Vapnik, 1995). For multi-class classification, a 'one-versus-one' approach with a Radial Basis Function (RBF) kernel was implemented. The values of the two crucial parameters for RBF kernel, $C = 0.5$ and $gamma = 0.5$, were chosen based on several iterations of experiments. The entire dataset was split into 80%–20% partitions for training and testing the model. Several text pre-processing techniques, such as stopwords removal, bad symbols removal, text lower-casing, etc., were applied to both the training and testing data.

[7] www.wikipedia.org

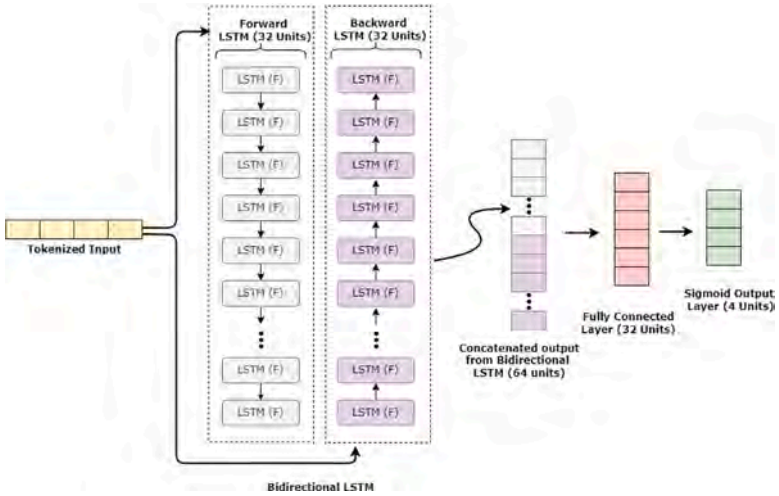


Fig. 3. Architecture of BiLSTM network.

5.2.2. Bidirectional LSTM (BiLSTM)

BiLSTM can preserve the sequence information in both directions, backward (right to left) or forward (left to right). To train the model, a bidirectional layer of 64 units was added after the word-embedding layer generated from the training data. The overall architecture of the BiLSTM network is illustrated in Fig. 3. Similar text pre-processing techniques like SVM were deployed for BiLSTM as well. During the training phase, the hyperparameters for this experiment were fine-tuned using cross-validation, adopting 10% of the data from the training samples as the validation set.

5.2.3. Fine-tuning BERT & DistilBERT

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.

Input Representation

Before being fed into the pre-trained models for embedding, each tweet text was 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 the WordPiece tokenizer (Wu et al., 2016), 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 the '[UNK]' token. Subsequently, the final input vector for the models was 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 were placed into the train set, and the rest of the tweets were equally distributed among the validation and test sets. Base-uncased[8] versions of the pre-trained models were implemented for fine-tuning with a total of 768 hidden output states. Categorical Cross-Entropy loss function with AdamW optimizer (Loshchilov & Hutter, 2019) was used that utilizes a fixed weight decay unlike common implementations of Adam optimizer (Kingma & Ba, 2015). Considering that the learning rate was set to 3×10^{-5} and 20% of the steps were 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 denote the total number of times when the model weights get updated during the fine-tuning phase.

Both of these models were 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. Fig. 4 depicts the process of predicting the severity of depression using the fine-tuned classifiers from a sample tweet.

5.3. Evaluation metrics

Evaluation metrics play a crucial role in quantifying the performance of a predictive classifier (Sun et al., 2009). Since the choice of metrics depends on the characteristic of the dataset, this can often lead to a 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 the small classes at all since the majority predictions are from the densely populated classes (Leevy et al., 2018).

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

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

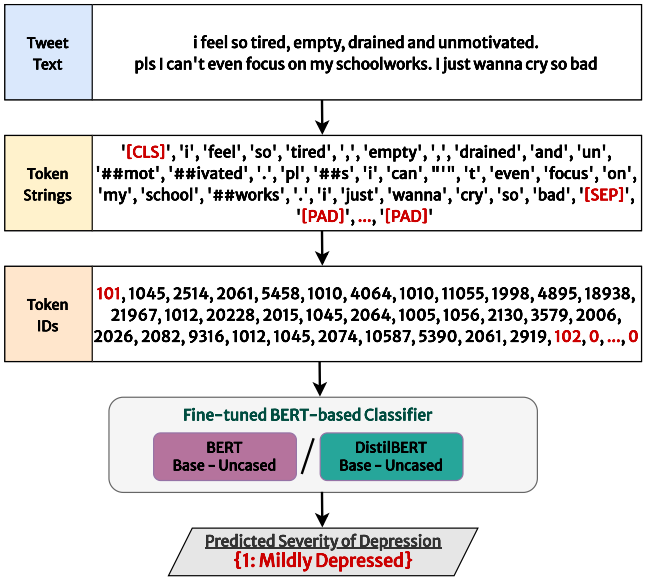


Fig. 4. Severity of depression prediction from a sample tweet.

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.

To tackle this issue, the Receiver Operating Characteristic (ROC) curve and the area under the ROC curve (AUC-ROC) were 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.

In the ROC graph, each probability threshold is represented by a point, linked to form a curve. A model 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 upper left corner of the plot.

6. Results and discussions

The performance of the baseline models on our dataset will be discussed in this section, followed by the potential unintended bias of this study.

6.1. Classification performance & analysis

According to the results shown in Table 5, it can be observed that SVM and BiLSTM were outperformed by the two transformer-based models by a large margin. Transformer-based models that are used in this study can learn each word's context from the words that appear before and after it and are also pre-trained on a large corpus. 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 (Radford et al., 2018) where each token is capable of managing only the preceding tokens in the transformer's self-attention layers. As BiLSTM can also learn contexts of words in both directions, it seems to achieve decent performance in some classes as well.

It can also be observed that DistilBERT outperformed BERT in all classes. Since DistilBERT is pre-trained under the supervision of its parent model, BERT through knowledge distillation, it is able to preserve 95% performance of the base uncased BERT (Sanh et al., 2020) which is divergent from the experimental results shown in this study. The experiments were conducted in a computationally limited environment with comparatively smaller batch size and fine-tuned only for 10 epochs. It is likely that BERT will outperform DistilBERT if the models are fine-tuned for a higher number of iterations with further hyper-parameter tuning.

Table 5
Performance comparison of baseline models.

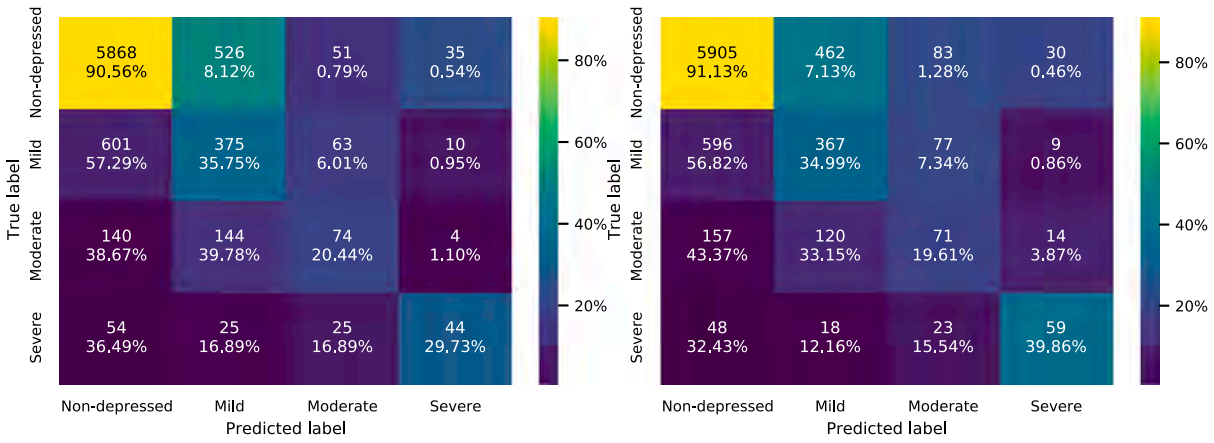
Model	Class name	ROC AUC Score
SVM	Non-depressed	0.514816
	Mild	0.511343
	Moderate	0.512785
	Severe	0.547684
BiLSTM	Non-depressed	0.692522
	Mild	0.565517
	Moderate	0.795351
	Severe	0.755356
BERT	Non-depressed	0.763699
	Mild	0.740019
	Moderate	0.748115
	Severe	0.826488
DistilBERT	Non-depressed	0.788841
	Mild	0.747211
	Moderate	0.787959
	Severe	0.866003

As seen from Table 3, 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 a relatively smaller number of samples for other classes as well. From the confusion matrices in Fig. 5, 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, in a 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 leaves room for further improvement through error analysis.

For the proposed dataset, ROC curves using the test predictions from the baseline classifiers is presented in Fig. 6. These plots are summarized by calculating the area under the ROC curve (AUC-ROC) in Table 5. The better performance of DistilBERT and BERT are also distinguishable from the class-wise AUC-ROC curves in Fig. 6.

6.2. Potential unintended bias

Fig. 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



(a) BERT (b) DistilBERT

Fig. 5. Confusion matrix obtained by evaluating test set using fine-tuned classifiers.

Table 6
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 are 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 do not like to brag. BUT, I do not think there is a soul on this earth that does self destruction like I do.	*severe*	*non-depressed*

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 was selected.

Another challenge that appeared in a similar context for the annotators was to avoid their individual biases 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 Vidgen et al. (2019), especially for complex multi-class tasks of this kind.

The data extension tool used for this study is Wordnet, which was initially released in the mid-1980s. Though it has been updated over time, due to the continuous evolution of language, people today often use a vocabulary on social media that can differ significantly from the one that Wordnet represents. Moreover, some of the semantic relations enlisted in Wordnet are more suited to concrete notions than to abstract ones (Rudnicka et al., 2018). For example, it is easy to create hyponyms/hypernym relationships to illustrate that a "Pinaceae" is a type of "tree", a "tree" is a type of "plant", and a "plant" is a type of "organism", but it is difficult to classify emotions like "anxiety" or "delight" into equally deep and well-defined associations. Finding appropriate seed terms that best capture the depressive emotion of people on social media might be substantially hindered by these limitations.

7. Conclusions and future work

This work introduced a new typology for diagnosing depression severities from social media texts, as well as a unique dataset of labeled tweets with a confidence score for each label. The dataset was constructed based on strong ground truths and clinical validation, and it 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. The experimental results indicated that existing state-of-the-art models often fail to understand the contextual undertone of the data samples. Developing a model that is capable of comprehending the subdued relationship and differences among depression severities can result in an even better understanding of human cognition. Moreover, analysis of the classification performance indicates that there is no distinct division of keywords among different depression severities. The same keyword might be used differently to express different emotions, rather it is more important to understand the context of the tweet to diagnose the severity of depression. Broader implications of this research may include personalizing and directing preventative and awareness messages by health professionals to the users in need.

The seed terms for each symptom of PHQ-9 in this study were extended by Wordnet (Miller, 1995). Considering the fast-evolving nature of languages in social media, future studies can utilize more recent lexical databases with a larger semantic network to extend the seed terms. For example, CMU pronouncing dictionary [9], MRC Psycholinguistic Database (Coltheart, 1981), and The Verb Semantics Ontology Project (Fukushima, 1984) are other available lexical databases that can be used in seed term extension. Additionally, authors can also develop their own domain-specific lexical database by vector

[9] http://www.speech.cs.cmu.edu/cgi-bin/cmudict

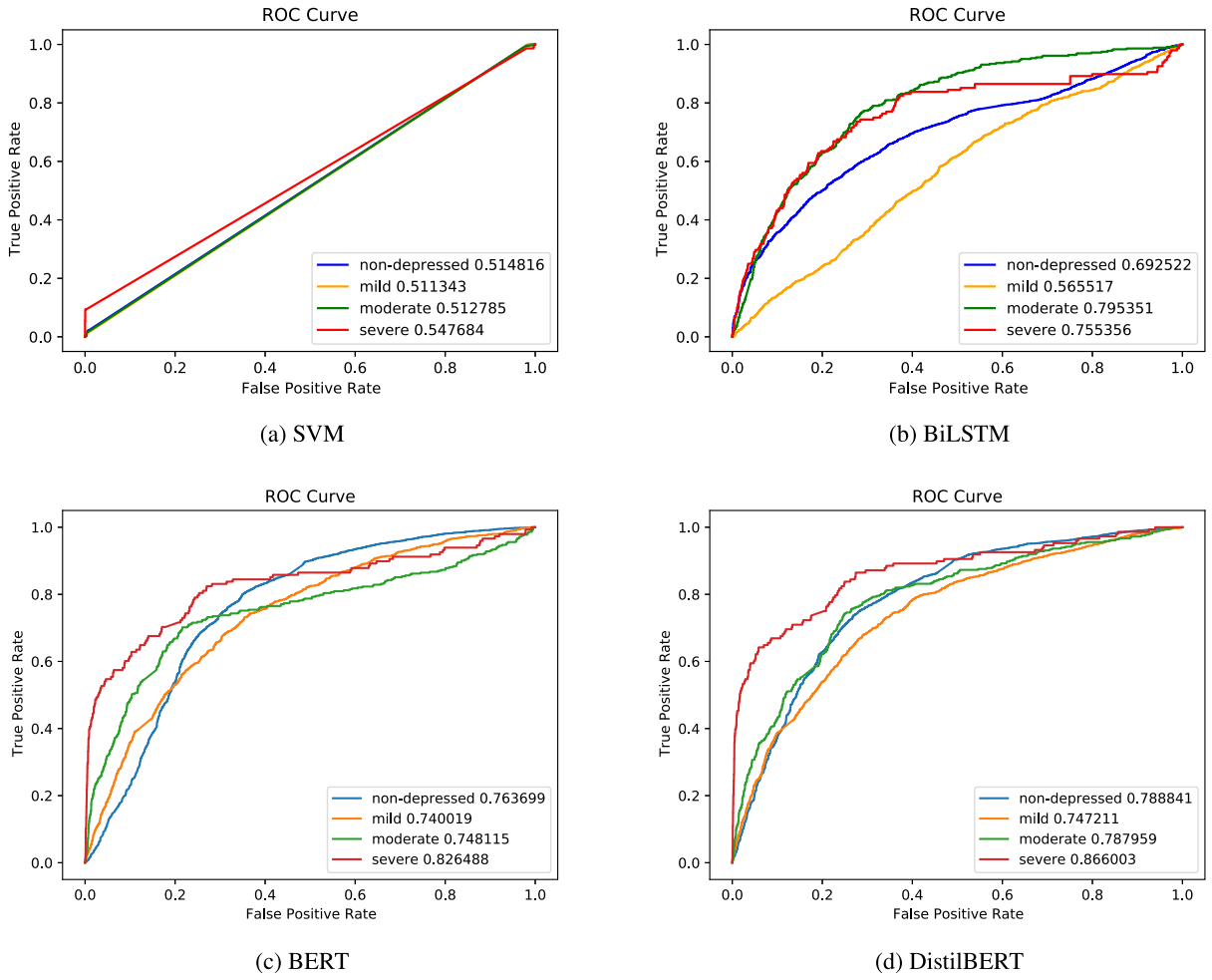


Fig. 6. Class-wise AUC-ROC curves.

proximity using a domain-specific corpus as a starting point. These approaches can build a keyword list that better extracts depression-related symptoms posted on social media nowadays. 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, a 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, 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.

Funding

This research is supported by Islamic University of Technology Research Seed, Bangladesh Grant (IUT-RSG/2021/OL/07/013).

CRediT authorship contribution statement

Mohsinul Kabir: Conceptualization, Methodology, Software, Formal analysis, Investigation, Data curation, Writing – original draft. **Tasnim Ahmed:** Methodology, Software, Formal analysis, Investigation, Writing – original draft. **Md. Bakhtiar Hasan:** Methodology, Formal analysis, Investigation, Writing – original draft. **Md Tahmid Rahman Laskar:** Resources, Writing – review & editing, Investigation. **Tarun Kumar Joarder:** Supervision, Methodology, Validation. **Hasan Mahmud:** Writing – review & editing, Resources, Supervision, Project administration, Funding acquisition. **Kamrul Hasan:** Writing – review & editing, Resources, Supervision, Project administration, Funding acquisition.

Data availability

The data associated with the research has been shared in a public git repository.

References

Ahmed, T., Kabir, M., Ivan, S., Mahmud, H., & Hasan, K. (2021). Am I being bullied on social media? An ensemble approach to categorize cyberbullying. In *2021 IEEE international conference on big data (Big data)* (pp. 2442–2453). http://dx.doi.org/10.1109/BigData52589.2021.9671594, URL: https://ieeexplore.ieee.org/abstract/document/9671594.

Ahmed, U., Mukhiya, S. K., Srivastava, G., Lamo, Y., & Lin, J. C.-W. (2021). Attention-based deep entropy active learning using lexical algorithm for mental health treatment. *Frontiers in Psychology*, *12*, http://dx.doi.org/10.3389/fpsyg.2021.642347, URL: https://www.frontiersin.org/article/10.3389/fpsyg.2021.642347.

Amir, S., Coppersmith, G., Carvalho, P., Silva, M. J., & Wallace, B. C. (2017). Quantifying mental health from social media with neural user embeddings. In F. Doshi-Velez, J. Fackler, D. Kale, R. Ranganath, B. Wallace, & J. Wiens (Eds.), *Proceedings of machine learning research*: *vol. 68*, *Proceedings of the 2nd machine learning for healthcare conference* (pp. 306–321). PMLR, URL: https://proceedings.mlr.press/v68/amir17a.html.

Arbanas, G. (2015). Diagnostic and statistical manual of mental disorders (DSM-5). *Alcoholism and Psychiatry Research*, *51*, 61–64, URL: https://www.amberton.edu/media/Syllabi/Spring%202022/Graduate/CSL6798_E1.pdf.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146, URL: https://transacl.org/ojs/index.php/tacl/article/view/999.

Bozyiğit, A., Utku, S., & Nasibov, E. (2021). Cyberbullying detection: Utilizing social media features. *Expert Systems with Applications*, *179*, Article 115001. http://dx.doi.org/10.1016/j.eswa.2021.115001.

Bucci, S., Schwannauer, M., & Berry, N. (2019). The digital revolution and its impact on mental health care. *Psychology and Psychotherapy: Theory, Research and Practice*, *92*(2), 277–297. http://dx.doi.org/10.1111/papt.12222, URL: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/papt.12222.

Cavazos-Rehg, P. A., Krauss, M. J., Sowles, S., Connolly, S., Rosas, C., Bharadwaj, M., & Bierut, L. J. (2016). A content analysis of depression-related tweets. *Computers in Human Behavior*, *54*, 351–357. http://dx.doi.org/10.1016/j.chb.2015.08.023, URL: https://www.sciencedirect.com/science/article/pii/S0747563215300996.

Cer, D., Yang, Y., Kong, S.-y., Hua, N., Limtiaco, N., St. John, R., Constant, N., Guajardo-Cespedes, M., Yuan, S., Tar, C., Strope, B., & Kurzweil, R. (2018). Universal sentence encoder for english. In *Proceedings of the 2018 conference on empirical methods in natural language processing: System demonstrations* (pp. 169–174). Brussels, Belgium: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/D18-2029, URL: https://aclanthology.org/D18-2029.

Chen, X., Sykora, M. D., Jackson, T. W., & Elayan, S. (2018). What about mood swings: Identifying depression on Twitter with temporal measures of emotions. In *Companion proceedings of the the web conference 2018* WWW '18, (pp. 1653–1660). Republic and Canton of Geneva, CHE: International World Wide Web Conferences Steering Committee, http://dx.doi.org/10.1145/3184558.3191624, URL: https://dl.acm.org/doi/abs/10.1145/3184558.3191624.

Cocarascu, O., & Toni, F. (2018). Combining deep learning and argumentative reasoning for the analysis of social media textual content using small data sets. *Computational Linguistics*, *44*(4), 833–858. http://dx.doi.org/10.1162/coli_a_00338, arXiv:https://direct.mit.edu/coli/article-pdf/44/4/833/1809934/coli_a_00338.pdf.

Coltheart, M. (1981). The MRC psycholinguistic database. *Quarterly Journal of Experimental Psychology*, *33*, 497–505.

Conway, M., & O'Connor, D. (2016). Social media, big data, and mental health: Current advances and ethical implications. *Current Opinion in Psychology*, *9*, 77–82.

Coppersmith, G., Dredze, M., & Harman, C. (2014). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51–60). Baltimore, Maryland, USA: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W14-3207, URL: https://aclanthology.org/W14-3207.

Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). Clpsych 2015 shared task: Depression and PTSD on Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 31–39). Denver, Colorado: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W15-1204, URL: https://aclanthology.org/W15-1204.

Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016). Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology* (pp. 106–117). San Diego, CA, USA: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W16-0311, URL: https://aclanthology.org/W16-0311.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*(3), 273–297. http://dx.doi.org/10.1023/A:1022627411411.

De Choudhury, M., Counts, S., & Horvitz, E. (2013). Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference* WebSci '13, (pp. 47–56). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/2464464.2464480, URL: https://dl.acm.org/doi/abs/10.1145/2464464.2464480.

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Human language technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)* (pp. 4171–4186). Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/n19-1423.

Ernala, S. K., Birnbaum, M. L., Candan, K. A., Rizvi, A. F., Sterling, W. A., Kane, J. M., & De Choudhury, M. (2019). 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* (pp. 1–16). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3290605.3300364, URL: https://dl.acm.org/doi/abs/10.1145/3290605.3300364.

Faghihi, H. R., Guo, Q., Uszok, A., Nafar, A., Raisi, E., & Kordjamshidi, P. (2021). Domiknows: A library for integration of symbolic domain knowledge in deep learning. CoRR abs/2108.12370. URL: https://arxiv.org/abs/2108.12370. arXiv:2108.12370.

Fleiss, J. L., Levin, B., & Paik, M. C. (2013). *Wiley series in probability and statistics*, *Statistical methods for rates and proportions* (3rd ed.). John Wiley & Sons, Inc., http://dx.doi.org/10.1002/0471445428, URL: https://onlinelibrary.wiley.com/doi/book/10.1002/0471445428.

Fuchs, C. (2015). *Culture and economy in the age of social media*. New York: Routledge, http://dx.doi.org/10.4324/9781315733517, URL: https://www.taylorfrancis.com/books/mono/10.4324/9781315733517/culture-economy-age-social-media-christian-fuchs.

Fukushima, N. (1984). Ontology in the verb semantics.

d'Avila Garcez, A. S., Gori, M., Lamb, L. C., Serafini, L., Spranger, M., & Tran, S. N. (2019). Neural-symbolic computing: An effective methodology for principled integration of machine learning and reasoning. *FLAP*, *6*(4), 611–632, URL: https://collegepublications.co.uk/ifcolog/?00033.

Garg, S., Vu, T., & Moschitti, A. (2020). TANDA: Transfer and adapt pre-trained transformer models for answer sentence selection. In *AAAI*.

Gaur, M., Kursuncu, U., Alambo, A., Sheth, A., Daniulaityte, R., Thirunarayan, K., & Pathak, J. (2018). "Let me tell you about your mental health!": Contextualized classification of reddit posts to DSM-5 for web-based intervention. In *Proceedings of the 27th ACM international conference on information and knowledge management* CIKM '18, (pp. 753–762). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3269206.3271732, URL: https://dl.acm.org/doi/abs/10.1145/3269206.3271732.

González-Carvajal, S., & Garrido-Merchán, E. C. (2020). Comparing BERT against traditional machine learning text classification. arXiv preprint arXiv:2005.13012.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Jamil, Z., Inkpen, D., Buddhitha, P., & White, K. (2017). Monitoring tweets for depression to detect at-risk users. In *Proceedings of the fourth workshop on computational linguistics and clinical psychology — from linguistic signal to clinical reality* (pp. 32–40). Vancouver, BC: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W17-3104, URL: https://aclanthology.org/W17-3104.

Kansara, D., & Sawant, V. (2020). Comparison of traditional machine learning and deep learning approaches for sentiment analysis. In *Advanced computing technologies and applications* (pp. 365–377). Springer.

Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific Reports*, *10*(1), 1–6. http://dx.doi.org/10.1038/s41598-020-68764-y, URL: https://www.nature.com/articles/s41598-020-68764-y.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In Y. Bengio, & Y. LeCun (Eds.), *3rd international conference on learning representations (ICLR)*. URL: http://arxiv.org/abs/1412.6980.

Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2001). The PHQ-9. *Journal of General Internal Medicine*, *16*(9), 606–613. http://dx.doi.org/10.1046/j.1525-1497.2001.016009606.x, URL: https://onlinelibrary.wiley.com/doi/abs/10.1046/j.1525-1497.2001.016009606.x.

Laskar, M. T. R., Hoque, E., & Huang, J. (2020a). Query focused abstractive summarization via incorporating query relevance and transfer learning with transformer models. In C. Goutte, & X. Zhu (Eds.), *Advances in artificial intelligence* (pp. 342–348). Cham: Springer International Publishing.

Laskar, M. T. R., Hoque, E., & Huang, J. X. (2020b). WSL-DS: Weakly supervised learning with distant supervision for query focused multi-document abstractive summarization. CoRR abs/2011.01421. URL: https://arxiv.org/abs/2011.01421. arXiv:2011.01421.

Laskar, M. T. R., Hoque, E., & Huang, J. X. (2021). Domain adaptation with pre-trained transformers for query focused abstractive text summarization. CoRR abs/2112.11670. URL: https://arxiv.org/abs/2112.11670. arXiv:2112.11670.

Laskar, M. T. R., Huang, X., & Hoque, E. (2020). Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task. In *LREC*.

Leard Statistics (2019). Fleiss' kappa in SPSS statistics. URL: https://statistics.laerd.com/spss-tutorials/fleiss-kappa-in-spss-statistics.php.

Leevy, J. L., Khoshgoftaar, T. M., Bauder, R. A., & Seliya, N. (2018). A survey on addressing high-class imbalance in big data. *Journal of Big Data*, *5*(1), 1–30. http://dx.doi.org/10.1186/s40537-018-0151-6, URL: https://link.springer.com/article/10.1186/s40537-018-0151-6.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2019). BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. CoRR abs/1910.13461. URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

Liu, Y., & Lapata, M. (2019). Text summarization with pretrained encoders. ArXiv abs/1908.08345.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. ArXiv abs/1907.11692.

Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, URL: https://openreview.net/forum?id=Bkg6RiCqY7.

Low, D. M., Rumker, L., Talkar, T., Torous, J., Cecchi, G., & Ghosh, S. S. (2020). 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*, *22*(10), Article e22635. http://dx.doi.org/10.2196/22635, URL: http://www.jmir.org/2020/10/e22635/.

Marouf, A. A., Hasan, M. K., & Mahmud, H. (2020). Comparative analysis of feature selection algorithms for computational personality prediction from social media. *IEEE Transactions on Computational Social Systems*, *7*, 587–599.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. In Y. Bengio, & Y. LeCun (Eds.), *1st international conference on learning representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop track proceedings*. URL: http://arxiv.org/abs/1301.3781.

Miller, G. A. (1995). WordNet: A lexical database for english. *Communications of the ACM*, *38*(11), 39–41. http://dx.doi.org/10.1145/219717.219748, URL: https://dl.acm.org/doi/abs/10.1145/219717.219748.

Moon, J., Cho, W. I., & Lee, J. (2020). Beep! Korean corpus of online news comments for toxic speech detection. In *SOCIALNLP*.

Moshkin, V., Konstantinov, A., & Yarushkina, N. (2020). Application of the BERT language model for sentiment analysis of social network posts. In S. O. Kuznetsov, A. I. Panov, & K. S. Yakovlev (Eds.), *Artificial intelligence* (pp. 274–283). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-59535-7_20, URL: https://rd.springer.com/chapter/10.1007/978-3-030-59535-7_20.

Mukhiya, S. K., Ahmed, U., Rabbi, F., Pun, K. I., & Lamo, Y. (2020). Adaptation of IDPT system based on patient-authored text data using NLP. In *2020 IEEE 33rd international symposium on computer-based medical systems (CBMS)* (pp. 226–232). http://dx.doi.org/10.1109/CBMS49503.2020.00050, URL: https://ieeexplore.ieee.org/abstract/document/9183294.

Ofek, N., Katz, G., Shapira, B., & Bar-Zev, Y. (2015). Sentiment analysis in transcribed utterances. In T. Cao, E.-P. Lim, Z.-H. Zhou, T.-B. Ho, D. Cheung, & H. Motoda (Eds.), *Advances in knowledge discovery and data mining* (pp. 27–38). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-18032-8_3, URL: https://rd.springer.com/chapter/10.1007/978-3-319-18032-8_3.

Pedersen, T. (2015). 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* (pp. 46–53). Denver, Colorado: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W15-1206, URL: https://aclanthology.org/W15-1206.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, a Meeting of SIGDAT, a Special interest group of the ACL* (pp. 1532–1543). ACL, http://dx.doi.org/10.3115/v1/d14-1162.

Price, I., Gifford-Moore, J., Flemming, J., Musker, S., Roichman, M., Sylvain, G., Thain, N., Dixon, L., & Sorensen, J. (2020). Six attributes of unhealthy conversations. In *Proceedings of the fourth workshop on online abuse and harms* (pp. 114–124). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.alw-1.15, URL: https://aclanthology.org/2020.alw-1.15.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training. URL: https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf.

Radloff, L. S. (1977). The CES-D scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, *1*(3), 385–401. http://dx.doi.org/10.1177/014662167700100306, URL: https://journals.sagepub.com/doi/abs/10.1177/014662167700100306.

Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.-A., & Boyd-Graber, J. (2015). Beyond LDA: Exploring supervised topic modeling for depression-related language in Twitter. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 99–107). Denver, Colorado: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W15-1212, URL: https://aclanthology.org/W15-1212.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. http://dx.doi.org/10.1162/tacl_a_00349, URL: https://aclanthology.org/2020.tacl-1.54.

Rudnicka, E., Bond, F., Grabowski, Ł., Piasecki, M., & Piotrowski, T. (2018). Lexical perspective on wordnet to wordnet mapping. In *GWC*.

Salminen, J. O., Al-Merekhi, H. A., Dey, P., & Jansen, B. J. (2018). Inter-rater agreement for social computing studies. In *2018 fifth international conference on social networks analysis, management and security (SNAMS)* (pp. 80–87). IEEE, http://dx.doi.org/10.1109/SNAMS.2018.8554744, URL: https://ieeexplore.ieee.org/abstract/document/8554744.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv:1910.01108.

Schockaert, S., & Gutiérrez-Basulto, V. (2022). Modelling symbolic knowledge using neural representations. In M. Šimkus, & I. Varzinczak (Eds.), *Reasoning web. Declarative artificial intelligence* (pp. 59–75). Cham: Springer International Publishing.

Schuster, M., & Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, *45*, 2673–2681.

Schwartz, H. A., Eichstaedt, J., Kern, M. L., Park, G., Sap, M., Stillwell, D., Kosinski, M., & Ungar, L. (2014). Towards assessing changes in degree of depression through facebook. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 118–125). Baltimore, Maryland, USA: Association for Computational Linguistics, http://dx.doi.org/10.3115/v1/W14-3214, URL: https://aclanthology.org/W14-3214.

Shoeybi, M., Patwary, M., Puri, R., LeGresley, P., Casper, J., & Catanzaro, B. (2019). Megatron-lm: Training multi-billion parameter language models using model parallelism. arXiv preprint arXiv:1909.08053.

Singh, S., Roy, D., Sinha, K., Parveen, S., Sharma, G., & Joshi, G. (2020). Impact of COVID-19 and lockdown on mental health of children and adolescents: A narrative review with recommendations. *Psychiatry Research*, *293*, Article 113429. http://dx.doi.org/10.1016/j.psychres.2020.113429, URL: https://linkinghub.elsevier.com/retrieve/pii/S016517812031725X.

Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719. http://dx.doi.org/10.1142/S0218001409007326, URL: https://www.worldscientific.com/doi/abs/10.1142/S0218001409007326.

Tian, X., Yu, G., & He, F. (2016). An analysis of sleep complaints on sina weibo. *Computers in Human Behavior*, *62*, 230–235. http://dx.doi.org/10.1016/j.chb.2016.04.014, URL: https://www.sciencedirect.com/science/article/pii/S0747563216302795.

Tolentino, J. C., & Schmidt, S. L. (2018). DSM-5 criteria and depression severity: Implications for clinical practice. *Frontiers in Psychiatry*, *9*, 450. http://dx.doi.org/10.3389/fpsyt.2018.00450, URL: https://www.frontiersin.org/article/10.3389/fpsyt.2018.00450/full.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Proceedings of the 31st international conference on neural information processing systems* NIPS '17, (pp. 6000–6010). Red Hook, NY, USA: Curran Associates Inc., URL: https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Vermeulen, A., Vandebosch, H., & Heirman, W. (2018). #Smiling, #venting, or both? Adolescents' social sharing of emotions on social media. *Computers in Human Behavior*, *84*, 211–219. http://dx.doi.org/10.1016/j.chb.2018.02.022, URL: https://www.sciencedirect.com/science/article/pii/S0747563218300803.

Vidgen, B., Harris, A., Nguyen, D., Tromble, R., Hale, S., & Margetts, H. (2019). Challenges and frontiers in abusive content detection. In *Proceedings of the third workshop on abusive language online* (pp. 80–93). Florence, Italy: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/W19-3509, URL: https://aclanthology.org/W19-3509.

Vincent, J. (2017). Facebook is using AI to spot users with suicidal thoughts and send them help. URL: https://www.theverge.com/2017/11/28/16709224/facebook-suicidal-thoughts-ai-help.

Watson, J. (2007). Big huge thesaurus. URL: https://words.bighugelabs.com/.

World Health Organization (1993). 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* (pp. xiii–248).

Wu, L., Petroni, F., Josifoski, M., Riedel, S., & Zettlemoyer, L. (2020). Zero-shot entity linking with dense entity retrieval. In *EMNLP*.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Kaiser, Ł., Gouws, S., Kato, Y., Kudo, T., Kazawa, H., ... Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. CoRR abs/1609.08144. URL: http://arxiv.org/abs/1609.08144.

Yadav, S., Chauhan, J., Sain, J. P., Thirunarayan, K., Sheth, A., & Schumm, J. (2020). Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework. In *Proceedings of the 28th international conference on computational linguistics* (pp. 696–709). Barcelona, Spain (Online): International Committee on Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.coling-main.61, URL: https://aclanthology.org/2020.coling-main.61.

Yamada, I., Asai, A., Shindo, H., Takeda, H., & Matsumoto, Y. (2020). LUKE: Deep contextualized entity representations with entity-aware self-attention. In *Proceedings of the 2020 conference on empirical methods in natural language processing* EMNLP, (pp. 6442–6454). Online: Association for Computational Linguistics, http://dx.doi.org/10.18653/v1/2020.emnlp-main.523, URL: https://aclanthology.org/2020.emnlp-main.523.

Yazdavar, A. H., Al-Olimat, H. S., Ebrahimi, M., Bajaj, G., Banerjee, T., Thirunarayan, K., Pathak, J., & Sheth, A. (2017). Semi-supervised approach to monitoring clinical depressive symptoms in social media. In *Proceedings of the 2017 IEEE/ACM international conference on advances in social networks analysis and mining 2017* ASONAM '17, (pp. 1191–1198). New York, NY, USA: Association for Computing Machinery, http://dx.doi.org/10.1145/3110025.3123028.

Zhu, Y., Kiros, R., Zemel, R. S., Salakhutdinov, R., Urtasun, R., Torralba, A., & Fidler, S. (2015). 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* (pp. 19–27). IEEE Computer Society, http://dx.doi.org/10.1109/ICCV.2015.11, URL: https://www.cv-foundation.org/openaccess/content_iccv_2015/html/Zhu_Aligning_Books_and_ICCV_2015_paper.html.

Mohsinul Kabir is currently working as an Assistant Professor in the Department of Computer Science & Engineering in Islamic University of Technology (IUT), Bangladesh. He received his B.Sc. and M.Sc. degrees in Computer Science from IUT in 2018 and 2022, respectively. He is extremely motivated to work in interdisciplinary domains where technology can have a positive impact on human life. His research interests include NLP, HCI, Religion, and Astronomy.

Tasnim Ahmed received B.Sc in Computer Science and Engineering degree from Islamic University of Technology (IUT) in 2019. Currently he is working as a Lecturer in the same university. His research interests include Biomedical Sequence and Signal Processing, Natural Language Processing, and Computer Vision. Besides teaching and research works, he is very enthusiastic in maintaining his YouTube channel which contains educational videos on various computer science topics such as, Computer Programming, Application Development, OOP, System Design, etc.

Md. Bakhtiar Hasan received his B.Sc.Engg. and M.Sc.Engg. degrees in Computer Science and Engineering (CSE) from Islamic University of Technology (IUT), in 2018 and 2022, respectively. Upon graduation, he joined IUT as a Lecturer. Since 2022, he has been working as an Assistant Professor with the Department of Computer Science and Engineering, IUT. His research interest includes the use of deep learning and computer vision techniques in human biometrics and smart agriculture.

Md Tahmid Rahman Laskar is currently working as an Applied Scientist at Dialpad, Canada. Previously, he did his M.Sc. from York University, Canada and B.Sc. from Islamic University of Technology, Gazipur, Dhaka. His research interest is within the area of Natural Language Processing and Machine Learning. He also published several research papers in reputed venues like Computational Linguistics, COLING, ACM Transactions on Cyber Physical Systems, LREC, and Canadian AI.

Tarun Kumar Joarder is working as an Associate Professor in the Department of Psychology, University of Rajshahi. He has a Ph.D. in Psychology and loves to work with Disaster psychology, Industrial Psychology, Sports Psychology, Mental Health, Aggressive Behavior, Psychology of Crime, etc.

Hasan Mahmud is working as an Assistant Professor in the CSE department of IUT. He has been involved in HCI research since 2009. His specialization is in the area of gesture-based interaction through machine learning approaches, affective computing, and assistive technology for the physically impaired.

Kamrul Hasan has received his Ph.D. from Kyung Hee University, South Korea. Currently, he is working as a Professor of CSE, IUT. He has expertise in intelligent systems and AI, software engineering, data mining applications, and social networking. He is the founding director of SSL research lab, IUT.