Monitoring Depression Severity and Symptoms in User-Generated Content: An Annotation Scheme and Guidelines

Falwah AlHamed[1,2], **Rebecca Bendayan**[3], **Julia Ive**[4], and **Lucia Specia**[1]

[1]Department of Computing, Imperial College London, London, UK
[1]{f.alhamed20,l.specia}@imperial.ac.uk
[2]King Abdulaziz City for Science and Technology(KACST), Riyadh, Saudi Arabia
[3]King's College London, London, UK
[4]Queen Mary University of London, London, UK
[4]j.ive@qmul.ac.uk

Abstract

Depression is a highly prevalent condition recognized by the World Health Organization as a leading contributor to global disability. Many people suffering from depression express their thoughts and feelings using social media, which thus becomes a source of data for research in this domain. However, existing annotation schemes tailored to studying depression symptoms in social media data remain limited. Reliable and valid annotation guidelines are crucial for accurately measuring mental health conditions for those studies. This paper addresses this gap by presenting a novel depression annotation scheme and guidelines for detecting depression symptoms and their severity in social media text. Our approach leverages validated depression questionnaires and incorporates the expertise of psychologists and psychiatrists during scheme refinement. The resulting annotation scheme achieves high inter-rater agreement, demonstrating its potential for suitable depression assessment in social media contexts.

1 Introduction

Within the domain of mental health, a multitude of disorders exists, each characterized by distinct symptoms that influence cognitive processes, emotional states, and behavioural patterns. This study directs its focus toward depression, a prevalent condition acknowledged by the World Health Organization (WHO) as a significant contributor to global disability (McManus et al., 2009). According to WHO estimates, approximately 264 million individuals worldwide suffer from the burdens of depression. Understanding the occurrence and severity of depression in online platforms can offer valuable insights for early detection, intervention, and support (Association, 2013). However, extracting meaningful information about depression from social media posts presents significant challenges due to the unstructured and nuanced nature of the content.

Labelling social media data for mental disorders like depression is a common practice in research, yet it presents notable challenges. Unlike clinical data, social media lacks validated indicators of sadness or formal diagnoses, necessitating the development of labelling techniques. Achieving consensus on and applying these labels proves challenging due to the subjective nature of mental health evaluation and the need for nuanced annotation schemes. However, existing schemes tailored to studying depression symptoms in social media data are limited. Ensuring the reliability and validity of such guidelines is important to accurately measure mental health issues in social media studies.

2 Related Work

The accurate annotation of user-generated content (UGC) is essential for developing reliable datasets to train machine learning models for various mental health applications (De Choudhury et al., 2013; Chancellor et al., 2021). Prior research has recognized the importance of creating annotation schemes specifically for labelling mental health data extracted from social media platforms(Benetka et al., 2020; Mowery et al., 2015; Straton et al., 2020). The first pilot study on annotation scheme for depression was conducted by Mowery et al. (2015) leveraging the DSM-5 criteria for item definition. However, their study employed a relatively small dataset for annotation, in which the dataset was collected only based on "depression" keywords in social media, which might have resulted in the inclusion of data from non-depressed users, potentially compromising the scheme's accuracy in reflecting true depression. Additionally, they reported low inter-annotator agreement, raising concerns regarding the scheme's applicability and reliability. Another study by (Yao et al., 2021) investigated the development of an annotation scheme for depression in online discussions

Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, pages 227–233

on the Chinese social media platform Sina Weibo. Their work focused specifically on Chinese forums and employed accuracy as the metric for inter-annotator agreement. It is important to note that accuracy can be inflated by chance agreement, potentially overestimating the scheme's reliability. These limitations necessitate further refinement of annotation schemes for depression to ensure their robustness and broad applicability. A recent study by Chancellor et al. (2021) addressed the challenge of annotating suicide risk and protective factors within online support forums. Their work yielded an annotation scheme and guidelines that achieved high inter-annotator agreement. Their approach emphasized incorporating the expertise of psychologists during guideline design. Additionally, they identified key considerations for developing robust annotation schemes, which informed the methodology employed in the present study.

3 Dataset

For the dataset, the study targeted platforms with a significant volume of textual content in English. The source of the data used in this experiment was introduced in (Alhamed et al., 2024). It consisted of tweets from users who self-disclosed being diagnosed with depression. The dataset underwent manual inspection to select only original tweets, excluding replicated tweets or narratives about others. Only users who specified the month and year of their diagnosis were included. Posts preceding and following the diagnosis date were extracted. The final dataset contains more than 1 million posts of people who self-reported being diagnoed with depression, with each post labelled as "before" or "after" depression diagnosis.

4 Depression Standardized Questionnaires

Depression is characterized by several symptoms, substantially impairing people's ability to function at work or school and to cope with daily life. At its most severe, depressive symptoms can be linked to suicidal ideation and are associated with a high risk for suicide. In the context of diagnostic and screening protocols for this illness, clinicians commonly administer standardized questionnaires to patients. These questionnaires consist of a series of questions about the patient's emotional state and daily activities over a designated timeframe. Based on the patient's responses, sometimes a score is generated to ascertain whether the patient exhibits symptoms of depression and to determine the severity level. To build the annotation scheme, we relied on three popular standardized questionnaires: the Patient Health Questionnaire (PHQ-9), the Beck Depression Inventory (BDI), and the Center for Epidemiologic Studies Depression Scale (CES-D). PHQ-9 is a validated depression screening tool developed by Kroenke et al. (2001). It comprises nine questions corresponding to the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV) criteria for major depressive disorder. Individuals rate the frequency of experiencing depressive symptoms over the past two weeks on a scale from 0 to 3. Scores are summed to indicate the severity of depressive symptoms, with higher scores suggesting greater impairment. CES-D is a validated questionnaire developed by Radloff (1977) to measure the presence and severity of depressive symptoms in the general population. Consisting of 20 items covering various aspects of depressive symptomatology, such as depressed mood, feelings of guilt and worthlessness, sleep disturbance, and loss of appetite, the CES-D provides a reliable assessment of depression severity. BDI is a widely used and well-validated tool for measuring depressive symptomatology in adults (Beck et al., 1961). BDI is a self-report questionnaire consisting of 21 items, each addressing a specific cognitive or behavioural symptom of depression (Beck et al., 1996). Respondents rate the severity of each symptom on a 4-point Likert scale, resulting in a total score that reflects the level of depression present. These questionnaires are widely used in clinical practice and research due to their brevity, simplicity, and demonstrated reliability and validity in assessing depression severity across diverse populations.

Our scheme draws upon the foundations laid by these three validated questionnaires. We extracted symptoms from these questionnaires and we explored their usability and appropriateness to extract symptoms from social media. We did this by collecting feedback from a panel of experts of psychologists and psychiatrists. The aim is to evaluate which symptoms could be discerned from social media platforms.

5 Study Design

In this section, we provide an overview of the procedures and methodologies employed in the development of our annotation scheme. The main

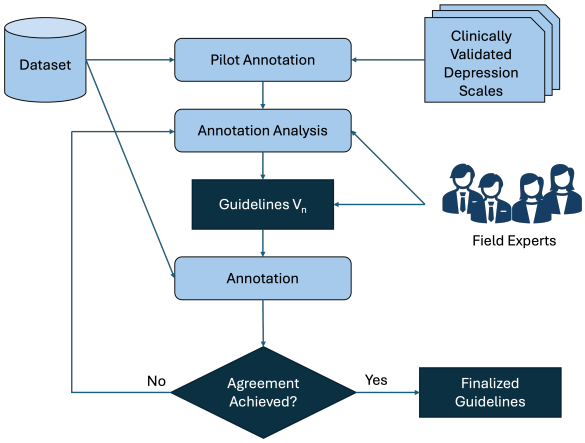

Figure 1: Annotation Process

goal of our work is to lay a foundation towards building a valid and reliable annotating scheme for depression that is able to: 1) indicate symptoms of depression from user-generated content in social media, 2) indicate the occurrence and severity of depression from user-generated content in social media. We used Labelstudio [1] as a labelling interface for all experiments in this work. Within Label Studio, we designed a custom labelling interface to meet the specific needs of our task, as none of the available templates offered a suitable match.

5.1 Annotators

The annotators for this task are five clinical psychologists, each possessing a minimum of three years of specialized experience in diagnosing depression and/or anxiety disorders. Their participation in this task is entirely voluntary, without any incentives, and motivated by a commitment to improve mental health research.

6 Scheme Development

In this section we are explaining in detail the procesudres and methods taken to develop the annotation scheme. First we looked into the most used questionnaires for diagnosing depression in clinical practice and in research. Then, we created a survey based on all symptoms occurred in these questionnaires. The survey was for psychologits/psychiatrsits to narrow down the symptoms and choose which of these symptoms can be detectd from texts posted on a social media platform. After that, the selected symptoms was categorized in a shorter list to facilitate the annotation procedure by annotators. each step is thoroughly described in the following sections.

6.1 Annotating Depression Symptoms

Initially, we constructed a survey encompassing all symptoms identified in validated depression questionnaires, namely: PHQ-9, CES-D, and BDI. This survey was then distributed to psychologists and psychiatrists, who were tasked with identifying symptoms potentially detectable from textual content posted on social media platforms. Following this, the selected symptoms were refined and categorized into a concise list to streamline the annotation process for annotators. Each of these steps is detailed in the subsequent sections.

6.2 Psychologists/Psychiatrists Survey

To ascertain the ability of each symptom to indicate depression from textual descriptions, we engaged psychology experts in a survey. We conducted a survey containing all symptoms from all three questionnaires, which resulted in 50 symptoms. The survey then was distributed to 17 psychologists and psychiatrists. The aim was to select items from questionnaires that reflect symptoms of depression that can be manifested within users' text on social media platforms. The survey aimed to determine whether each symptom could independently signify depression or if it necessitated accompanying symptoms for diagnostic clarity, or whether the item by itself can be used to identify clinically significant depression. Based on the collective insights garnered from this survey, we refined the initial list to include only symptoms that could effectively identify depression either in isolation or when coupled with other symptoms, 40 symptoms resulted from this step.

[1] https://labelstud.io/

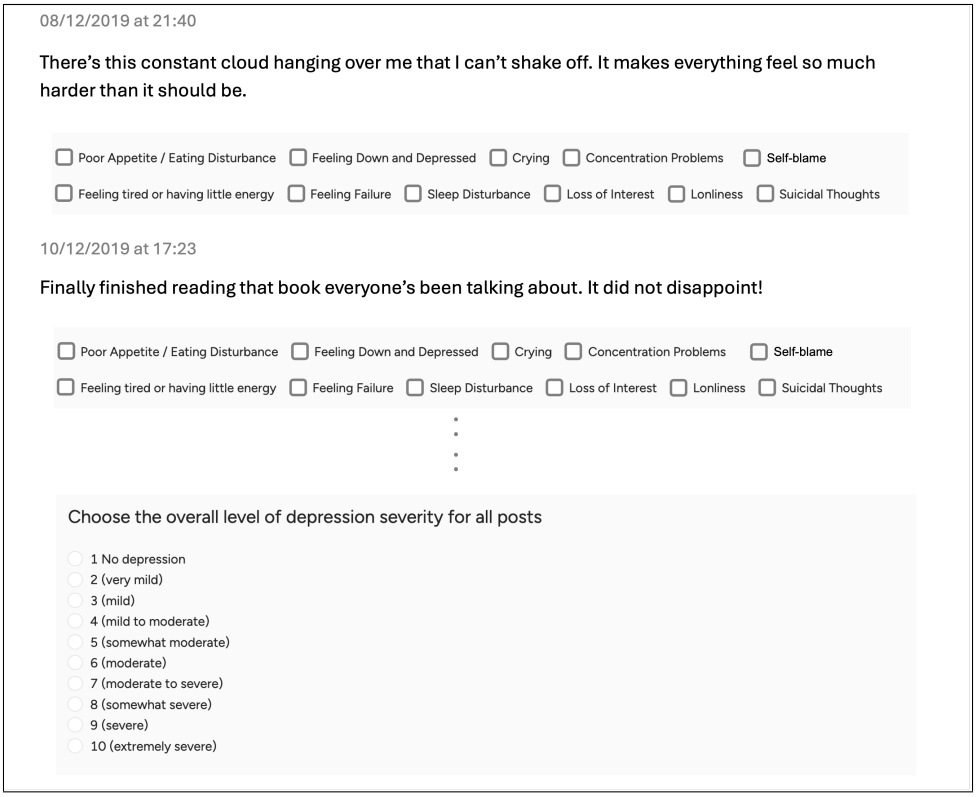

Figure 2: Example of annotating a chunk that contains multiple posts. Depression symptoms annotation task is post-level, while severity annotation task is chunk-level. Posts are generated using the OpenAI GPT-4 model, closely mimicking original posts to protect users' privacy.

6.3 Categorizing and Refining Annotation Items

We conducted a comprehensive analysis of depression symptoms derived from survey results which are 40 distinct symptoms. Considering the potential overlap of symptoms across the aforementioned questionnaires, and to streamline and condense the extensive list of symptoms, we proceeded to categorize them into a concise set of symptom categories (details can be found in Appendix A). This categorization process aimed to facilitate a more efficient and manageable list for the identification and assessment of depression symptoms. The final list consists of 11 symptom categories: poor appetite/eating disturbance, feeling down and depressed, crying, concentration problems, feeling tired or having little energy, feeling failure, sleep disturbance, loss of interest, self-blame and shame, loneliness, and suicidal thoughts.

6.4 Annotation Process

The annotation process employed a cyclical approach to ensure validity and annotator agreement. First, a pilot scheme was conducted utilizing established depression scales to develop an initial annotation framework. Subsequently, five independent annotators applied this framework to the data (450 posts). Following this initial annotation round, a collaborative analysis was undertaken. This analysis involved both the annotators (clinical psychologists) and field experts. The field experts included a clinical psychology consultant with expertise in annotation guideline development and a computer science specialist experienced in annotating mental health applications. Their feedback on the annotation process and scheme informed subsequent modifications. With these revisions incorporated, a second round of annotation was conducted utilizing the refined scheme. This iterative process of annotation, analysis, and refinement was repeated for a total of three rounds. This cyclical approach led to the establishment of final depression annotation guidelines and a scheme deemed to be valid by the field experts and achieved a satisfactory level of annotator agreement. The annotation process is illustrated in Figure 1.

6.5 Annotating Depression Severity

In the context of annotating depression severity in social media posts, experts in psychology have

recommended incorporating a timeframe of one to two weeks to accurately measure the occurrence and severity of depressive symptoms. Aligning with the Center for Epidemiologic Studies Depression Scale (CES-D), we segmented tweets into 1-week intervals, herein referred to as "chunks." Each chunk represents tweets spanning a week, requiring annotators to review all tweets within the chunk to assess depression occurrence and severity. Initially, our depression severity scheme comprised four categories: **No depression** (indicating absence of depressive symptoms), **Mild depression** (denoting mild indications of depression), **Moderate depression** (suggesting moderate manifestations of depression), and **Severe depression** (representing severe symptoms or inclinations towards suicidal thoughts).

In a pilot study involving 45 chunks, annotators encountered challenges in accurately categorizing depression severity, particularly when it fell between two predefined categories, such as mild and moderate. Consequently, a suggestion emerged to enhance the granularity of the severity scale. To address this, we transitioned to a **10-point severity scale** (0 to 9 where 0 indicates no depression and 9 indicates extremely severe depression). This adaptation aims to provide a more nuanced framework, facilitating a finer alignment of observed symptoms with corresponding severity levels. An example of our final annotation scheme with example posts is shown in Figure 2. It is noteworthy that the adjustment of severity levels from 4 to 10 scale might lead to decreased inter-rater agreement. However, the primary aim is to enhance the reliability and precision of data annotation, thereby fostering more meaningful insights for psychologists and psychiatrists analyzing social media posts concerning depression severity.

7 Results

To assess the consistency of our annotations, we employed Cohen's kappa (κ), a well-established statistic for measuring inter-rater agreement for nominal data (Cohen, 1960). This metric accounts for agreement that may occur by chance, providing a more robust concordance measure than simple accuracy agreement. In our study, we achieved a pairwise kappa score of 0.67 for 45 chunks of tweets, encompassing a total of 450 individual posts annotated. This value falls within the range typically interpreted as indicating "substantial" agreement (Landis and Koch, 1977). The high level of agreement achieved through kappa analysis strengthens the reliability of our findings and underscores the consistency with which the annotation scheme was applied. We posit that the high inter-annotator agreement achieved in this work stems, at least in part, from the collaborative approach involving computer scientists and field experts with psychological and psychiatric backgrounds. This collaborative effort ensured that the annotation scheme and guidelines were grounded in both technical expertise and clinical knowledge. This work has the potential to significantly contribute to the field of digital health and social media. The proposed scheme and guidelines can serve as a robust baseline for collecting and labelling high-quality, gold-standard datasets. Machine learning models trained on such datasets could be developed to detect depression symptoms and assess their severity. If integrated with established mental health support systems, these models could potentially function as preventive tools by facilitating early intervention.

8 Conclusion

This work addressed the critical need for robust annotation schemes for detecting depression symptoms and severity in social media text. We presented an annotation scheme and corresponding guidelines informed by validated depression questionnaires (PHQ-9, CES-D, and BDI) and refined through collaboration with psychologists and psychiatrists. The resulting scheme demonstrates good inter-rater agreement (Cohen's kappa = 0.67), signifying its potential for reliable depression assessment in social media contexts. This scheme and its accompanying guidelines can serve as a valuable foundation for collecting and labelling high-quality, gold-standard datasets. Future research will leverage this scheme to create a labelled dataset and develop machine learning models capable of accurately detecting depression symptoms from social media data.

Limitations

The annotation scheme is designed for screening purposes and is not intended for clinical diagnosis of depression. While tested on a specific dataset, further validation across diverse datasets, including different languages, is necessary to establish its applicability and reliability. Moreover, the scheme is specifically tailored to evaluate depression symp-

toms and severity, potentially limiting its application to other mental health conditions.

Ethics Statement

This study has received ethics approval from the Science Engineering Technology Research Ethics Committee at Imperial College London (SETREC Reference: 21IC7222).

References

Falwah Alhamed, Julia Ive, and Lucia Specia. 2024. Classifying social media users before and after depression diagnosis via their language usage: A dataset and study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3250–3260.

American Psychiatric Association. 2013. *Diagnostic and statistical manual of mental disorders : DSM-5*, 5th ed. edition. American Psychiatric Association Arlington, VA.

Aaron T Beck, Michael H Steer, and Gregory K Brown. 1996. *Manual for the Beck Depression Inventory-II*. Psychological Corporation.

Aaron T Beck, Charles H Ward, Morris Mendelsohn, John Mock, and James Erbaugh. 1961. An inventory for measuring depression. *Archives of General Psychiatry*, 4(6):561–571.

David Benetka, Alicia Moreno-Moral, Lorena Romero-Fombuena, and Juan Lopez-Gazpio. 2020. An annotation scheme for mental health discussions in social media. In *International Conference on Computational Linguistics (Proceedings of the Conference: Long Papers, 2020)*, pages 2617–2627. Association for Computational Linguistics.

Stevie Chancellor, Steven A Sumner, Corinne David-Ferdon, Tahirah Ahmad, and Munmun De Choudhury. 2021. Suicide Risk and Protective Factors in Online Support Forum Posts: Annotation Scheme Development and Validation Study. *JMIR Ment Health*, 8(11):e24471.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Munmun De Choudhury, Shanika M De Silva, K Wiemer-Hastings, and James W Pennbaker. 2013. Identifying depression using social media. In *The future of mental health: An international perspective*, pages 170–180.

Kurt Kroenke, Robert L Spitzer, and Janet B Williams. 2001. The phq-9: validity of a brief depression severity measure. *Journal of General Internal Medicine*, 16(9):606–613.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

S. McManus, H. Meltzer, T. Brugha, P. E. Bebbington, and R. Jenkins. 2009. Adult psychiatric morbidity in england: results of a household survey.

Danielle L Mowery, Craig Bryan, and Mike Conway. 2015. Towards developing an annotation scheme for depressive disorder symptoms: A preliminary study using twitter data. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 89–98.

Lenore S Radloff. 1977. The ces-d scale: A self-report depression scale for research in the general population. *Applied Psychological Measurement*, 1(3):385–401.

Nadiya Straton, Hyeju Jang, and Raymond Ng. 2020. Stigma annotation scheme and stigmatized language detection in health-care discussions on social media. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1178–1190, Marseille, France. European Language Resources Association.

Xiaoxu Yao, Guang Yu, Jingyun Tang, and Jialing Zhang. 2021. Extracting depressive symptoms and their associations from an online depression community. *Computers in Human Behavior*, 120:106734.

A Symptoms Categories List

A list of symptoms and the corresponding symptom category is shown in Table 1.

I did not feel like eating; my appetite was poor	Poor Appetite / Eating Disturbance
I was bothered by the problem: Poor appetite or overeating	
My appetite is much worse than before.	
I have lost/gained noticeable weight	
I felt that I could not shake off the blues even with help from my family or friends.	Feeling Down and Depressed
I felt depressed	
I felt sad	
I was bothered by the problem: feeling down, depressed, or hopeless	
I am sad all the time and I can't snap out of it.	
I had crying spells	Crying
I cry all the time	
I used to be able to cry, but over the last 2 weeks I can't cry even though I want to.	
I had trouble keeping my mind on what I was doing.	Concentration Problems
I was bothered by the problem: Trouble concentrating on things, such as reading the newspaper or watching tv	
I feel irritated all the time.	
I have greater difficulty in making decisions more than I used to	
I felt that everything I did was an effort.	Feeling tired or having little energy
I could not get "going"	
I was bothered by the problem: Feeling tired or having little energy	
I was bothered by the problem: Moving or speaking so slowly that other people could have noticed. Or the opposite being so fidgety or restless that I have been moving around	
I have to push myself very hard to do anything. or I can't do any work at all.	
I get tired from doing almost anything	
I thought my life had been a failure.	Feeling Failure
I was bothered by the problem: Feeling bad about myself, or that I am a failure or let myself or my family down	
I feel I have nothing to look forward to.	
As I look back on my life, all I can see is a lot of failures.	
My sleep was restless	Sleep Disturbance
I was bothered by the problem: Trouble falling or staying asleep, or sleeping too much	
I wake up several hours earlier than I used to and cannot get back to sleep.	
I was bothered by the problem: Little interest or pleasure in doing things	Loss of Interest
I don't get real satisfaction out of anything anymore. or I am dissatisfied or bored with everything.	
I have lost most of my interest in other people	
I have almost no interest in sex.	
I feel quite guilty most of the time.	Self-Blame and Shame
I expect to be punished.	
I am disgusted with myself.	
I blame myself all the time for my faults.	
I felt lonely	Loneliness
I was bothered by the problem: Thoughts that I would be better off dead, or of hurting myself	Suicidal Thoughts
I would like to kill myself.	

Table 1: Shortened list of depression symptoms with the finalized categories.