JCSHD COGNITIVE SCIENCES AND HUMAN DEVELOPMENT

Classifying depression severity in online chats through human-coded psycholinguistic analysis using DSM-5 and Beck Depression Inventory

Ross Azura Zahit[1], Amalia Madihie[1], Salmah Mohamad Yusoff[1], Ida Juliana Hutasuhut[1], Mohamad Azhari Abu Bakar[1], Mohamad Hardyman Barawi[1], Syahrul Nizam Junaini[2] & Nur Haziyah Amni Raimaini[*1]

[1]Faculty of Cognitive Sciences and Human Development, Universiti Malaysia Sarawak, 94300, Kota Samarahan, Malaysia.

[2]Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, 94300, Kota Samarahan, Malaysia.

ABSTRACT

Despite advances in artificial intelligence, accurately detecting the severity of depression in online communications remains a challenge, underscoring the need for expert-led psycholinguistic analysis. This study employs such an approach to examine depression and other mental health issues in online chat data. Depression severity was classified using DSM-5 and Beck's Depression Inventory by five mental health professionals, with results tested for inter-rater reliability. Human-coded psycholinguistics adds clinical nuance to the classification. A random sample of 4,000 chat entries was analysed, with five professionals independently categorising each entry based on the DSM-5 and Beck Depression Inventory criteria into the categories of no depression, mild, moderate, severe, or unknown. The analysis showed a high average inter-rater reliability, indicating substantial agreement among raters. Results revealed that 7% of chats exhibited some level of depression (2% mild, 2% moderate, 3% severe), 19% indicated other mental health issues such as anxiety, and 58% were ambiguous. These findings suggest that psycholinguistic analysis of online communication has strong potential for early detection of mental health issues. Integrating such features into digital tools could enhance early identification on online platforms, enabling timely intervention and better mental health support within communities.

Keywords: online chats, depression, mental health, social media, psycholinguistic

ARTICLE INFO
Email address: amnirai21@gmail.com (Nur Haziyah Amni Raimaini)
*Corresponding author
https://doi.org/10.33736/jcshd.10058.2025
e-ISSN: 2550-1623

Manuscript received: 26 June 2025; Accepted: 20 September 2025; Date of publication: 30 September 2025

1 INTRODUCTION

In today's era of rapid technological advancement, social media has become an integral part of daily life, providing unprecedented access to platforms like Instagram, Facebook, Twitter, Reddit, and Discord. These platforms have evolved into popular venues where individuals express opinions, thoughts, and feelings on a myriad of topics, including mental health issues such as depression. A meta-analysis and review conducted by Solmi et al. (2022) has shown a trend that the onset of any mental health disorder starts at 14.5 years old. The study also shows that most mental health issues emerge during childhood, and depression usually starts during early teens; however, it has a broad window of onset, but typically it occurs before the age of 30 years old. Hence, it is no surprise that depression, which was once considered a taboo subject, has become more openly discussed within online communities. The anonymity afforded by online profiles encourages individuals to share their struggles and experiences with depression more freely. Ostic et al. (2021) emphasise that social media has become a prominent platform for voicing concerns about mental well-being (Chancellor & De Choudhury, 2020; Ostic et al., 2021). For introverted individuals, these platforms serve as preferred channels to address personal issues they might hesitate to discuss in face-to-face interactions.

The growing interest among data scientists in analysing language within online mental health communities has led to the application of natural language processing (NLP) techniques to text data from blogs, tweets, and messages related to mental health (Li et al., 2020; Low et al., 2020). A study on detecting late-life depression highlighted the use of NLP in analysing the linguistic and acoustic elements derived from text and speech (DeSouza et al., 2021). NLP offers advantages over traditional psychological methods (Jackson et al., 2022)' holding the potential to uncover valuable insights into how individuals discuss mental health issues in digital contexts.

However, many researchers have utilised NLP to differentiate between depression and non-depression using binary classification methods. This approach is insufficient because theoretical perspectives recognise varying levels of depression severity, mild, moderate, and severe, each with distinct indicators, as outlined in the Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5) by the American Psychiatric Association (2013). Therefore, it is crucial to consider the severity of depression when developing and applying NLP techniques to identify it within online communities. Moreover, Korhonen et al. (2022) emphasised the need for high construct validity and a clear classification process for identifying mental disorders to enhance diagnostic accuracy.

Psycholinguistic research has increasingly incorporated inter-rater methodologies to study linguistic behaviour (Bijou et al., 1986; Mehta et al., 2020; Pennebaker et al., 2003). With the proliferation of social media, text analysis has become a vital tool for understanding individuals (Amanat et al., 2022; Ziemer & Korkmaz, 2017). Recent studies have compared automated analyses of psychological and physical health to human raters (Ziemer & Korkmaz, 2017), detected depression by examining social media posts on platforms like Twitter, Facebook, Reddit, and electronic diaries (Chiong et al., 2021), and identified early signs of depression through the emotional content of multiple social media users (Amanat et al., 2022). Research has also consistently shown that language is able to indicate the mental well-being of an individual. As an example, the extensive use of first-person singular pronouns is associated with symptoms of depression and anxiety within the context of a joyful memory recall and not an unfortunate memory recall (Brockmeyer et al., 2015). This study by Brockmeyer et al. (2015) utilises Linguistic Inquiry Word Count (LIWC), which counted the singular pronouns used between clinically and non-clinically depressed groups.

Another recent study by Ren et al. (2023) suggested that the use of the singular pronoun "I" is more extensive among the clinically depressed group. The latter research is more focused on developing more informed tools to screen mental health issues, which utilises a newer method which includes contextual circumstances in determining the context of the text. Aspects like cognitive distortions are often visible in language as well. As an example, absolute thinking. "Absolute" refers to the ideas, words, and phrases that promote totality, either in terms of intensity or probability. Al-Mosaiwi and Johnstone (2018) have suggested in their findings that there is an elevated use of absolute words in groups with depression and anxiety as compared to the control group. The finding also suggests that these absolute markers are also more prominent in groups with suicidal ideation.

Additionally, there is a growing interest in incorporating digital data to understand better and monitor mental health conditions. Beyond traditional self-report measures and clinical interviews, researchers have examined social media posts, text messages, smartphone sensor data, and other digital traces as potential indicators of psychological states. Mobile sensing data, including sleep patterns, physical activity, and communication logs, have also been shown to predict fluctuations in mood and stress. In a meta-review conducted by Amin et al. (2025), six studies (67%) suggested that mobile sensing data can predict depressive symptoms and their severity. These approaches fall under the broader framework of digital phenotyping, which refers to the moment-by-moment quantification of behaviour and mental health using personal digital devices (Oudin et al., 2023). Through the incorporation of this digital phenotyping and mobile sensing data with expert-in-the-loop, it could holistically improve the readily available screening tools.

Recent research offers ground-breaking insights into how online communications can serve as indicators for detecting depression. Huang (2022) delved into creating intelligent chatbots capable of discerning depression from textual interactions, acknowledging the challenges in fully understanding human emotions. Similarly, Cai et al. (2023) integrated machine learning techniques to detect depressive symptoms among users through multivariate time series analysis, focusing on users' online behaviours, such as posting frequency, language use, and interaction patterns. Cacheda et al. (2019) explored machine learning models that analysed users' semantics, text, and writing styles to detect early signs of depression on social networking sites, resulting in a 10% improvement in performance using their proposed single-model approach.

Burkhardt et al. (2021) highlighted the role of behavioural activation in mitigating depression, showcasing how text-based therapy sessions can provide insights into patients' psychological states through their language use. Furthermore, Dewangan et al. (2022) explored the potential of machine learning in interpreting emotions from textual messages to diagnose depressive moods, emphasising the nuances of natural language processing in this context. Collectively, these studies affirm the promising potential of leveraging online communication in the early detection and intervention of depression.

The present study aims to explore affective and psycholinguistic patterns within online communities to identify indicators of depression and other mental health issues. Additionally, it seeks to determine the inter-rater reliability among mental health professionals in interpreting these patterns. Through this approach, the study endeavours to contribute to the burgeoning field of psycholinguistics in digital mental health discourse by offering new insights and methodologies for understanding and addressing mental well-being in the digital era. Prior to past NLP-based studies (Li et al., 2020; Low et al., 2020), this study approaches the topic of analysing the language through a deeper clinical nuance as it is more clinically grounded. The texts are analysed more thoroughly by the mental health experts, and this could bridge the gap that automated NLP might have missed in detecting the linguistic markers of depression and other mental health issues. The in-depth evaluation by the experts offers a new insight that the detection of mental health issues online could be done in a better and more clinically accurate way. The raters' feedback and judgments serve as a benchmark which ensures that the model's predictions align with the nuanced understanding of a human expert. This expert's inclusion approach strengthens the study's validity and provides a level of rigour that is often absent in purely computational studies. As demonstrated by Han et al. (2023), leveraging such domain expertise is critical for automated text categorisation tasks, particularly when dealing with complex, subjective, or specialised content. Han also added that the inclusion of domain expertise will significantly enhance automated categorisation, and this improvement depends on several factors: the experts' clear understanding of the categories, their confidence in their annotations, the amount of data available for a given category, and the presence of unique keywords that help identify a category. This expert-in-the-loop approach strengthens the study's validity and provides a level of rigour that is often absent in purely computational studies. Instead of just categorising the text based on depression or no depression, the texts are classified into no depression, mild depression, moderate, severe, and unknown. Thus, this study could suggest an improvement in NLP model training as well as the training of clinicians in the future.

2 METHODS

2.1 Design and Materials

This qualitative study was conducted in collaboration with a private consultant who offers online chat services for clients seeking mental health support through an online platform. The analysis of the chat items was carried out by a team of five mental health professionals, referred to as inter-raters, from a public university in Malaysia. These inter-raters, with expertise in clinical psychology, industrial psychology, and counselling, meticulously examined each chat entry for psycholinguistic patterns that could indicate depressive symptoms or other mental health concerns. Their diverse backgrounds equipped them with the necessary skills to accurately identify signs of mental health issues through language analysis.

This study was divided into two phases. In the first phase, inter-raters categorised the symptoms expressed in the chat into five categories: no depression, mild, moderate, severe, and unknown.

According to The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM–5) (American Psychiatric Association, 2013) and Beck Depression Inventory (Beck et al., 1961), the categorisation of depression is based on the severity and duration of occurrence. Symptoms such as depressed mood, loss of interest, feeling worthless, fatigue, insomnia, and suicidal ideation must be present for at least two weeks, which could affect their activities of daily living (ADL).

The five inter-raters made an agreed definition of depression classification, namely:

1. No depression: The chat only indicates the existence of anxiety, stress, and other emotional disturbances that arise due to certain triggers, and that person is still carrying out their daily functions properly.
2. Mild depression: The chat indicates the existence of ongoing anxiety, stress, and other emotional disturbances, and that person is still able to function appropriately to some degree or some of the time (approximately two weeks).
3. Moderate depression: The chat indicates the experience of one or two depressive symptoms that result in a decrease in function that creates problems in daily living activities (personal and work). Symptoms are assumed to persist for at least two weeks, as inferred from the text.
4. Severe depression: The chat indicates the experience of various depressive symptoms that result in a decrease in function that leads to feelings of worthlessness or not worth living or not being able to cope with the prolonged pain of depression that dominates the function of cognitive, emotional, physical, and physiological, so that the thoughts are focused on ending one's life.
5. Unknown: The chat indicates all statements, wishes, and greetings, as well as something that falls outside the characteristics of "no depression, mild, moderate, and severe depression", as mentioned above. The chat does not clearly fit the criteria for any of the categories above.

In the second phase, inter-raters analysed the classification of depression and other mental health issues using inter-rater reliability and descriptive analysis. As the nature of the chat items is cross-sectional, reflecting only a single period of time, it is crucial to note that the operational definitions do not include longitudinal patterns, and the cumulative symptom count is not available. Hence, observable language and indicators in the texts were analysed to interpret the contextually limited chat items.

2.2 Sample and Data Collection

A randomised sample of 4,000 online chat items was extracted from a larger dataset containing 22,500 chat entries provided by the consultant. The selection process involved using a random sampling technique to ensure that the chosen 4,000 chats were representative of the broader dataset, minimising bias and enhancing the reliability of the findings. Each of the five inter-raters was individually assigned this sample of chat items and instructed not to discuss the categorisation process with one another, ensuring independent analysis. The inter-raters were guided by a predefined and agreed-upon classification system for depression, categorising each chat into one of five categories: no depression, mild, moderate, severe, or unknown. To streamline this process, they assigned binary codes, either 1 or 0, to each chat entry based on the appropriate category. The categorisation was conducted over one month, after which each inter-rater submitted their coded data to the consultant for further analysis. This phase marked the second stage of the study, focusing on reliability analysis. By comparing the categorisations made by the different inter-raters, the study aimed to assess the consistency and agreement across the evaluations, which is critical for the overall validity of the findings

2.3 Data Analysis

Reliability checking was conducted to determine Inter-rater reliability (IRR), which refers to the consistency of subjective judgements made by two or more raters when assessing the same items (Tinsley & Weiss, 1975). IRR has been widely used in psychological research to observe behaviours (Geisinger et al., 2013) and to evaluate the consistency of clinical diagnoses (Lange, 2011). High IRR values closer to 1.0 indicate strong agreement among raters.

In this study, IRR was assessed using Spearman's rank-order correlation coefficient (ρ), as outlined by Barawi et al. (2017). This method measures the degree of association between two sets of ranked data. The coefficient was calculated using the following formula:

$$\rho = 1 - \frac{6\sum D^2}{n(n^2 - 1)}$$

Where *D* represents the difference between the two ranks for each observation, and *n* denotes the number of observations.

To calculate IRR, Spearman's ρ was first computed for each pair of raters based on their depression severity classifications. These ratings were then averaged across all five raters for each chat entry. Finally, the average ρ value was calculated to determine the overall level of agreement. A descriptive analysis was also conducted to examine further the consistency of the coded items among the raters.

3 RESULTS

3.1 Inter-Rater Reliability

Based on the mentioned formula for the IRR, the reliability between the five inter-raters for depression and other mental health issues ranged from moderate (ρ=0.6) to strong (ρ=0.9). The average value was ρ=0.72, indicating good interrater reliability.

3.2 Descriptive Analysis

Five mental health-related professionals closely inspected a total of a randomised sample of 4000 chat items. The online chat items were categorised based on the BDI and DSM-5 criteria for Major

Depressive Disorder, which were classified into five categories: no depression, mild, moderate, severe, and unknown.

Overall, the results of this study indicate that 84% (n=3, 355) of the agreed-coded items were usable. The high percentage of usable data signifies that three or more inter-raters similarly categorised the chat items. For instance, the chat item "*Every year I tell myself it is gonna get better. And every month I tell myself it is gonna be better. Every week I say I will be prepared. But every day I wish I were dead,*" was categorised as moderate depression by three inter-raters. Nevertheless, 16% (n=645) of the chat items were disputed because of disagreement, with fewer than three of the same categories coded by the inter-raters (see Table 1). An example of a disputed chat item was "*Aku rasa kosong kau tiada, kau tak rasa apa apa aku takde*" (I feel empty without you, you do not feel anything when I am not there), in which one of the inter-raters coded as no depression, two labelled as mild depression, and another two coded as unknown.

Table 1. Usable and disputed data.

Labels	Labelled items	Percentage (%)	Final Results
Mild	61	2%	
Moderate	90	2%	Usable data
Severe	130	3%	(n=3355, 84%)
No depression	756	19%	
Unknown	2318	58%	
No agreed label	645	16%	Disputed (n=645, 16%)

3.3 Prevalence of Depression and Other Mental Health Issues

The study classified chat items into various categories based on the presence and severity of depression. In the "No depression" category, the chat items reflected emotional disturbances like anxiety or stress triggered by specific events, but the individual was still functioning effectively in their daily life. These chats did not indicate symptoms of depression but highlighted other mental health challenges. The "Unknown" category, on the other hand, included chat items that did not fit into any of the established categories of depression, such as no depression, mild, moderate, or severe. These entries consisted of neutral statements, wishes, greetings, or conversations that lacked any clear indicators of mental health concerns, making them difficult to classify based on the depression-related criteria.

Table 2. Percentage of mild, moderate, and severe depression by the inter-raters.

Labels	Labelled items	Percentage (%)
Mild depression	61	2%
Moderate depression	90	2%
Severe depression	130	3%
Total		7%

Based on the usable data (n=3355), 7% of the chat items were classified as involving depression. Specifically, 2% of the items were categorised as mild depression (n=61) and moderate depression (n=90), while 3% (n=130) were classified as severe depression (refer to Table 2). Examples of chat items where there was a strong agreement between the inter-raters in identifying mild and severe depression are shown in Table 3, where the items were categorised similarly by multiple raters. Interestingly, 19% (n=756) of the chats were categorised as "No depression" but showed signs of other mental health issues such as anxiety or eating disorders (refer to Table 4). Examples of chat items with strong inter-rater agreement in labelling "No depression" and "Unknown," where three or more raters reached similar conclusions, are displayed in Table 5.

Table 3. Strong agreement between inter-raters in labelling the chat items.

Labels	Labelled items
Mild depression	Example 1: *I mengadu yg i selalu sakit kepala sgt one side...selalu nye belah kiri… kadang2 smpai ke mata…Doctor di KKIA bagi buat test depression, and resultnya mild depression*. (I complain that I always have one-sided headaches… it always blows to the left…sometimes it affects my eyes…the doctor at KKIA gave me a depression test, and the result was mild depression.)
Severe depression	Example 1: I did it, guys. I will now be going straight to the asylum. I spread pornographic material on my college chat group, and now they have reported me. I gave my best efforts to look as mentally unstable as possible and made it clear that I will not settle for less than rustication. Now I have to wait. They will probably call my parents and tell them I am mentally unstable (which I do not know they already think). Now I will be given anti-psychotics and sent to a mental care facility (I hope). Well, it was fun. Now I will be away from my phone for a while. Good for me, though, I seriously needed digital detox. So yes, this is my last post here. Needed an excuse good enough to kill myself, now I have got one. See You are in hell, guys!

Table 4. Percentage of "No Depression" and "Unknown" labels.

Labels	Labelled items	Percentage (%)
No depression	756	19%
Unknown	2,318	58%

Table 5. Examples of chat items for "No Depression" and "Unknown".

Label	Examples of chat items
No depression (but communicating other mental health issues)	Example 1: I guess in life, I just cannot always try to be nice to everyone and blame myself for when other people hurt me. Sometimes, it really is not my fault. Yeah, I agree that I always feel guilty and want to improve. However, to a certain extent, it has caused me stress. However, I will just learn not to be so hard on myself.
	Example 2: Colouring books- I am scared to stay between the lines, and then if the strokes and coverage are okay. Playing music- I do not play very well, and when you hear it, it is off-key. Exercise- it is okay, I am unable to maintain correct posture and stance. Movies/ content - I cry even at Adam Sandler's movies when he gets mistreated…which is stupid... I do not know how to manage anxiety.
	Example 3: Anxious after a panic attack
	Example 4: Anxiety from the job has me barely sleeping for over a week now.
	Example 5: My fiancé has been cheating for 4 weeks and now tells me.... I just want to end it all.
	Example 6: I am dealing with really serious stress from my anxiety, my OCD and paranoia. I cannot even function normally...I really need some guidance...
	Example 7: Worthlessness is when I cannot cope with everyone's expectations for my career. Office people say I can go further and can ace the yearly assessment. My husband also says so. The problem is that I cannot find time to study. During the day, I am occupied with work. At night, I am a mom and a wife. Moreover, I am the type who needs enough sleep. I will get cranky if I am sleep deprived.

Unknown	Example 1: My most recent experiences in the ER. (TW) Example 2: Today did 4 flights of stairs Example 3: hahahaha…I am controlling my food intake during CNY. Example 4: *Wow, bestnya, rock climbing di mana tu* (Wow, that is great! Where do you do rock climbing?) Example 5: Wow! This is good! Example 6: I really like the job that I have now. I am happy with my job. It takes up most of my time that I do not even think about my boyfriend when I work, haha. It somehow helps with my problem, which is that I miss my boyfriend constantly.

4 DISCUSSION

4.1 Inferential Analysis on Prevalence of Depression and Other Mental Health Issues

Based on Table 1 and Table 2, the result suggests that the raters were able to agree on the same items, indicating reliability in detecting mental health issues indicators. This result also suggests that online excerpts may contain clinically relevant signs that could support assessing one's mental well-being. In terms of practical values, the 7% result suggests that early identification of mental health issues in digital platforms is possible. The strong inter-rater agreement on the classification of depression, particularly across mild, moderate, and severe categories, suggests the reliability of the human evaluation process. This indicates that the raters were not only consistent with each other but were also able to identify subtle yet significant clinical nuances in the text. This is a crucial point, as it validates the use of human expertise as a benchmark for training and evaluating NLP models. Through the incorporation of human expertise in the evaluation, this result suggests that future NLP tools can be trained to go beyond detecting mental health issues and also to be able to determine the severity of the mental health issues. Thus, it offers an improvement for the current NLP tools.

In terms of other mental health issues, 58% (n=2318) of the chat items were labelled as "Unknown," indicating that users were neither exhibiting depression nor any other significant mental health concerns. This high result raises the issue of the complexity in analysing short texts. The ambiguity of the texts makes it difficult for the raters to categorise them explicitly. The high percentage also indicates the shared judgment between the raters on the data's inability to be classified. This result suggests that the chat excerpts contain very little information and context for them to be labelled, hence making it difficult for raters to categorise the severity of any mental health issues confidently. This finding is particularly significant for future research, as it highlights the need for more in-depth classification schemes that can move beyond simple and binary categorical labels. For NLP models, this vast "Unknown" class represents a major challenge, underscoring that while a model might become proficient at classifying an item that is not a mental health issue, it may still struggle with the subtle linguistic cues of more complex cases.

Overall, the findings also reveal that while a minority of online chats explicitly display depressive symptoms, a significant portion exhibits other mental health challenges. The good inter-rater reliability suggests that professionals can consistently identify these patterns. However, the "unknown" category highlights the difficulty of interpreting brief, context-limited messages. From the clinical point of view, the lack of context from the brief chat items does not permit a thorough interpretation of one's mental health condition. A mental health diagnosis requires a comprehensive assessment from human clinicians, which AI tools cannot possibly replicate in terms of expertise and skill. While the tools might help to screen for possible mental health conditions, there are still limitations to them. Future AI and NLP tools may improve in their classification, sensitivity, and linguistic markers; however, they should be constantly updated in accordance with the standards set by mental health professionals.

Most importantly, there are expressions such as non-verbal cues, cultural context, and tones that AI or NLP tools will never completely grasp. Ultimately, these technological advancements should not be regarded as substitutions but rather as complementary tools which aid in promoting a smoother and more accessible mental health service for the community. The results also suggest that psycholinguistic analysis of online communications can be a valuable tool for the early detection of mental health issues. However, the complexity of human language necessitates careful interpretation. Incorporating additional contextual information and leveraging professional judgment are essential for developing effective automated or semi-automated assessment tools in mental health. However, it is important to note that the inter-rater reliability score (ρ=0.72) among the mental health professionals, though it falls in the good range, can be improved. Inconsistencies among the mental health professionals could pose a risk of misclassification and implications for the users' mental well-being. The misclassification might lead to distress being unnoticed or vice versa. This disagreement can be improved through regular standardisation of operational definitions, training, and multidisciplinary inclusion in discussions. As aforementioned, contextual data alone is insufficient for a comprehensive evaluation, and although the IRR score is promising, the risk of misjudgement remains when relying solely on decontextualised text; hence, it is important to improve both the IRR score and the nuances of languages that the tool is able to detect.

4.2 Limitations of the Study

Despite the promising findings, this study has several limitations that must be acknowledged. First, the analysis was conducted on isolated chat items without access to the broader context of each conversation or the users' interaction histories. This lack of contextual information may lead to misinterpretations of the users' mental states, as brief messages can be ambiguous or insufficient for accurate assessment. Relying solely on single chat entries might overlook nuances that are critical for understanding the severity and nature of mental health issues. As mentioned earlier, this study emphasises the importance of seeking proper professional help, as the assessment conducted by the raters in this study cannot serve as a sole diagnosis; instead, it serves as a screening step to facilitate a comprehensive diagnosis. The steps taken in this study differ from the use of self-validated questionnaires such as the nine-item Patient Health Questionnaire (PHQ-9) (Kroenke et al., 2001), which is often accompanied by clinical interviews, and an in-depth look into one's history in order to make a comprehensive evaluation. On the contrary, the raters in this study were limited to classifying chat items that are isolated and decontextualised, with no access to the users' history, symptoms, and non-verbal cues. These limitations require raters to make judgments solely on the basis of restrictive information. The limitations of this method might produce judgments that could be less reliable, as exposure to misclassification is increased, and the result of the method cannot be used to substitute clinical diagnoses. These AI and NLP tools should be deemed as complementary tools rather than stand-alone for standard clinical assessment.

Second, the absence of demographic data such as age, gender, cultural background, and language proficiency limits the generalizability of the findings. Without this information, it is challenging to determine whether the sample is representative of the broader population or to account for cultural and linguistic nuances that influence how individuals express mental health concerns online. Language use and expressions of distress can vary significantly across different demographic groups, potentially affecting the accuracy of psycholinguistic analysis. Different populations express their feelings in varying ways. The extensive use of slang and emoticons in texts should also be considered. For example, individuals from Generation Z, Baby Boomers, and Millennials use emojis in different contexts and ways. Generation Z would use the emoji in sarcastic and ironic manners, such as the skull emoji, to impose something funny, while the Baby Boomers might interpret it as danger. Millennials, on the other hand, use emojis in a straightforward manner (Zahra & Ahmed, 2025). To capture these differences, future work should aim to include demographics, linguistic practices, and digital communication styles. Thus, it would improve the interpretation of the data through a holistic lens and a culturally informed frame.

Third, the study employed a manual categorisation process by five mental health professionals, which, while thorough, introduces subjectivity. Although an average inter-rater reliability of ρ=0.72 indicates good agreement, the use of Spearman's rank correlation coefficient may not be the most appropriate statistical measure for categorical data. Employing other statistical methods like Cohen's kappa or Krippendorff's alpha could provide a more robust assessment of inter-rater reliability for nominal scales, thereby strengthening the validity of the results. By addressing these limitations, subsequent studies can enhance the accuracy, ethical integrity, and applicability of psycholinguistic analyses in detecting mental health issues within online communities.

4.3 Future Work

To build upon the findings of this study and address its limitations, future research should focus on integrating more comprehensive contextual information into psycholinguistic analyses. Access to full conversation threads and users' interaction histories would provide deeper insights into language nuances, enabling more accurate assessments of mental health states. This holistic approach could capture the progression and severity of symptoms more effectively than isolated chat items.

Incorporating demographic data such as age, gender, cultural background, and language proficiency is essential for enhancing the generalizability of results. Understanding how different populations express mental health concerns online will allow for the development of culturally sensitive models that can accurately interpret diverse linguistic expressions. This inclusion would also help in identifying demographic-specific markers of mental health issues, contributing to more personalised interventions.

Advancements in natural language processing and machine learning offer promising avenues for future work. Developing automated systems that combine psycholinguistic features with sophisticated algorithms can enhance the scalability and efficiency of detecting mental health indicators in online communications. Implementing statistical methods better suited for categorical data, such as Cohen's kappa or Krippendorff's alpha, could improve the robustness of inter-rater reliability assessments in future studies. By addressing these areas, future work can significantly enhance the accuracy, ethical integrity, and applicability of psycholinguistic analyses in detecting and understanding mental health issues within online communities, ultimately contributing to improved mental health outcomes in the digital age.

4.4 Conclusion

The results emphasise the importance of thorough, context-aware assessments rather than swift diagnoses based solely on chat content. While users often exhibit a range of mental health challenges, relying solely on brief messages can overlook underlying issues. The overemphasis on isolated chat content might also risk misrepresentation of individuals' unique experiences. Hence, this study raises the issue that online tools should only be an addition, not a substitute, for individuals seeking thorough professional help. Mental health professionals must consider multiple factors to provide accurate and ethical evaluations, acknowledging that ambiguous or context-lacking messages can lead to misunderstandings. This study also highlights ethical concerns, including privacy and misclassification. As the study utilised chat items, the risk for misclassification and mislabelled items would pose a risk for mental health stigma to the users. Thus, it is crucial for the users' privacy to be protected and for their profiles to be confidential. Safeguarding anonymity not only protects individuals from potential harm but also upholds the ethical integrity of research in sensitive domains such as mental health. This research underscores the promise of combining expert analysis with advanced data mining techniques to improve mental health support in online communities.

ACKNOWLEDGEMENTS

We would like to thank Steffen Ruefer for his helpful and insightful feedback on earlier drafts of this manuscript. Thank you to the Faculty of Cognitive Sciences and Human Development

(FCSHD) and Research, Innovation & Enterprise Centre (RIEC, UNIMAS) for supporting and guiding the research and publication process.

AUTHOR CONTRIBUTIONS

Ross Azura Zahit was responsible for conceptualization, project administration, and data analysis and interpretation. Amalia Madihie contributed through funding acquisition as well as data analysis and interpretation. Salmah Mohamad Yusoff, Ida Juliana Hutasuhut, and Mohamad Azhari Abu Bakar were involved in writing the original draft and in data analysis and interpretation. Mohamad Hardyman Barawi contributed to writing, review, and editing, in addition to formal analysis. Syahrul Nizam Junaini and Nur Haziyah Amni Raimaini were responsible for writing, review, and editing. All authors reviewed and approved the final version of the manuscript.

CONFLICT OF INTEREST

The authors declare that there are no conflicts of interest related to this study.

DATA AVAILABILITY STATEMENT

Data is available upon reasonable request.

FUNDING

The research was financially supported by Naluri Hidup Sdn Bhd IG/F04/NALURI/2021. We would like to extend our sincere appreciation to Mr. Azhan Rani, the CEO of Naluri Hidup Sdn Bhd, for his generous funding and support of the aforementioned research project.

REFERENCES

Al-Mosaiwi, M., & Johnstone, T. (2018). In an absolute state: Elevated use of absolutist words is a marker specific to anxiety, depression, and suicidal ideation. *Clinical Psychological Science*, *6*(4), 529–542. https://doi.org/10.1177/2167702617747074

Amanat, A., Rizwan, M., Javed, A. R., Abdelhaq, M., Alsaqour, R., Pandya, S., & Uddin, M. (2022). Deep learning for depression detection from textual data. *Electronics*, *11*(5), 676. https://doi.org/10.3390/electronics11050676

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). https://doi.org/10.1176/appi.books.9780890425596

Amin, R., Schreynemackers, S., Oppenheimer, H., Petrovic, M., Hegerl, U., & Reich, H. (2025). Use of mobile sensing data for longitudinal monitoring and prediction of depression severity: Systematic review. *Journal of Medical Internet Research*, *27*, e57418. https://doi.org/10.2196/57418

Barawi, M. H., Lin, C., & Siddharthan, A. (2017). Automatically labelling sentiment-bearing topics with descriptive sentence labels. In F. Frasincar, A. Ittoo, L. Nguyen, & E. Métais (Eds.), *Lecture notes in computer science: Vol. 10260. Natural language processing and information systems* (pp. 478–484). Springer. https://doi.org/10.1007/978-3-319-59569-6_38

Beck, A. T., Ward, C. H., Mendelson, M., Mock, J., & Erbaugh, J. (1961). An inventory for measuring depression. *Archives of General Psychiatry*, *4*(6), 561–571. https://doi.org/10.1001/archpsyc.1961.01710120031004

Bijou, S. W., Umbreit, J., Ghezzi, P. M., & Chao, C.-C. (1986). Psychological linguistics: A natural science approach to the study of language interactions. *The Analysis of Verbal Behavior*, *4*, 23–29. https://doi.org/10.1007/bf03392812

Brockmeyer, T., Zimmermann, J., Kulessa, D., Hautzinger, M., Bents, H., Friederich, H.-C., Herzog, W., & Backenstrass, M. (2015). Me, myself, and I: Self-referent word use as an indicator of self-focused attention in relation to depression and anxiety. *Frontiers in Psychology*, *6,* 1564. https://doi.org/10.3389/fpsyg.2015.01564

Burkhardt, H. A., Alexopoulos, G. S., Pullmann, M. D., Hull, T. D., Areán, P. A., & Cohen, T. (2021). Behavioral activation and depression symptomatology: Longitudinal assessment of linguistic indicators in text-based therapy sessions. *Journal of Medical Internet Research*, *23*(7), e28244. https://doi.org/10.2196/28244

Cacheda, F., Fernandez, D., Novoa, F. J., & Carneiro, V. (2019). Early detection of depression: Social network analysis and random forest techniques. *Journal of Medical Internet Research*, *21*(6), e12554. https://doi.org/10.2196/12554

Cai, Y., Wang, H., Ye, H., Jin, Y., & Gao, W. (2023). Depression detection on online social network with multivariate time series feature of user depressive symptoms. *Expert Systems with Applications*, *217*, 119538. https://doi.org/10.1016/j.eswa.2023.119538

Chancellor, S., & De Choudhury, M. (2020). Methods in predictive techniques for mental health status on social media: A critical review. *NPJ Digital Medicine*, *3*, 43. https://doi.org/10.1038/s41746-020-0233-7

Chiong, R., Budhi, G. S., Dhakal, S., & Chiong, F. (2021). A textual-based featuring approach for depression detection using machine learning classifiers and social media texts. *Computers in Biology and Medicine*, *135*, 104499. https://doi.org/10.1016/j.compbiomed.2021.104499

DeSouza, D. D., Robin, J., Gumus, M., & Yeung, A. (2021). Natural language processing as an emerging tool to detect late-life depression. *Frontiers in Psychiatry*, *12*. https://doi.org/10.3389/fpsyt.2021.719125

Dewangan, D., Selot, S., & Panicker, S. (2022). Implementation of machine learning techniques for depression in text messages: A survey. *I-Manager's Journal on Computer Science*, *9*(4), 13–20. https://doi.org/10.26634/jcom.9.4.18549

Geisinger, K. F., Bracken, B. A., Carlson, J. F., Hansen, J.-I. C., Kuncel, N. R., Reise, S. P., & Rodriguez, M. C. (2013). *APA handbook of testing and assessment in psychology, Vol. 3: Testing and assessment in school psychology and education.* American Psychological Association. https://doi.org/10.1037/14049-000

Han, K., Rezapour, R., Nakamura, K., Devkota, D., Miller, D. C., & Diesner, J. (2023). An expert-in-the-loop method for domain-specific document categorization based on small training data. *Journal of the Association for Information Science and Technology*, *74*(6), 669–684. https://doi.org/10.1002/asi.24714

Huang, X. (2022). Ideal construction of chatbot based on intelligent depression detection techniques. *IEEE International Conference on Electrical Engineering, Big Data and Algorithms, China,* 511–515. https://doi.org/10.1109/EEBDA53927.2022.9744938

Jackson, J. C., Watts, J., List, J. -M., Puryear, C., Drabble, R., & Lindquist, K. A. (2022). From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, *17*(3), 805–826. https://doi.org/10.1177/17456916211004899

Korhonen, J., Axelin, A., Katajisto, J., Lahti, M., & MEGA Consortium/Research Team. (2022). Construct validity and internal consistency of the revised Mental Health Literacy Scale in South African and Zambian contexts. *Nursing Open*, *9*(2), 966–977. https://doi.org/10.1002/nop2.1132

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

Lange, R. T. (2011). Inter-rater reliability. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (pp. 1348). Springer. https://doi.org/10.1007/978-0-387-79948-3_1203

Li, I., Li, Y., Li, T., Alvarez-Napagao, S., Garcia-Gasulla, D., & Suzumura, T. (2020). What are we depressed about when we talk about COVID-19: Mental health analysis on tweets using natural language processing. In M. Bramer & R. Ellis (Eds.), *Lecture notes in computer science: Vol. 12498. Proceedings of the 2020 SGAI International Conference* (pp. 358–370). Springer. https://doi.org/10.1007/978-3-030-63799-6_27

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), e22635. https://doi.org/10.2196/22635

Mehta, Y., Fatehi, S., Kazameini, A., Stachl, C., Cambria, E., & Eetemadi, S. (2020). Bottom-up and top-down: Predicting personality with psycholinguistic and language model features. *IEEE International Conference on Data Mining, Italy,* 1184–1189. https://doi.org/10.1109/ICDM50108.2020.00146

Ostic, D., Qalati, S. A., Barbosa, B., Shah, S. M. M., Galvan Vela, E., Herzallah, A. M., & Liu, F. (2021). Effects of social media use on psychological well-being: A mediated model. *Frontiers in Psychology*, *12,* 678766. https://doi.org/10.3389/fpsyg.2021.678766

Oudin, A., Maatoug, R., Bourla, A., Ferreri, F., Bonnot, O., Millet, B., Schoeller, F., Mouchabac, S., & Adrien, V. (2023). Digital phenotyping: Data-driven psychiatry to redefine mental health. *Journal of Medical Internet Research*, *25*, e44502. https://doi.org/10.2196/44502

Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, *54*, 547–577. https://doi.org/10.1146/annurev.psych.54.101601.145041

Ren, X., Burkhardt, H. A., Areán, P. A., Hull, T. D., & Cohen, T. (2023). Deep representations of first-person pronouns for prediction of depression symptom severity. *Annual Symposium Proceedings Archive*, 1226–1235.

Solmi, M., Radua, J., Olivola, M., Croce, E., Soardo, L., Salazar de Pablo, G., Il Shin, J., Kirkbride, J. B., Jones, P., Kim, J. H., Kim, J. Y., Carvalho, A. F., Seeman, M. V., Correll, C. U., & Fusar-Poli, P. (2022). Age at onset of mental disorders worldwide: Large-scale meta-analysis of 192 epidemiological studies. *Molecular Psychiatry*, *27*, 281–295. https://doi.org/10.1038/s41380-021-01161-7

Tinsley, H. E., & Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, *22*(4), 358–376. https://doi.org/10.1037/h0076640

Zahra, T., & Ahmed, S. (2025). Generational differences in emoji interpretation: A study of millennial, Gen Z, and baby boomers. *Advance Social Science Archive Journal*, *3*(2), 857–864.

Ziemer, K. S., & Korkmaz, G. (2017). Using text to predict psychological and physical health: A comparison of human raters and computerised text analysis. *Computers in Human Behavior*, *76*, 122–127. https://doi.org/10.1016/j.chb.2017.06.038