An empirical content analysis of posts made to the r/depression subreddit

Marais, Kurt
Stellenbosch University
maraisk@sun.ac.za

Abstract

Online fora are one of the oldest spaces for internet communication. These fora provide internet users the opportunity to connect with others of similar interests and experiences. Online mental health communities are fora for individuals that wish to connect with others based on a shared mental health experience. These digital ecosystems offer an avenue for seeking advice or validation, as well as for freely sharing experiences with similar others within a culture of reduced stigma and social support.

Reddit is a popular and readily accessible platform that hosts moderated communities known as subreddits, which include communities for mental health concerns such as depression, anxiety, trauma and eating disorders. This study is an analysis of data collected from the r/depression subreddit by way of the Reddit application programming interface. A conceptual framework was developed for preparing and evaluating the data by way of an empirical content analysis. The content analysis was conducted on 11 975 posts generated from 9 593 unique user accounts over a 56-day period.

The analysis of user behaviour in this subreddit relates to general engagement, the temporal attributes of the dataset, and the demographic and geographic attributes of the users. This investigation provides insights into the words most frequently used in posts made to this subreddit, the time of day and day of the week that most individuals posted to the forum, the age and locational distribution of r/depression subreddit users accounted for in the dataset and how these insights relate to the global prevalence of depression.

Keywords: depression, mental health, social media, online mental health communities, Reddit

1 Introduction

Reddit is a community-based and largely peer-driven social media platform where users join online spaces known as subreddits that revolve around a topic of interest. These subreddits act as communities of practice and discourse, as information surrounding the common topic is exchanged, interacted with or debated in posts and comment threads where anyone can contribute. A Reddit post consists of a title that may contain up to 300 characters and the body of a post with a limit of 40 000 characters. Anyone is able to interact with a post by commenting or voting it up or down on the subreddit, given that the capability to interact with a Reddit user's content has been made available. Upvotes and downvotes are a way of signifying to members of the subreddit or members just viewing the subreddit that a post or comment is relevant and contributes to the subreddit positively. Any Reddit user may add a comment of at most 10 000 characters under a post, and posts may include additional subreddit-specific tags known as flairs.

Reddit is free to use, with the option to gain premium access at a monthly or annual subscription fee. Reddit Premium enables users to gain access to exclusive content and browse the platform without advertisements. There is also Reddit-exclusive currency called coins that may be used to purchase customisations, avatar gear and awards that may be gifted to other users' posts. Reddit users are incentivised with achievements and "karma" points, which is a system that reflects user's contributions on the platform through posting, commenting, giving awards and having their posts be interacted with.

Reddit users may decide how anonymous they would like to be by generating a profile and an avatar in the form of the Reddit mascot named Snoo. Reddit identifies its users by the prefix "*u/*" followed by their username of choice, and subreddits are identified by the prefix "*r/*" followed by the name of the subreddit, in the same way that an X (the platform

formerly known as Twitter) user is identified by a commercial 'at' symbol (@) and a topic known as a hashtag is identified by an octothorpe (#).

1.1 Online mental health communities

Subreddits may represent broad topics such as sports, gaming, crypto, television and business. Other subreddits may be more niche topics that may or may not form part of the discourse of the broader subreddits, like the Premier League, Minecraft, Bitcoin, The Amazing Race and Tesla. Most subreddits are user-generated and may be as ambiguous or as specific as is deemed appropriate. The subreddit communities exist independent of one another; each subreddit has its own users, rules and content that may overlap with another subreddit. There are more than 100 thousand active subreddits on the platform, and Reddit users have been diligent in posting the appropriate content to the appropriate subreddit. Upvoting and downvoting posts allows users to inform what is relevant and important for a subreddit or not, which further encourages and incentivises users to post relevant content. However, the platform is not immune to the presence of unrelated and "clickbait" content.

All online spaces are always subject to antisocial behaviours in the form of spam, trolling and harassment. This includes community-based platforms that are considered to be safe havens for avid members or individuals seeking a sense of community and social support. There are often rules of engagement proposed within each space relative to the nature of the subreddits. Fora are typically monitored by moderators that ensure that the content shared adheres to the subreddit guidelines, that it is credible and does not promote harmful behaviour (Saha et al. 2020).

In the case of online mental health communities (OMHCs), strict moderation is required to protect the well-being of members of those communities. These spaces allow for individuals to disclose experiences that may be sensitive or stigmatising, as well as to facilitate peer-to-peer support. Moderation within OMHCs requires that support and privacy for disclosure is upheld, with Saha et al. (2020) raising the need for counsellors and psychiatrists to be present in these communities to ensure the quality and credibility of the content being shared.

1.2 Community guidelines for r/depression

It is imperative to the integrity of any analytical research that the context of the environment from where the text originates is disclosed. This study considers the OMHC that is the r/depression subreddit, which has amassed more than 1.04 million subscribers as of August 2024. The subreddit was established on 1 January 2009 and in the subreddit description it is stated that the community is strictly focused, as depression is difficult to talk about. Individuals posting to the subreddit do not need to be formally diagnosed with depression, but it is further stated in the description that individuals who do post must request support. The rules of the r/depression subreddit, as compiled by the moderators of this space, are paraphrased as follows:

1. Individuals who choose to post must request support for depressive disorders relating to themselves or for someone close to them.
2. Replies to any posts must show empathy towards the original poster (OP). There should not be any attempt to display sentiments of "tough love" or to encourage debate.
3. Anyone who would like to offer help or support should do so publicly in the comments and not through private messages or through the chat function.
4. Posts and responses must demonstrate understanding or an attempt to understand and must include sufficient information to provide meaningful support.
5. While stories of personal success may be inspiring to some, others in the subreddit may find

such content painful and demoralising. Users are prohibited from posting content that puts themselves above others.

6. Users are not allowed to share patronising content of any kind, which includes messages following the sentiment of "it gets better" that would encourage an unhealthy outcome-based mindset.
7. Users may not give or request diagnoses or any clinical advice, nor advocate for or against treatments and self-help strategies. Other subreddits relating to this have been recommended, should that be a space that someone needs.
8. Posts sharing information about methods for inciting any forms of self-harm or validating self-destructive intent are strongly prohibited. Other subreddits with information specifically relating to self-harm have been hyperlinked.
9. Community members should not encourage rule-breaking or posting of unrelated content.
10. There should be no forms of activism, debate or unapproved surveys, which includes informal questions and polls. It is emphasised that the purpose of this particular subreddit is to provide a support space.
11. There should be no self-promotion of any kind, including works of creative writing or any other art.
12. No one is allowed to request or offer money, goods, or services.

Members of the subreddit are encouraged to report any behaviour that transgresses the rules that have been set in place to maintain the emotional and physical safety of other members of the subreddit. As of August 2024, there were eleven moderators that manage the subreddit and update the page and additional sources [1] regularly. Moderators would typically be community members of that subreddit and serve on a voluntary basis without compensation (Matias 2019). This may be considered beneficial to the subreddit, since the moderators are of the community and understand the complexities of interactions on the OMHC, but may also be detrimental to their own psychological distress and are at risk of burnout resulting from engaging with triggering or harmful content and users (Schöpke-Gonzalez et al. 2024).

2 Related work

Data generated in Reddit OMHCs have resulted in research primarily concerned with the detection and prediction of health issues (Proferes et al. 2021). Balani & De Choudhury (2015) used data generated on Reddit for the detection of mental health self-disclosure levels. Self-disclosure is said to be an important social behaviour that increases social support (Bak et al. 2014), and this data can prove very useful for examining health-related discourse online. Aladağ et al. (2018) used publicly available Reddit post data to apply sentiment analysis and other tools measuring linguistic characteristics to detect suicidal ideation in Reddit users' posts. De Choudhury et al. (2016) also investigated Reddit data in the detection of suicide ideation, and provided a statistical methodology to derive distinct markers which identify and predict users that are likely to engage in suicidal ideation.

There exists the stigma that social media use is only a contributing factor to poor mental health and symptoms of depression. Cunningham et al. (2021) conducted a meta-analytic study on social network site (SNS) use and depression symptoms, with a focus on time spent using SNSs, the extent of use and the problematic use of SNSs. They recommended that research considering problematic SNS use as the primary construct driving this relation should focus on the ways in which individuals engage in patterns of problematic use of social media. A study by Seabrook et al. (2016) found that there is a correlation between SNS use and mental illness and well-being, and that, whether this relationship is positive or negative, it is dependent on the quality of social factors within the SNS environment. In other words, positive interactions and supportive social cohesion related to lower levels of anxiety and de-

pression, whereas negative experiences and social comparisons (as opposed to social connectedness) related to higher levels of depression and anxiety. The study concluded that a greater understanding of these relationships can unlock the potential of SNSs to positively influence mental health. These contrasting perspectives converge to the same conclusion; that there is an opportunity to leverage insights from engagement on SNSs to promote mechanisms for improving mental well-being.

The role of SNSs, and particularly OMHC, is not to replace the role of treatment that would be facilitated by a mental health practitioner, but may provide therapeutic support in moments of crisis. Moreover, engaging with similar others helps reduce feelings of loneliness, encourages individuals to be more informed, and may lead to self-efficacy of individuals in seeking professional support (Prescott et al. 2020). Suler (2004) posits that online fora offer less inhibited spaces for mental health-related disclosure and social support by affording users a chance at anonymity, invisibility and asynchronicity. Saha et al. (2020) proffers that effective support on OMHCs manifests as complex language factors, such as verbosity, diversity, expressions of positive affect, adaptability in responses, a dynamic writing style, and both emotional and informational support. Chen & Xu (2021) found that users that experience social support and empathy in OMHCs tend to express support and empathy to others, and encourages users to post more.

3 Data collection

Data from the r/depression subreddit were obtained through the Reddit application programming interface (API) (Reddit Inc. 2005). An application with unique client credentials was registered within Reddit's development framework and OAuth 2.0 authentication protocols were employed to ensure secure access to the API. Reddit's API infrastructure was used in the scripting for accessing data from the subreddit. This was done to ensure that data extraction aligned with the platform's data access policies and API governance framework. This approach is also resilient to changes made to access via third-party services such as Pushshift, where access via these retrieval systems may change subject to Reddit's terms of service and API usage guidelines. Accessing the API using Reddit's API infrastructure does mean that limited data can be extracted per request. For this study, this meant that the most recent posts made to the subreddit were extracted once every few days.

The data collection process began on 16 April 2023 and ended on 11 June 2023. Data were scraped periodically within this time period and a total of 12 133 unique posts were obtained over the 56 days. The statistics of the Reddit data used in this study are summarised in Table 1. The conceptual framework for data evaluation and content analysis (Drisko & Maschi 2016) is represented in Figure 1.

Data were collected every few days since posts made

Table 1: A summary of the statistics of the Reddit posts collected for this study.

Final number of Reddit posts	11 975.00
Original number of posts collected	12 133.00
Posts from deleted accounts removed	80.00
Non-English posts removed	78.00
Number of unique user accounts that posted during the data collection period	9 593.00
Mean number of posts made per day	213.84
Number of unique words contained in the corpus	12 857.00
Mean word length per post	178.07
Maximum word length of a post	5 192.00
Maximum number of upvotes for a post	1 340.00
Maximum number of comments under a post	188.00

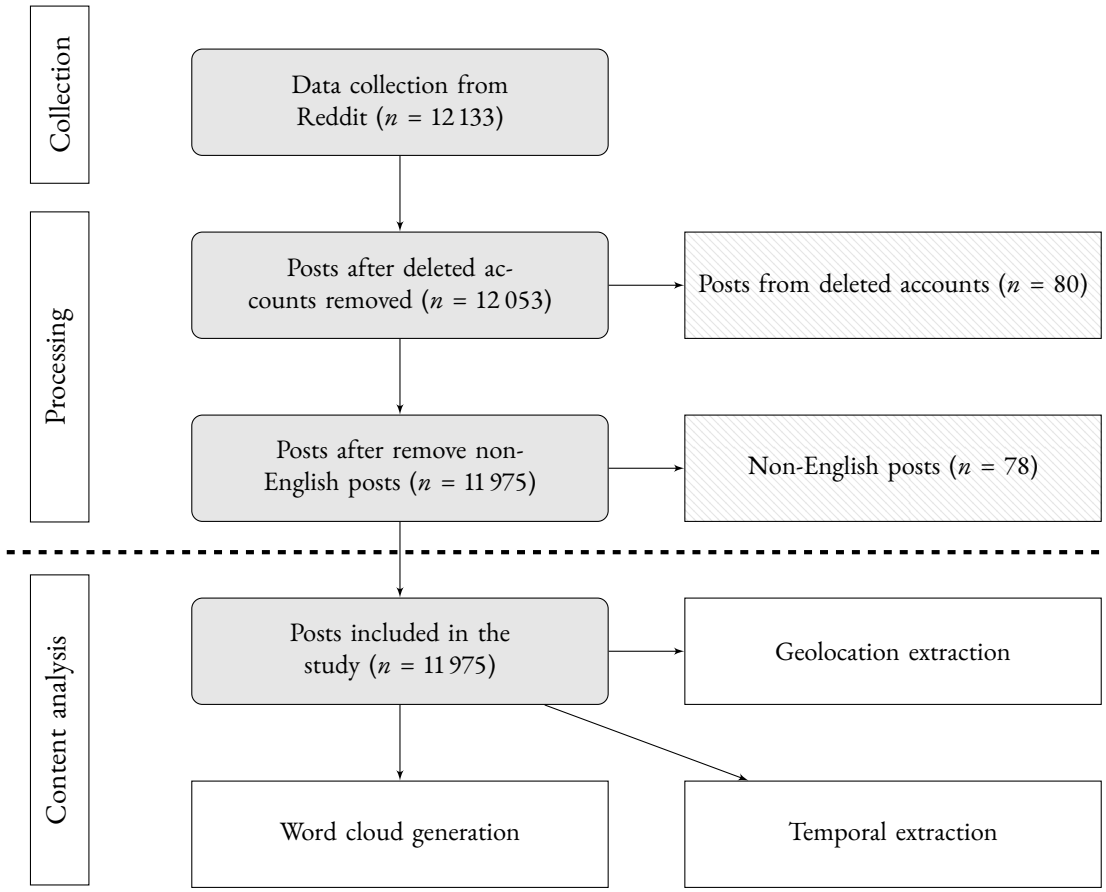

Figure 1: A conceptual framework of the content analysis of the r/depression subreddit.

to a subreddit were not made as frequently as posts generated on non-community-based social media platforms. Additionally, there is a limited rate of 1 000 data entries per request that may be retrieved from a particular subreddit via the Reddit API. This limit is imposed to assist with server load, data storage and improved user experience. No comments or comment threads on posts were collected during the collection period.

There were 80 accounts that were deleted by users during the data collection process. The posts that were made from the later deleted accounts were subsequently removed from the dataset to ensure the integrity of each user's privacy was maintained. Thereafter, a further 78 posts were identified as non-English posts by way of the *'langdetect'* automated language detector in Python. The languages of the posts that were removed along with the ISO 639-1 codes are summarised in Table 2. It was assumed that all posts that were identified as English posts were written in English. The language detector was not accurate at identifying the languages of the non-English posts, especially in the case where posts were too short or when special characters or emoticons were used as text. The language detector would classify URL links or symbols it was unfamiliar with into an "unknown" category. Thus, the non-English posts that were removed were manually evaluated as to whether they should or should not be included in the final dataset. The final dataset contained 11 975 posts from 9 593 unique accounts.

4 Data analysis

Further analysis of the Reddit data was conducted to investigate the ways in which the subreddit community members that were captured during data collection use the platform for sharing their expe-

Table 2: A summary of the non-English languages posts removed from the original dataset.

Language	ISO 639-1 code	Posts removed
Catalan	ca	3
Danish	da	3
German	de	3
Spanish	es	24
Estonian	et	3
French	fr	1
Indonesian	id	1
Korean	ko	1
Portuguese	pt	6
Romanian, Moldavian, Moldovan	ro	1
Somali	so	3
Swedish	sv	5
Turkish	tr	3
Unknown	–	20
Vietnamese	vi	1

riences of depression and whether there are underlying trends in the behaviour of these users in this community when posting to Reddit. The analysis focuses on attributes of general engagement, temporal, demographic and geographic attributes observed in the corpus.

4.1 General engagement

The body of Reddit posts from the r/depression subreddit were used as the data for the investigations conducted in this study. However, it is important to state that Reddit users in the dataset often used the title of the post for the main messaging. Individuals may opt to type posts with concise titles rather than more substantial messages in the body of the post. It was often the case that users would just refer anyone viewing their post to the title to satisfy the requirement of having text in the title and in the main text. There is a degree of information that is lost due to only considering the main text of Reddit posts, but it is assumed that the information that exists in the character limit of a post title is not as rich as the text that exists within the character limit of the main post text.

The data collected contained no duplicates, which is defined as the same post generated by the same user at the same time. However, there were multiple posts made at different timestamps that contained identical information in either the title or the main post text and were posted in close succession within the span of a few minutes. In some instances, these anomalous posts were posted from the same account with minor alterations either to the original or from the previous post. This allows the OP to gain more engagement across multiple posts, with the hopes that other individuals in this OMHC may reach out to them and offer support. In other instances, the dataset contained posts with identical text made from two or more separate accounts. It is assumed that these identical posts were made by the same user, and may have felt comfortable seeking help through increased anonymity by posting from multiple accounts. Only 13.25% of the 9 593 user accounts that posted during the data collection period posted more than once on the subreddit, with one Reddit account posting to the r/depression subreddit a maximum of 41 times during the 56-day data collection period.

The corpus of posts contained 12 857 unique words. Figure 2 shows the 200 most frequently occurring words out of the corpus of posts, excluding commonly occurring stop words. Posts from the dataset extracted from this subreddit contained an average of 178.07 (σ = 195.84) words per post, and the longest post consisted of 5 192 words.

The most frequently used words in the cor-

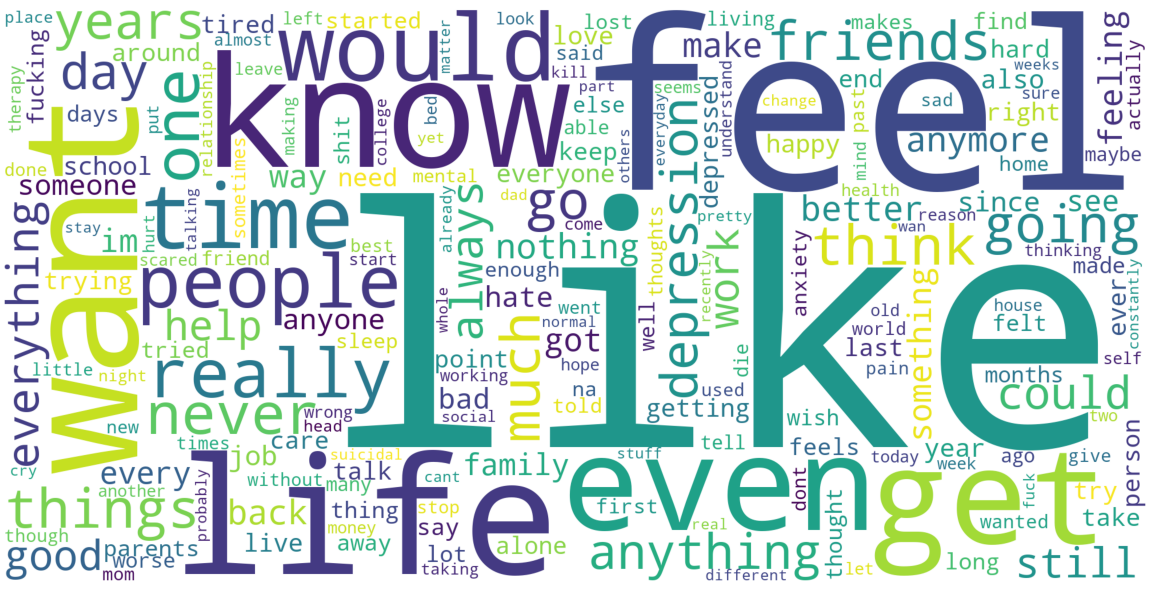

Figure 2: A word cloud of the 200 most frequently occurring words from all of the r/depression dataset.

pus (excluding stop words) were "like" (16 078), "feel" (14 513), "life" (10 044) and "want" (9 884). Posts on this subreddit often centred around how the OP feels and the desires for something other than feelings associated with depression. This is reflected in the frequent use of first-person stop words such as "I" (173 581), "am" (31 611), "my" (46 587) and "me" (29 335). The text has not been stemmed for this analysis, and hence words such as "feeling" (3 140), "depression" (4 367) and "thinking" (1 260) appear in the most frequently used word list alongside "feel" (14 513), "depressed" (2 421) and "think" (4 735), respectively.

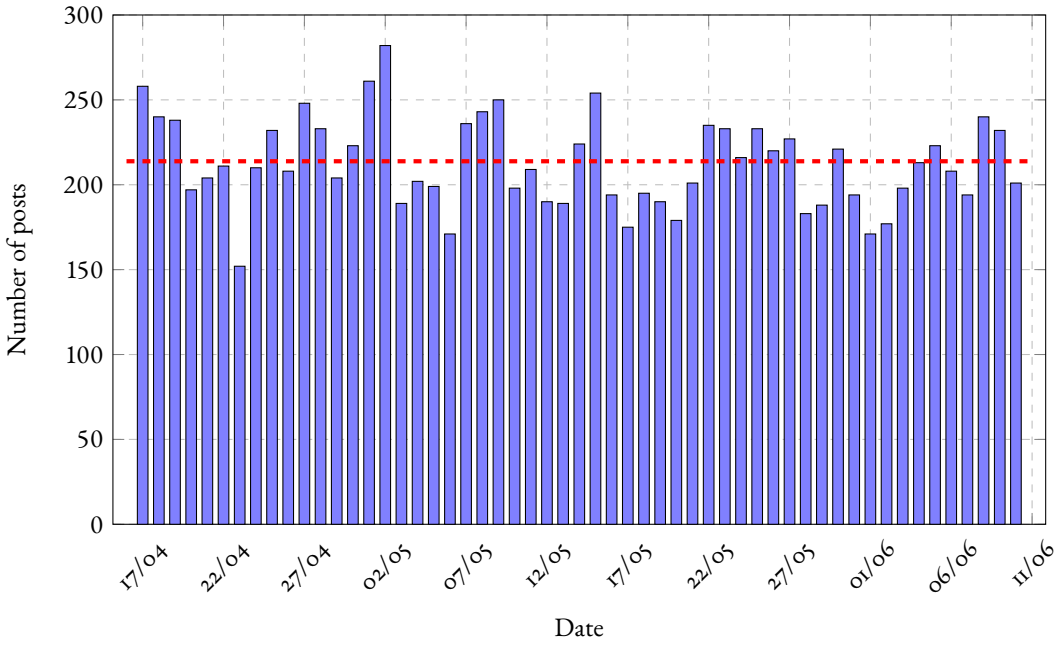

Figure 3: The number of posts made each day on the r/depression subreddit.

4.2 Temporal attributes

Figure 3 is a bar chart of the number of posts made to the subreddit of each day of data collection, excluding the first and last days. The days excluded did not include all posts made within the 24-hour cycle of that day.

All times and dates are relative to South African Standard Time (SAST), which is two hours ahead of Coordinated Universal Time (UCT+02:00). The dotted red line represents the mean number of posts per day over the 56-day data collection period. The mean number of daily posts made to the subreddit during this period was 213.84 posts per day. The day that incurred the most posts of the days observed was 2 May 2023 with 282 posts. There is no prominent consensual theme of the posts made on this day, but this date is significant to most countries in the world that observe International Labour Day or Workers' Day. Countries in different time zones would have observed a public holiday around this time and may explain the reason for the increased number of posts made on this day.

The total number of posts from the considered dataset made on different days of the week are plotted in Figure 4. Increased posting activity on the r/depression subreddit were seen on Mondays and Tuesdays. Saturdays were the days that fewer posts were made to the community.

Engagement with the posts on Reddit take place asynchronously and over time, as is the nature of online fora (Hunt & Brookes 2020). Once a user has generated their post, users are able to engage with the content for as long as it is visible on the forum page. The forms of engagement with Reddit posts are through upvotes and comments. Downvotes are also recorded by Reddit but not explicitly publicly available to Reddit users. The post with the most upvotes obtained 1 340 upvotes and the post with the most comments generated 188 comments.

Figure 5 is a bar chart of the total upvotes and total comments that posts made on particular days of the week have accumulated. Posts made on Mondays received more engagement in the form of upvotes, and posts made on Saturdays received more engagement in the form of comments, relative to posts made on other days of the week. Posts within OMHCs do not necessarily strive for increased engagement for traditional social acclaim, since these spaces generally serve as a platform for support and validation.

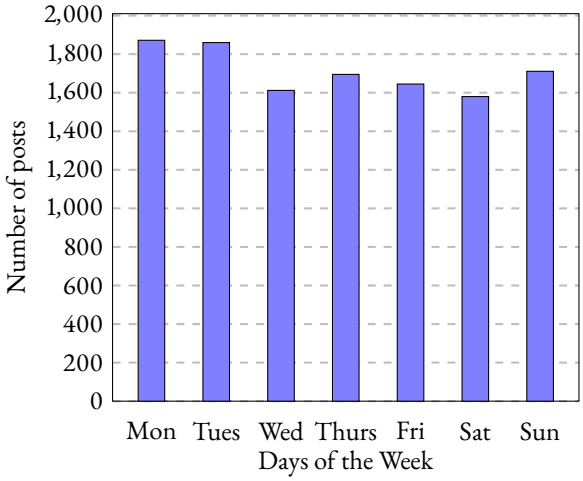

Figure 4: A bar chart of the total number of posts made to r/depression on different days of the week.

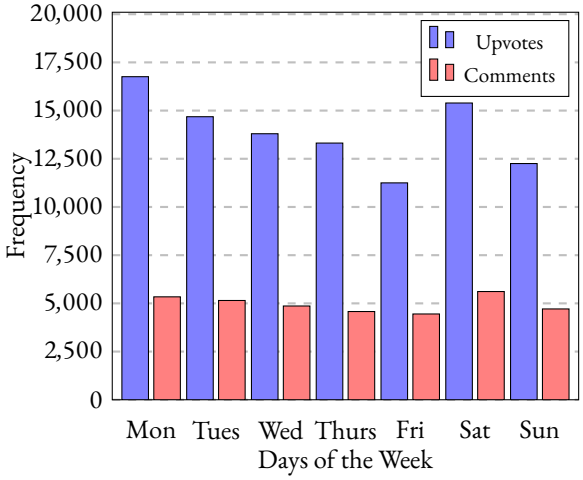

Figure 5: A bar chart of the total engagement on posts made to r/depression on different days of the week.

The time of day of when posts are made is a significant factor to consider in the case of a depression-centred community. Figure 6 is a radar plot of the

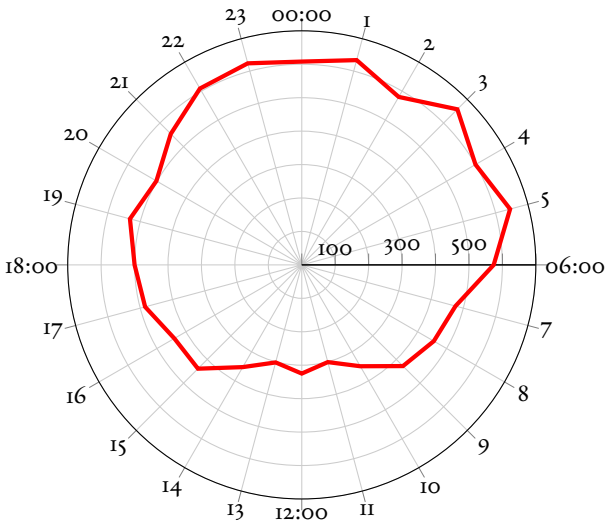

Figure 6: A radar plot depicting the frequency of posts made to r/depression relative to the time of day in South African Standard Time.

total number of posts made over the duration of the data collection period relative to the hour of the day it was posted in SAST. It is noted that there is a sustained increase in the number of posts made from 10 PM until 5 AM. De Choudhury et al. (2013) have shown that people living with depression have increased social media activity later at night. Becker & Lienesch (2018), as well as Woods & Scott (2016), found an association between nighttime social media use and depressive symptoms.

The appeal of an online forum such as r/depression is the concept of asynchronicity, where interactions on the forum do not require the simultaneous participation of all of the members (Hunt & Brookes 2020). Suler (2004) explained that asynchronicity provides flexible temporal demands, such that more time and thought can be dedicated to crafting contributions to the forum. Additionally, users sharing information on the subreddit may gain a sense of comfort in having the option to disclose as much or as little information as they are comfortable with. Thus, there is no time that is universally opportune for all users of the subreddit.

The frequency of posts relative to the time of day is subject to bias, since the posts would have been made at different times of the day for different time zones. The hours of the day with the highest and the lowest frequency of posting activity relative to different time zones across the world are indicated in Figure 7 and Figure 8, respectively.

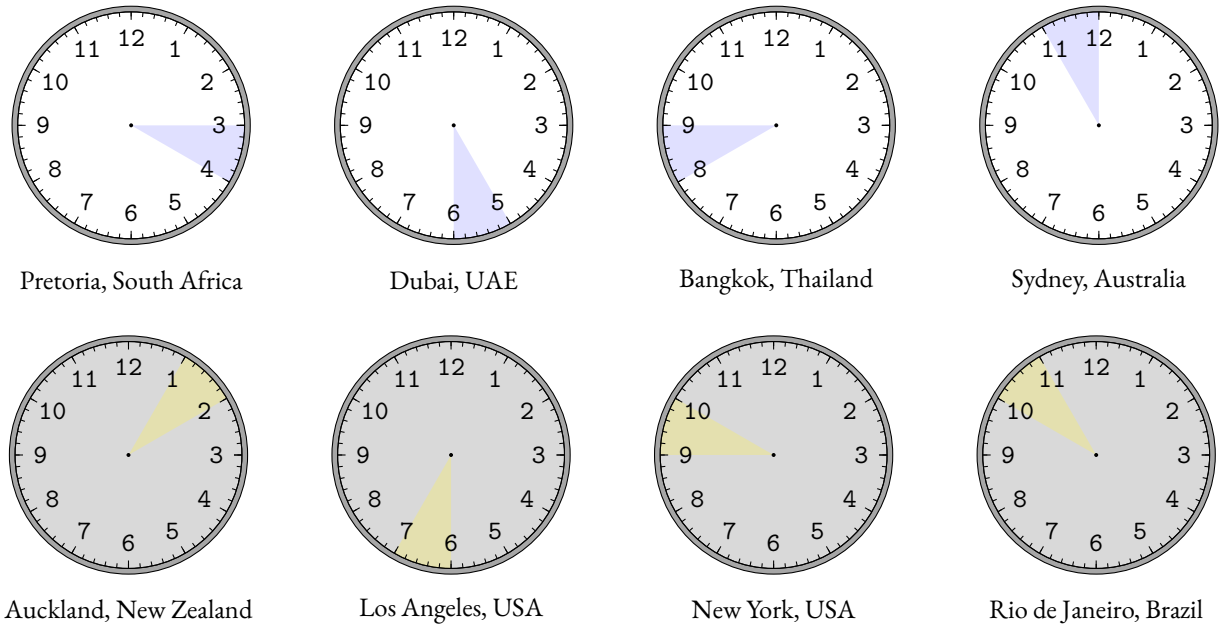

Figure 7: World clocks depicting the hour of highest posting frequency to the r/depression subreddit. The white clock faces depict the ante meridiem (AM) time and the grey clock faces represent post meridiem (PM) time.

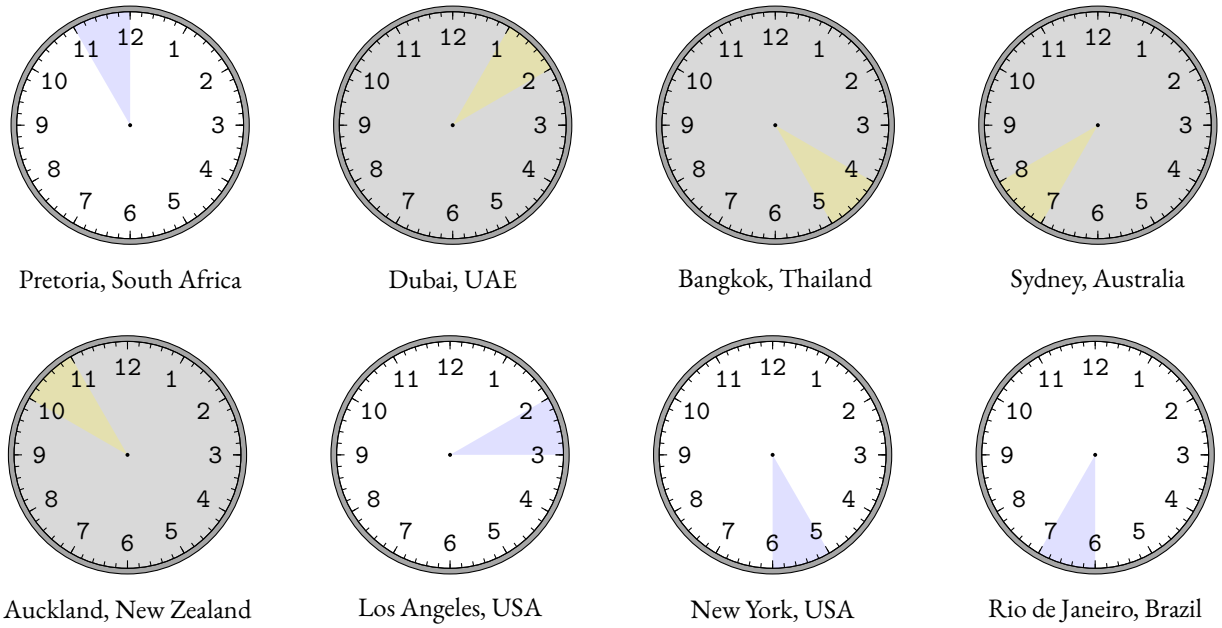

Figure 8: World clocks depicting the hour of lowest posting frequency to the r/depression subreddit. The white clock faces depict the ante meridiem (AM) time and the grey clock faces represent post meridiem (PM) time.

The busiest time for posting to the r/depression subreddit from the collected dataset took place between 3 AM and 4 AM SAST, followed closely by between 5 AM and 6 AM SAST. These times correspond to nighttime activity for individuals from the United States of America (USA) who form the majority of Reddit users and the majority of this subreddit. The activity would thus be busiest between 6 PM and 7 PM Los Angeles time (UTC−07:00) and between 9 PM and 10 PM New York time (UTC−04:00). The times with the lowest degree of posting activity on the subreddit was between 11 AM and 12 PM SAST, which corresponds to 2–3 AM Los Angeles time and 5–6 AM New York time. The second lowest degree of posting activity to the subreddit took place between 1 PM and 2 PM SAST.

4.3 Demographic and geographic attributes

Demographic and geographic location are important attributes to consider in the analysis of Reddit user behaviour. The Reddit user base is predominantly male users (64.9%) and users tend to use Reddit largely to look for entertaining content (DataReportal 2023). Figure 9 contains a pie chart of the percentage of all active Reddit users aged 16 to 64 that

Figure 9: The proportion of activities that active Reddit users aged 16 to 64 prefer engaging in on the platform.

specify using the platform for each kind of activity based on information collected by DataReportal (2023).

The ages of the r/depression subreddit members that posted during the data collection period was acquired through the extraction of the ages disclosed in the Reddit posts. There is no convention for how or why individuals decide to disclose their age and is often shared as an additional layer of context to the information provided in the post. Reddit is a platform for younger users, with the posts considered in this study made by individuals as young as 10 years and as old as 70 years. The age distribution of the 1 707 individuals that disclosed their ages in their posts is plotted in Figure 10.

The Reddit API does not allow access to geolocation data of where the post originated. Inferences about the location of users were made based on the post text and any mention of geography. Geopolitical entities and locations were extracted from each Reddit post in the dataset to identify which regions were mentioned and were further evaluated manually to determine which country each post may have originated from. The locations that were mentioned were tallied and are represented in Figure 11 as the shaded regions.

Not all Reddit posts in the dataset contained information regarding cities, countries or continents. Some countries may not be represented in the shaded areas of the map. This is a result of the removal of non-English languages spoken in these regions having been removed during the preprocessing of the corpus data or some posts that would refer to wider regions like Europe or the Middle East as opposed to a specific country. Reddit is also more popular in certain countries than in others, which may inform bias as to the users that are on the platform. Figure 12 contains a diagram of the rankings of countries with the highest prevalence of depression in 2023 (World Population Review 2023) and the largest global traffic to Reddit (Semrush 2023) compared to the rankings of the countries that were referenced most in the extracted Reddit posts.

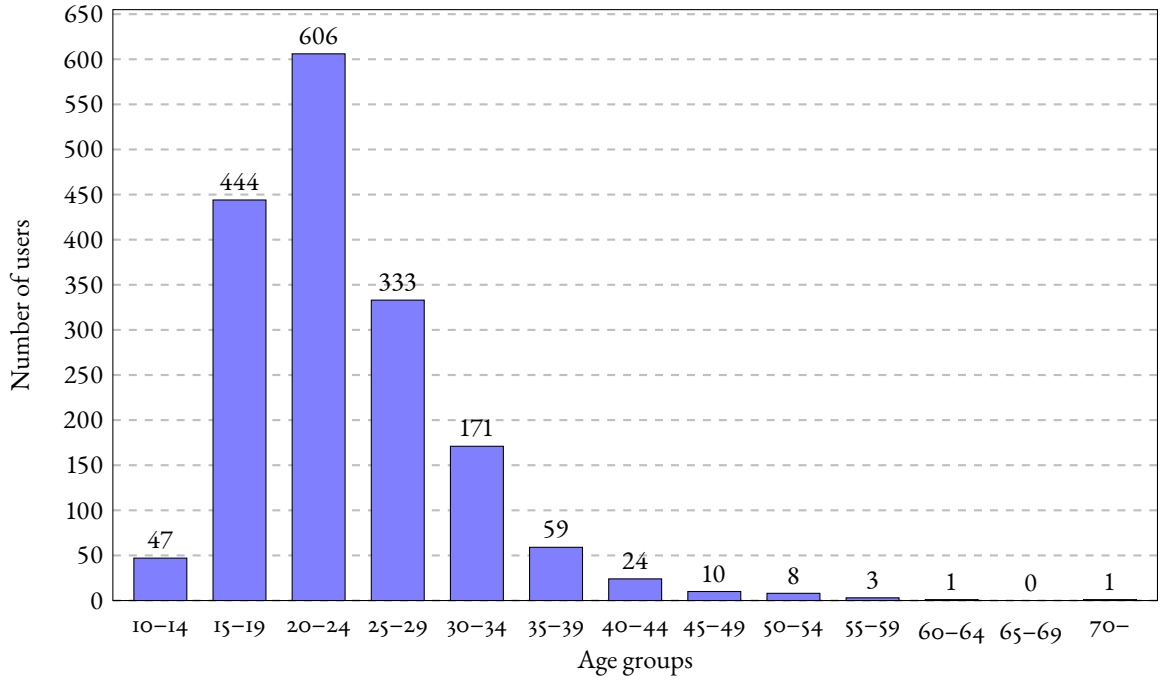

Figure 10: A bar chart of the number of individuals per age interval who disclosed their ages in posts on the r/depression subreddit.

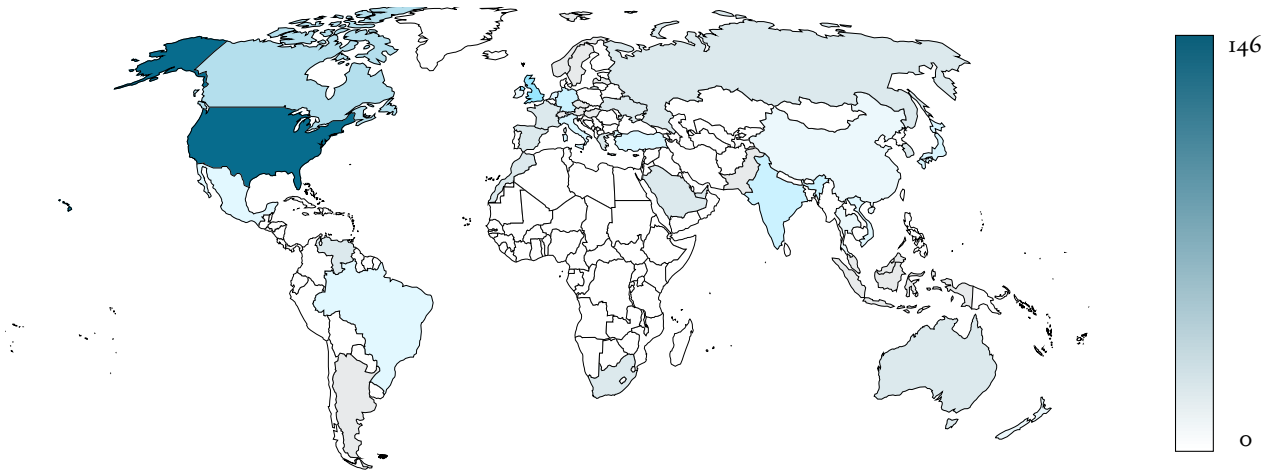

Figure 11: A world map of the frequency of countries of origin mentioned in posts in the r/depression subreddit dataset.

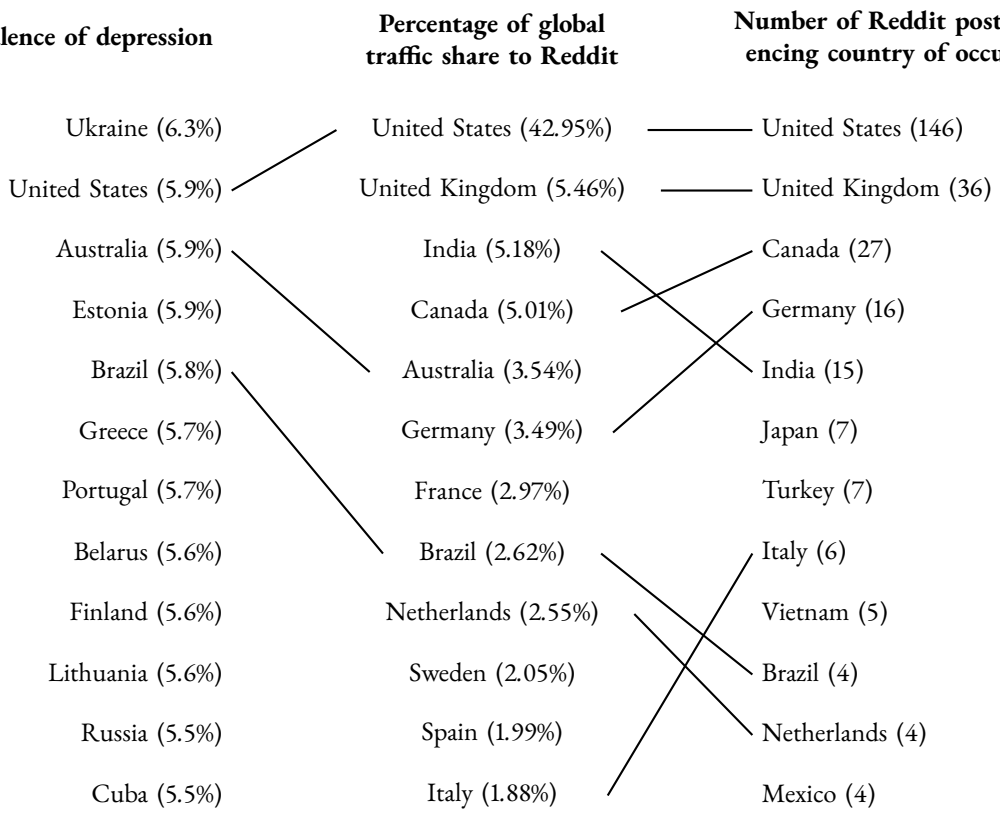

Figure 12: A connection diagram mapping the placement of the top 12 countries in terms of prevalence of depression, the percentage of global traffic share to Reddit and number of Reddit posts which reference country of occupancy.

The United States of America, the United Kingdom, Canada, India and Germany were of the countries that were explicitly mentioned in the dataset most often. These countries also appear in the top 12 countries with the highest traffic share to Reddit (Semrush 2023). The United States appears

near the top of all of the rankings in terms of prevalence of depression, global traffic to Reddit and references made in the subreddit dataset. Countries with increased prevalence of depression that are not mentioned in the Reddit corpus are Ukraine, Estonia, Belarus, Finland and Lithuania. This is, in part, because Estonian posts were removed from the corpus, and largely because these countries were implicated in the Russo-Ukrainian War in 2023. It is not possible to accurately determine the degree of proportionality between prevalence of depression, use of Reddit and participation in the r/depression subreddit. This is because the geographic locations of Reddit users were inferred from their text only.

The map in Figure 11 and the information in Figure 12 is compared with the map in Figure 13 which depicts the prevalence of depression for all countries (World Population Review 2023). Countries with higher rates of prevalence of depression are shaded darker than countries with lower rates of depression prevalence. Correlation statistics based on the frequency that countries are mentioned in the dataset with the prevalence of depression in that country and the percentage of traffic share from that country to Reddit is summarised in Table 3.

The Spearman and Kendall correlation coefficients indicate a weak positive correlation between the mentioned countries and their corresponding prevalence of depression with high statistical significance ($p < 0.01$). These correlation coefficients may be a result of limited geolocation data available and extrapolating meaning from direct mentions of locations in the Reddit posts. The correlation statistics also indicate a moderately strong positive correlation between the mentioned countries in the dataset and the corresponding percentage of global traffic share to Reddit from those countries with high statistical significance. Thus, there exists some correlation between the prevalence of depression within a country and the likelihood for someone from that country to post about their experience about depression on the subreddit.

5 Discussion and conclusion

This investigation explored the ways in which individuals on the r/depression subreddit interacted with and in an OMHC. People of diverse ages, habits and geographies meet in this community to share and to provide and receive support for them-

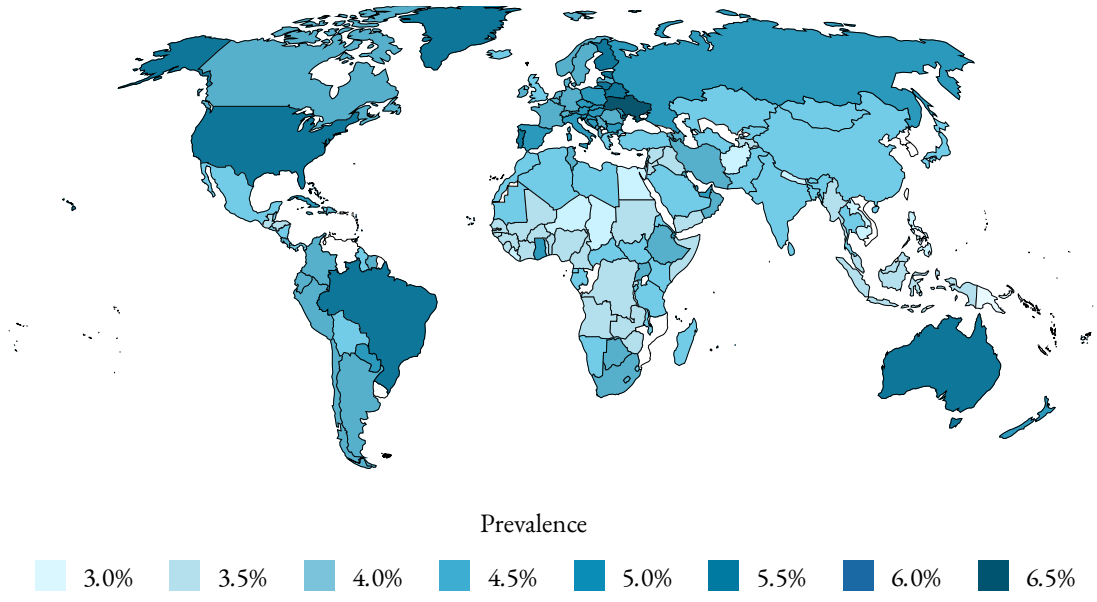

Figure 13: A world map of the prevalence of depression in 2023.

Table 3: The correlation statistics comparing the frequency of countries mentioned in the dataset with the country's prevalence of depression and the percentage of global traffic share to Reddit.

			Prevalence	**Reddit share**
Frequency	Spearman's rho	Correlation coefficient	0.273	0.692
		Significance level	$p < 0.01$	$p < 0.01$
		N	180	76
	Kendall's tau	Correlation coefficient	0.217	0.552
		Significance level	$p < 0.01$	$p < 0.01$
		N	180	76

selves or for individuals they know are living with depression. The subreddit is a forum dedicated to a culture of safety and a source of support where individuals use the space in ways that make them feel most comfortable. The analysis of general engagement showed that some individuals who post to the r/depression subreddit may have multiple user accounts. This is based on at least one instance where two posts, identical in title and main text, were made nine minutes apart from two different accounts. Feelings of anonymity provide individuals with the confidence to engage from a position of safety, to express freely and openly, and to overcome social barriers by exploring a new or other facet of their identity without fear of prejudice.

There were 8 322 accounts that only posted once in the data collection period, signifying that it is less common to post often in this subreddit. There are users who have posted to this subreddit multiple times, often to seek support or validation. Individuals on this subreddit tended to share information that may be relevant or helpful to others, but also as a means of release and venting. The average Reddit post in the dataset consisted of 178.07 (σ = 195.84) words per post. Some posts were more substantial and very detailed, with the longest post in terms of word count containing 5 192 words. Shorter posts tended to express a single sentiment or ask a question. These questions may have been rhetorical exclamation or an earnest plea. The most common words used on this subreddit (excluding stop words) were "like" (16 078), "feel" (14 513), "life" (10 044) and "want" (9 884).

The investigation of temporal attributes to posting on the subreddit yielded observations regarding posting habits. An average of 213.84 posts were made per day to the r/depression subreddit. Most posts were made at the start of the week, which were on Mondays and Tuesdays. Posts made on Mondays received more upvotes and posts made on Saturdays received more comments. This is indicative of individuals being on Reddit and engaging with other posts through upvotes and comments over weekends and at the start of the week. Upvotes also occur at a higher frequency to comments being generated. The most frequent time of posts being made to the subreddit occurred from 10 PM to 5 AM SAST. This result may be interpreted differently when considering the time zones representative of the majority user demographic of the subreddit, which are individuals from the USA. The increased activity shows more posts occurring between 1 PM and 8 PM for users in Los Angeles, with the most frequent posting activity taking place between 6–7 PM. For individuals in New York, the increased posting activity may be seen between 4–11 PM, with the highest posting frequency hour taking place between 9–10 PM. It is assumed that the percentages of global traffic share to Reddit is representative of the proportions of individuals that post in the r/depression subreddit based on the correlation statistics indicating a significant positive relationship between these two variables. It follows that the times that the most frequent posting activity took place between the early afternoon to late evenings for most of the Reddit users. Following this, fewer posts were made during the early morning hours of the day.

The analysis of the demographic and geographic at-

tributes of users' posts revealed that the age distribution of Reddit platform users is skewed to younger audiences, which is also represented in the age distribution of individuals posting to the r/depression subreddit. While geographic data of the Reddit users were not accessible via the Reddit API, it was possible to extract regional information from the text of posts if a user has explicitly disclosed this information. There were no assumptions made that a region mentioned in a post represented the actual location of the user, hence a manual evaluation of each post with regional information was conducted and the location at a country-level was ascertained. There was a strong positive correlation between the frequency of countries mentioned in Reddit posts and the countries with the highest global traffic share to Reddit. There was a weak positive correlation between a country's prevalence of depression and the frequency of countries mentioned. This weak correlation may be due to the posts from those countries being excluded from the dataset.

The insights from this investigation provide a snapshot of the interactions of people engaged on the r/depression subreddit during a particular point in time. The nature of the OMHC changes as the circumstances of the forum members change. Convenience to post and engage with posts may be the reason for an increase in the number of posts made in a day around the same time as a multinational public holiday compared to the rest of the days considered in the study. Similarly, the relationship between depressive symptoms and nighttime social media use is represented in the increased activity during post meridiem times for the majority of r/depression subreddit users. Results may vary from other datasets captured at different times of the year and in the context of other current events. The use of automated tools to extract meaningful information may include translation of non-English texts. Data may also be collected over greater periods to investigate whether global and country-specific news impacts users' participation in OMHCs.

This investigation only explores the meta-properties of the subreddit posts. Future work will consider the content of the posts by way of natural language processing (NLP). Text data, particularly from social media, are often unstructured and noisy because it is sparse or incomplete (Egger & Yu 2022). Despite this, the short form of posts published on social media is appealing to some when it comes to expression (Sabate et al. 2014). Regardless of platform preferences and the kinds of content that social media users choose to consume, it is clear that stories and messages are being exchanged. Future work will be to apply topic modelling on the subreddit data, which is a form of text mining by which to identify main themes or topics within a collection of text documents. Interpretation of these models, as with most NLP models, relies on manual evaluation and domain knowledge to validate the results (Egger & Yu 2022). Advancements in large language models have made it accessible for researchers to employ topic models beyond the traditional probabilistic approaches such as latent Dirichlet allocation (Blei et al. 2003). The use of linear algebraic models such as non-negative matrix factorisation (Lee & Seung 2001) and embedding models such as Top2Vec (Angelov 2020) and BERTopic (Grootendorst 2022) will be explored in later research. Findings from this study will assist in validating the outputs from the topic models.

Notes

[1] The full list of information and links to sources regarding depression have been posted on the subreddit at `https://www.reddit.com/r/depression/wiki/what_is_depression/`.

Acknowledgements

The author thanks his doctoral supervisors, Lieschen Venter, Stephan Visagie and Stephanie Merchant for their academic support. The author is supported by the South African Department of Higher Education and Training's University Staff Doctoral Programme 2021–2025.

Ethical clearance

Research ethics approval to conduct this study was reviewed and approved by the Social, Behavioural and Education Research Ethics Committee of Stellenbosch University (project ID: 26373).

References

Aladağ, A. E., Muderrisoglu, S., Akbas, N. B., Zahmacioglu, O. & Bingol, H. O. (2018), 'Detecting suicidal ideation on forums: proof-of-concept study', *Journal of Medical Internet Research* **20**(6), 1–3.

Angelov, D. (2020), 'Top2Vec: Distributed representations of topics', *arXiv preprint arXiv:2008.09470* .
URL: *https://doi.org/10.48550/arXiv.2008.09470*

Bak, J., Lin, C.-Y. & Oh, A. (2014), Self-disclosure topic model for classifying and analyzing Twitter conversations, *in* 'Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing', pp. 1986–1996.

Balani, S. & De Choudhury, M. (2015), Detecting and characterizing mental health related self-disclosure in social media, *in* 'Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems', pp. 1373–1378.
URL: *https://doi.org/10.1145/2702613.2732733*

Becker, S. P. & Lienesch, J. A. (2018), 'Nighttime media use in adolescents with ADHD: Links to sleep problems and internalizing symptoms', *Sleep Medicine* **51**, 171–178.
URL: *https://doi.org/10.1016/j.sleep.2018.06.021*

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003), 'Latent dirichlet allocation', *Journal of machine Learning research* **3**(Jan), 993–1022.

Chen, Y. & Xu, Y. (2021), 'Exploring the effect of social support and empathy on user engagement in online mental health communities', *International Journal of Environmental Research and Public Health* **18**(13), 1–18.
URL: *https://doi.org/10.3390/ijerph18136855*

Cunningham, S., Hudson, C. C. & Harkness, K. (2021), 'Social media and depression symptoms: a meta-analysis', *Research on child and adolescent psychopathology* **49**(2), 241–253.

DataReportal (2023), 'Digital 2023 July Global Statshot Report', [Online], [Accessed 21 September 2023], Available at `https://datareportal.com/reports/digital-2023-july-global-statshot`.

De Choudhury, M., Gamon, M., Counts, S. & Horvitz, E. (2013), 'Predicting depression via social media', *International Conference on Web and Social Media* **13**, 1–10.
URL: *https://doi.org/10.1609/icwsm.v7i1.14432*

De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G. & Kumar, M. (2016), Discovering shifts to suicidal ideation from mental health content in social media, *in* 'Proceedings of the 2016 CHI conference on human factors in computing systems', pp. 2098–2110.

Drisko, J. W. & Maschi, T. (2016), *Content analysis*, Oxford University Press, New York.

Egger, R. & Yu, J. (2022), 'A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts', *Frontiers in Sociology* **7**.
URL: *https://doi.org/10.3389/fsoc.2022.886498*

Grootendorst, M. (2022), 'BERTopic: Neural topic modeling with a class-based TF-IDF procedure', *arXiv preprint arXiv:2203.05794* .
URL: *https://doi.org/10.48550/arXiv.2203.05794*

Hunt, D. & Brookes, G. (2020), *Corpus, discourse and mental health*, Bloomsbury Publishing.

Lee, L. & Seung, D. (2001), 'Algorithms for non-negative matrix factorization', *Advances in neural information processing systems* **13**, 556–562.

Matias, J. N. (2019), 'The civic labor of volunteer moderators online', *Social Media+ Society* **5**(2), 1–11.
URL: *https://doi.org/10.1177/2056305119836778*

Prescott, J., Rathbone, A. L. & Hanley, T. (2020), 'Online mental health communities, self-efficacy and transition to further support', *Mental Health Review Journal* **25**(4), 329–344.
URL: *https://doi.org/10.1108/MHRJ-12-2019-0048*

Proferes, N., Jones, N., Gilbert, S., Fiesler, C. & Zimmer, M. (2021), 'Studying Reddit: A Systematic Overview of Disciplines, Approaches, Methods, and Ethics', *Social Media+ Society* **7**(2), 3.
URL: *https://doi.org/10.1177/20563051211019004*

Reddit Inc. (2005), 'About Reddit', [Online], [Accessed 6 September 2020], Available at `https://www.redditinc.com`.

Sabate, F., Berbegal-Mirabent, J., Cañabate, A. & Lebherz, P. R. (2014), 'Factors influencing popularity of branded content in Facebook fan pages', *European Management Journal* **32**(6), 1001–1011.

Saha, K., Ernala, S. K., Dutta, S., Sharma, E. & De Choudhury, M. (2020), Understanding moderation in online mental health communities, *in* 'International Conference on Human-Computer Interaction', Springer, pp. 87–107.
URL: *https://doi.org/10.1007/978-3-030-49576-3_7*

Schöpke-Gonzalez, A. M., Atreja, S., Shin, H. N., Ahmed, N. & Hemphill, L. (2024), 'Why do volunteer content moderators quit? Burnout, conflict, and harmful behaviors', *New Media & Society* **26**(10), 5677–5701.
URL: *https://doi.org/10.1177/14614448221138529*

Seabrook, E. M., Kern, M. L. & Rickard, N. S. (2016), 'Social networking sites, depression, and anxiety: a systematic review', *JMIR Mental Health* **3**(4), 1–2.
URL: *https://doi.org/10.2196/mental.5842*

Semrush (2023), 'Reddit.com Web Traffic Statistics', [Online], [Accessed 21 September 2023], Available at `https://www.semrush.com/website/reddit.com/overview/`.

Suler, J. (2004), 'The online disinhibition effect', *Cyberpsychology & behavior* **7**(3), 321–326.
URL: *https://doi.org/10.1089/1094931041291295*

Woods, H. C. & Scott, H. (2016), '# Sleepyteens: Social media use in adolescence is associated with poor sleep quality, anxiety, depression and low self-esteem', *Journal of Adolescence* **51**, 41–49.
URL: *https://doi.org/10.1016/j.adolescence.2016.05.008*

World Population Review (2023), 'Depression Rates by Country 2023', [Online], [Accessed 11 September 2023], Available at `https://worldpopulationreview.com/country-rankings/depression-rates-by-country`.