Understanding depression and anxiety through social media

December 14, 2023

Contents

1 Introduction

What users post on social media can be a reflection of the mental health of not only themselves, but also their wider community. This is especially true for content on Twitter, where people can post their true feelings and thoughts freely. Unfortunately not all exhibits of social media behaviour are optimistic - some are a true and honest reflection of mental health struggles that people are going through. For researchers aiming to understand mental health difficulties it is of utmost importance to analyse such posting behaviours and gain insight from them. In this project, this topic will be delved into deeper, but first the effort to understand the problem at hand will be aided with relevant literature.

Choudhury et al., 2013 aims to predict depression through social media posts. Tweets from individuals who have received a clinical depression diagnosis, were collected. The tweets were collected from the year preceding the onset of the individual's depression. The depression metrics included user engagement and emotion, their network graph, linguistic style, depressive language use, and mentions of antidepressant medications. This project was particularly informed by the method of determining depressive words, which involved calculating the tf.idf score of frequent terms and then employing a depression lexicon mined from Yahoo! to determine depressive tweets.

Another interesting aspect of this problem relate to posting habits. In Ríssola et al., 2022 it was stated that the posting habits of individuals with mental health problems and those without are markedly different. It was found that monthly posting variance is much higher for the mental health group. Topic-specific vocabulary and emotional content were also found more in their posts. Another important point mentioned in the paper is that individuals suffering from depression exhibit a much higher post frequency in the evenings. This may be an interesting analysis point, as being active on social media especially at night is a behavioural pattern that differs from the norm. As mentioned by ten Thij et al., 2014, for all languages and time zones examined in their study, posting activity on Twitter calmed down significantly during the nighttime. Although there is no direct way for this project to determine whether users truly suffer from clinical depression, time stamps can be used to determine posting frequency.

For further research that delves deeper into the analysis of mental health on social media, in Coppersmith et al., 2018 natural language processing methods are used to determine the suicide risk of individuals. Two datasets containing posts from multiple websites were analysed - one with posts by individual users who had previously attempted suicide, and the other from different users who had not. Supervised and unsupervised machine learning methods were employed to classify tweets as either suicidal or not. Although this research goes beyond the scope of this project, it is an important note on the current state-of-the-art. It can be concluded that employing sentiment analysis on mental health data is no straightforward task, and that the variety of methods presented in literature simply indicate the large scope of the task.

This project aims to approach the problem of better understanding mental health conditions such as depression and anxiety, and in particular what vocabulary and behavioural patterns they might be associated with. A dataset containing tweets indicating depression and anxiety will be observed to reveal these associations and patterns. After preprocessing and plotting immediate visual information, LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., n.d.) was used to determine frequencies of certain emotions present in the data, and to examine whether posters use more future-oriented or past-reminiscing language. VADER (Valence Aware Dictionary and sEntiment Reasoner) Hutto, 2023 was used for sentiment analysis. The results of the analysis were mapped against the times of day of the tweets to determine whether any trends could be found. Finally, tf-idf scoring was applied on the data to find frequent words and their scores.

It was found that the sentiments in both datasets were overwhelmingly negative, with the depression data mainly exhibiting sadness and the anxiety data exhibiting general negative feelings. Neither dataset exhibited future-oriented language. There was found to be an association between tweets related to both topics, and posting more frequently at late hours of the night. The tf-idf scoring mainly highlighted some potential shortcomings of the dataset, but nevertheless raised words related to medical interventions that raised important discussion points on self reporting compared to diagnosis of mental health.

2 Problem formulation

The main problem at hand is discovering words associated with the respective themes, depression and anxiety, as well as discovering other patterns of behaviour in the tweets. These would namely be time of posting, as well as more general themes such as ”future-orientedness”, which signifies the proportion of tweets discussing and looking forward to future events.

The hypothesis is that users experiencing both anxiety and depression will be more active in late hours. The contents of the tweets will also be expected to be negative, although the exact nature of it will have to be analysed more closely. This problem scope was chosen after doing research into the papers mentioned above. It appears that while Ríssola et al., 2022 conducted analysis into both topic-specific vocabulary and posting times, the mention of posting times was not brought to the forefront of the analysis. And not to forget preliminary data analysis steps - this project aims to get to know the data at hand on a deeper level, to really understand it and guide potential future academic research into pain points found here.

3 Dataset

The course-provided dataset for this project task is two csv files containing posts from Twitter about anxiety, and about depression. Both sets are relatively recent, with the anxiety tweets falling between 2021 and 2022, and the depression tweets spanning from 2018 to 2022.

A preliminary analysis reveals that many tweets seem to include a variation of the word ’anxiety’ or ’depression’ for each respective file. This may be a potential limitation of the data - in real life, not all expressions of poor mental health use such explicit language, with a lot of posters commonly using metaphors and euphemisms to express their feelings. The thematic range of the tweets also seems to be vast, ranging from less serious (”i have depression after watching season 4 of dexter”) to very serious (”At 23 I was forced into making the decision to move into residential care the bullying stopped at the age of 40. I was diagnosed with depression and developed autistic trates.”). These points will be revisited in the data exploration section.

Both files also include columns for interaction metrics, as well as other metadata. The date column indicates the date and time when the tweet was posted, including the timezone information. The user column contains information about the user who posted the tweet, including their username and ID.

Having all this context and metadata about the tweets in the dataset will potentially prove informative for this analysis. In particular, the date and time posted, could reveal interesting patterns and insights. A preliminary hypothesis is that people exhibiting mental health difficulties would most likely be active on social media late at night.

The depression dataframe has 35101 entries, whereas the anxiety dataframe has more than double, with 69946 entries. This means that both of them are relatively large datasets, as usually datasets with over 10000 entries can be considered medium-sized. As illustrated

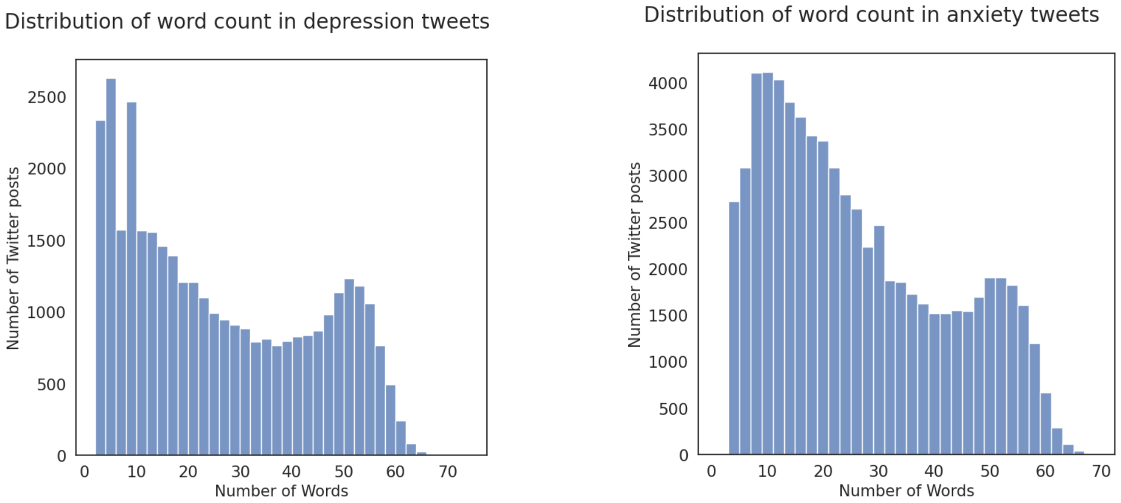
Column	Datatype	Description
url	object	The url of the tweet
date	object	The date that the tweet was posted
rawContent	object	The raw tweet text
renderedContent	object	The rendered tweet text
id	float	The unique tweet id
user	object	A dictionary of the username, id, and display name
replyCount	float	Count of replies to the tweet
retweetCount	float	Count of retweets of the tweet
likeCount	float	Count of likes the tweet has
quoteCount	float	Number of quote tweets
conversationId	float	The id of the conversation the tweet is part of
lang	object	Language of the tweet
source	object	The name of the app used to create the tweet
sourceUrl	object	URL of the source
sourceLabel	object	Labelled name of the source
links	object	Any links that the tweet contains
media	object	Links to any media part of the tweet (e.g. images)
retweetedTweet	float	Any tweet the tweet is retweeting
quotedTweet	object	Any tweet the tweet has quote tweeted
inReplyToTweetId	float	Id of the tweet the tweet is in reply to
inReplyToUser	object	User the tweet is replying to
mentionedUsers	object	Users mentioned in the tweet
coordinates	object	Coordinates of the poster's location
place	object	Location of the poster
viewCount	float	Viewcount of the tweet's video

Table 1: Datatypes in the dataframes

in table 1, the datatypes in both dataframes are the same. Although this analysis mostly concerns itself with the text content of the tweets, which will utilise the raw text column, there are also other columns of interest.

When investigating the datatypes of the columns, it was revealed that for both dataframes, the latter columns from links onwards, contain a lot of missing values. In particular for links, the depression dataframe contains over 33916 missing values, and the anxiety dataframe contains over 68240 missing values respectively. It can thus be said that these features are extremely sparse. Because of this it was decided that the latter features would not be used for analysis at all. The main features used for this analysis will be the rawContent column which will be further transformed, and the date column for extracting information about the time and date of the tweet.

More importantly, there was found to be one missing value for the raw text column in the anxiety dataframe. This entire row was removed from the dataframe.



(a) Depression

(b) Anxiety

Figure 1: Word counts in tweets

3.1 Preprocessing

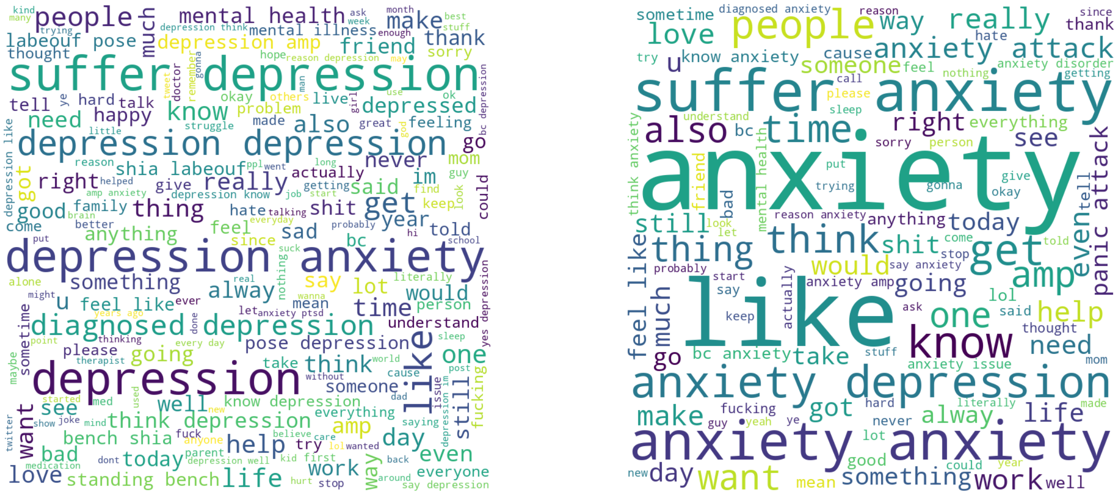
After investigating the nature of the data and potential missing values, preprocessing was applied on the rawContent text column. Using the **re** Python library, regular expressions are used to remove hyperlinks, mentions of users (i.e. the @ character), and newline characters that might occur in the middle of a tweet. All characters that aren't letters, numbers, and hashtags also get replaced with a space (for instance, emojis would get removed in this step), and at the end any extra spaces are removed. A new column called 'tweetCleaned' is created from this preprocessing. An additional check was performed to check whether all the text cleaning resulted in any empty tweets in the new column (e.g. tweets that were only a link or a mention), however there were none.

The next step of the preprocessing involved removing stopwords, which are defined as common words that do not add anything informative to a sentence, such as "the" or "to". After investigating some of the rows in the dataframe, it can be seen that the remaining words all add meaning to the sentences. However, since the dataframes are so large, it is possible that some uninformative words have made it through this stage. If revealed necessary in later steps, this step could be repeated with some custom stopwords.

3.2 Exploratory data analysis

In this section, the data is visualised to better understand the relationships of the different variables. Of particular interest are the most common words, as well as the time of day that most people tweeted.

As seen in Figure 1, most of the tweets are not particularly wordy, with peaks occurring at around 10 words per tweet and less. This may be problematic, as tweets which aren't very wordy might be more difficult to analyse, or may not be as informative. This potential setback will be kept in mind for subsequent steps.



(a) Depression

(b) Anxiety

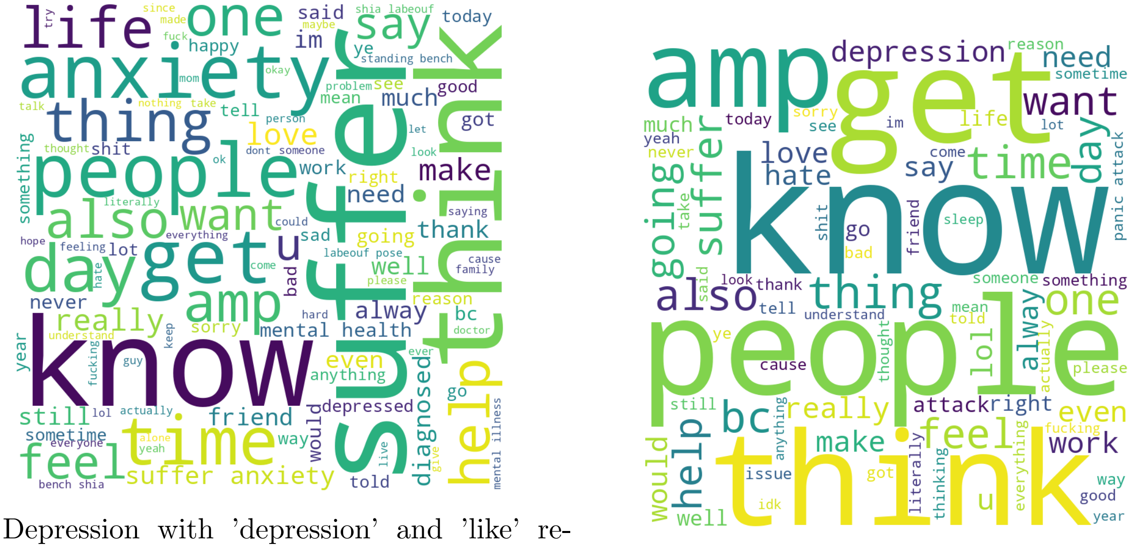
Figure 2: Wordclouds

Next, the distribution of words is analysed using wordclouds - a quick and visual assessment that will tell us more about the distribution of our words and reveal any future steps needed. The sentences were split into words, which were then fed to the WordCloud Python function to generate the visualisation.

From Figure 2 the same glaring problem that was mentioned in the preliminary analysis can be seen: the words depression and anxiety dominate their respective dataframes. This hinders our understanding of the data, as the aim of the wordclouds would be to reveal potential other words that frequently occur in these themes. To solve this, we are returning back to the stopword removal step, to remove these two words from their respective dataframes. However, the words will not be removed from the opposite dataframes as they are forecasted to provide important insight into the overlap between the two. It was also elected to remove the word 'like' from both datasets, as from the wordclouds it became evident that it was also disproportionately dominant in both datasets. From visual inspection of the raw text content, it could be seen that 'like' was used both in its intended uses (e.g. " I strive to be like this"), but also as a discourse particle as is common in informal English (e.g. "are you, like, really tired?"). Especially as an informal affective it does not add to the informativeness of the dataset.

From the modified wordclouds presented in Figure 3 we can start to unpack the true nature of the data at hand. The most common word in the depression dataset is "suffer", and in the anxiety dataset it is "people" and "know". Also, "anxiety" is among the top words in the depression set, and similarly with "depression" in the anxiety dataset. This indicates that there would be significant overlap between posters who self-report both of these conditions.

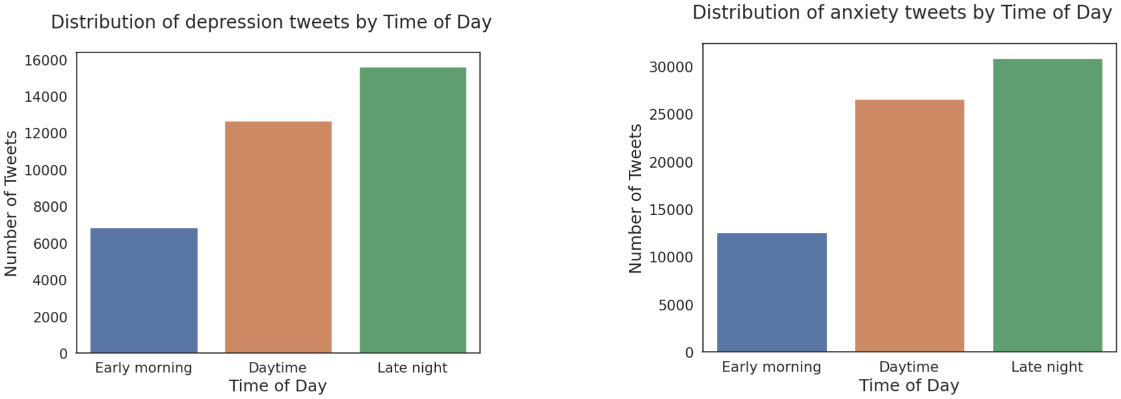
The date column was converted to datetime, and the time of the tweet was extracted. The times were split into three bins, with early morning being between 5am and 12pm, daytime being from 12pm included until 8pm, and 8pm to 5am being late night. As Figure 4 indicates, most of the tweets were posted late at night, which supports our hypothesis that



(a) Depression with 'depression' and 'like' removed

(b) Anxiety with 'anxiety' and 'like' removed

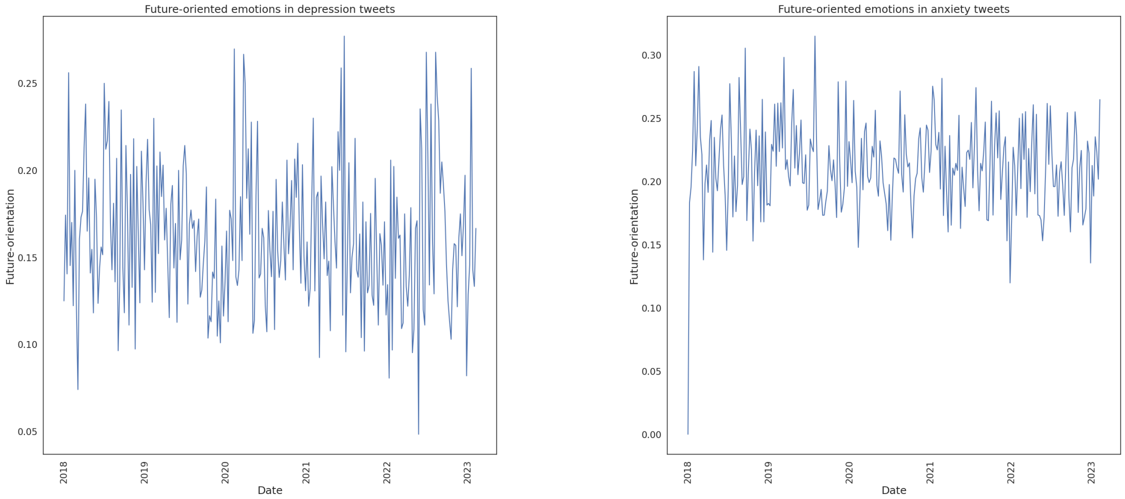
Figure 3: Modified wordclouds



(a) Depression

(b) Anxiety

Figure 4: Time of day distribution of tweets



(a) Depression

(b) Anxiety

Figure 5: Future-orientedness

individuals with mental health difficulties would spend more time being active online at late night hours.

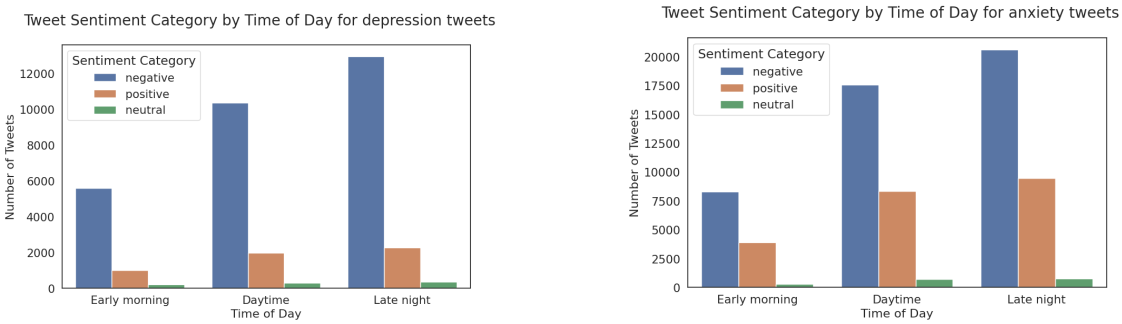
4 Methods

In this section, the data analysis tasks conducted in this report will be analysed.

The methods chosen for this task were LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., n.d.), VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto, 2023), as well as tf-idf scores. The choice to use VADER instead of the recommended Textblob library came from a comparison between their attributes and a conclusion that our aims could just as well be conducted with VADER.

Firstly, the LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., n.d.) Python library is used to conduct a linguistic analysis on the tweets, to map out positive and negative emotions present, and to plot any findings. After creating a new dataframe that calculates the normalized frequency of LIWC lexicon words in the original dataframes, these frequencies were grouped by emotion ('anger', 'sad', 'fear', 'negative', 'positive') and plotted. Quite unsurprisingly, the depression data had high frequencies of negative and sad emotions, and low frequencies of positive emotions and anger. On the other hand, the anxiety data had extremely high frequencies of negative emotions, but low frequencies of other emotions. It could be thus noted that anxiety is a distinct, still negative, emotion which has little overlap with emotions like sadness and anger.

Still using LIWC, the future-orientedness (i.e. the amount of language discussing and preparing for the future) of the tweets was analysed. The LIWC methods had added a column to the dataframe called 'future' and 'past', which were summed in order to figure out the proportion of tweets with future-oriented language. Figure 5 demonstrates the results of plotting this. For both datasets, the values fall into a range well below 0.5 (1 being the maximum), meaning that most tweets in the dataset are not discussing future events, and



(a) Depression (b) Anxiety

Figure 6: Sentiment analysis per time of day

instead could be seen to reminisce on past events instead. The values in the depression data oscillate strongly between 0 and 0.30, while the values in the anxiety data stay relatively stable at around a 0.30 frequency. This finding for the anxiety dataset is surprising, as anxiety tends to be a future-oriented condition, with worries generally concerning future events.

The next analysis step involved using the VADER (Valence Aware Dictionary and sEntiment Reasoner) (Hutto, 2023) method for sentiment analysis. It is a sentiment analysis tool capable of handling social media-specific language. It was applied directly on the cleaned tweet data to determine a sentiment score for each tweet, the sentiment categories being negative, neutral and positive. A column for this, as well as another column for the dominant sentiment label. Using the previous steps where tweets were split into times of day, the sentiment was plotted against the time of day that the tweet occurred. Figure 6 demonstrates that the number of negative tweets steadily increased during the late night for both depression and anxiety tweets. It can also be observed that the number of positive tweets is significantly larger for the anxiety tweets than for the depression tweets, with the depression tweets remaining overwhelmingly negative.

As a third and final measure, tf-df scores were employed in hopes to uncover deeper relationships between variables. Tf-idf vectorization is a popular technique used to measure the frequency of a term in each document (in our case documents being individual tweets), contrasted with their inverse document frequency (i.e. how frequent a term is across all documents). For this part of the analysis, only the tweet text was needed, and it was required to be in list form. Some additional preprocessing was needed, including lemmatization, which transforms words back into their basic form (e.g. ”am” mapped to ”be”). As such, the preprocessing steps were repeated on the lists with lemmatization as the last step. Originally both stemming and lemmatization were tested, and it was found that lemmatization provided more effective results for later steps. The tf-idf analysis identified the words as presented in Table 2 as having the top tf-idf values. The tf-idf scores were also calculated for the anxiety data, however, the only two words raised in this were ”would” and ”go”, which have limited informational value for our analysis. Despite tweaking with the parameters of the TfidfVectorizer such as minimum and maximum frequency, no additional results were able to be extracted from the anxiety data. It should be noted that also for the depression data, the max features was set to 1000, min document frequency to 5, and max document frequency to 0.5.

word	tf-idf score
autistic	0.44
stopped	0.43
move	0.43
age	0.40
making	0.35
care	0.32
diagnosed	0.24

Table 2: Tf-idf scores for depression data

5 Results

The above sections presented data analysis on two datasets, containing tweets about depression and anxiety. It was found that users posting about both of the topics tend to post more late at night, although activity could also be seen during other times of the day. It was also found, using the LIWC Python library, that neither the depression tweets nor the anxiety tweets contained significant amounts of future-oriented language, possibly indicating that users reminisce on the past and don't look forward to things.

Sentiment analysis using the VADER method revealed that the large majority of tweets in both datasets are overwhelmingly negative. The anxiety dataset is slightly more balanced, with slightly less than 10000 positive tweets in both the daytime and late night. Our results, most of all, indicate that the posters in these datasets are at most active during the night.

From the wordclouds in the exploratory data analysis we can find more qualitative insight into other words and as such, themes that correspond to these two mental health conditions. ”Suffer” being the top word for depression makes sense, with the common expression ”I suffer from depression” probably occurring in the dataset. Other words, like ”anxiety”, ”people”, and ”think”, reveal that there may be correlation between the two themes. ”People” is also one of the top words in the anxiety tweets, which may indicate an asocial theme in the tweets.

Unfortunately the tf-idf had some problems with scoring the anxiety data, with only ”would” and ”go” showing up as frequent terms. The results are better for the depression data, with seven frequent words found: ”autistic”, ”stopped”, ”move”, ”age”, ”making”, ”care” and ”diagnosed”. It could be said that ”care” and ”diagnosed” indicate a medical intervention aspect, potentially pointing to users that are seeking support for their mental health problems. The word ”autistic” may indicate some overlap with other diagnoses. It's unclear what context ”stopped”, ”move”, ”age” and ”making” may have without further context. ”Stopped” may have to do with medication, and ”move” may have to do with lack of movement associated with depressive episodes. These raise an important contrast to the wordcloud words, which had to do with more subjective experiences such as ”people”, ”day”, and ”know”. The differences in the tf-idf frequent words and the wordcloud highlights an important distinction: whether depression or anxiety refer to clinical, diagnosed conditions, or self-reported states of mind. While both are indicators of mental health struggles, only one of them has measures for clinical interventions and alleviation.

One potential explanation for the low performance with the anxiety dataset could be the aforementioned shortness of the tweets. If the term frequency of a word is very low (i.e. once per tweet or less), this term will have low importance even if it appears in a lot of the tweets. Tf-Idf also penalises words for being frequent in all documents, and from the previous analysis steps we can tell that the two datasets are quite homogeneous with many tweets including very similar vernacular and vocabulary. Thus one might consider that tf-idf analysis may have been a misstep for this particular problem. It was also considered to construct frequent unigrams (one-word sets) and bigrams (two-word sets), however, after seeing the tf-idf results it was elected to not continue with this analysis method.

6 Conclusion and Discussion

This analysis presents a worrying trend in online behaviour which causes concern and serves as motivation for deeper investigation.

Social media is for the most part something that most people intrinsically use in their free time, and individuals who engage in posting at late hours of the night are ostensibly not following a regular sleeping pattern - something which could be another sign of poor mental health.

Indeed, Nakshine et al., n.d. lays out the negative effects of spending too much time with digital technology and on social media: they increase stress, anxiety, and sleep issues. It's linked to depression, suicidal tendencies, and screen-time-induced poor sleep quality. Excessive screen time can cause hyper-arousal, disrupt circadian rhythms, and negatively affect mental energy and development. Internet addiction is also a growing concern mentioned by the article.

6.1 Limitations

The main limitation occurred in the tf-idf stage, with the lack of results for the anxiety data. The tf-idf values for the depression are reasonably informative, but the question remains, whether there is something fundamentally mismatched that causes such few tf-idf values to appear.

The same could also be said for the sentiment analysis per time of day-analysis. On one hand it is interesting to see the distribution of tweets, but a more extensive sentiment analysis really dissecting further themes surrounding depression and anxiety could have been an extension of the existing analysis.

6.2 Future work

Potential future analysis into this topic could reconsider the aforementioned issue with the tf-idf values, potentially investigating whether another similarity measure could have been more powerful for this specific data at hand.

As mentioned above, more extensive research into the overarching themes surrounding these topics of mental health could be conducted. For example, analysing whether the negativity in the tweets is directed towards one's self, another person, or something more general, could be a starting point in collecting a framework for warning signs of depression or suicidal ideation.

The discussion also raises a point on the difference between clinical mental health conditions and self-reported mental health states. Further research would also be required to investigate the overlap between the two and how to most appropriately provide support and interventions, even with the absence of a diagnosis.

6.3 Conclusion

In this project, the relationship between mental health tweets, surrounding terms, and time of tweeting was investigated. LIWC, VADER and tf-idf measures were conducted to reveal patterns in the data. The report concludes with a note on the role that social media plays in mental health, in particular with circadian rhythms and internet addiction. Future work should direct itself in designing prevention measures using this myriad of social media data available.

7 References

References

Choudhury, M. D., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media [Number: 1]. *Proceedings of the International AAAI Conference on Web and Social Media*, *7*(1), 128–137. https://doi.org/10.1609/icwsm.v7i1.14432

Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical Informatics Insights*, *10*, 117822261879286. https://doi.org/10.1177/1178222618792860

Hutto, C. J. (2023, December 12). *Cjhutto/vaderSentiment* [original-date: 2014-11-17T16:31:45Z]. Retrieved December 14, 2023, from https://github.com/cjhutto/vaderSentiment

Nakshine, V. S., Thute, P., Khatib, M. N., & Sarkar, B. (n.d.). Increased screen time as a cause of declining physical, psychological health, and sleep patterns: A literary review. *Cureus*, *14*(10), e30051. https://doi.org/10.7759/cureus.30051

Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (n.d.). The development and psychometric properties of LIWC2007.

Ríssola, E. A., Aliannejadi, M., & Crestani, F. (2022, February 7). Mental disorders on online social media through the lens of language and behaviour: Analysis and visualisation. https://doi.org/10.48550/arXiv.2202.03291

ten Thij, M., Bhulai, S., & kampstra peter, p. (2014). Circadian patterns in twitter.