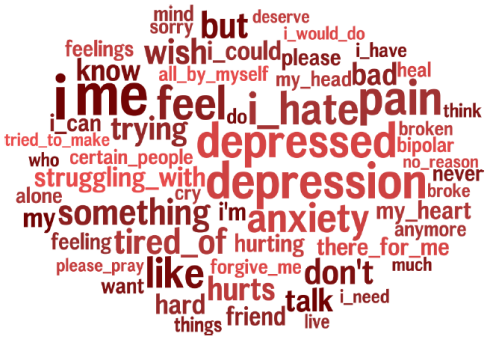
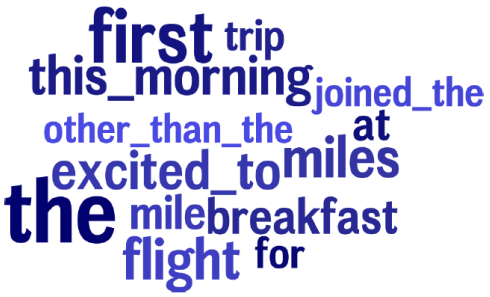
Chang An Le Harry Jr
GET1030 Individual Assignment





For this assignment, I will be analysing word clouds from a research article published in 2019, which aimed to understand and measure psychological stress levels based on social media posts.

In today's digital age, people are increasingly using social media platforms to inform others of their mental states, garner social support, as well as to record their daily activities. Despite previous studies which investigate the underlying factors behind stresses in our daily lives, many researchers believe that there is still a gap in the scientific understanding of how psychological stress is expressed on social media. In particular, they believe that mental health conditions, such as depression and anxiety, can be predicted from the social media language of its users.

In 2019, a study was implemented to explore how to differentiate high-stress users from low-stress users by performing natural language processing methods on text in social media posts.

To facilitate data collection, the study's researchers deployed a survey on Qualtrics, which consisted of several demographic questions (age, gender, race, education, income) and the Perceived Stress Scale questionnaire. Each item in the scale is scored from 0 to 4, with an absolute maximum summing to 40. The stress scores for each survey participant range from 6 to 39, with a mean value of 30.

Interestingly, the researchers chose to obtain their data from Facebook and Twitter, as they are amongst the most widely used social media platforms worldwide. Survey participants were invited to share access to their Facebook and/or Twitter posts. Among the participants, 601 users completed the survey and had active accounts with more than 900 words on Facebook and Twitter. Their social media posts were then downloaded by using the Facebook Graph and Twitter APIs. All participants who completed the survey were based in the United States.

The posts were then processed using the HappierFunTokenizer available with the DLATK package in Python. The language of each user and county is then represented as a set of features. In the dictionary-based method, social media language is transformed into numerical features representing percentage proportions of lexical categories in an existing dictionary. In the data-driven method, the language is morphed into numerical features which represent the proportions of word clusters that are statistically similar according to their frequency distributions.

Afterwards, 1-,2-, and 3-grams were extracted from all posts to analyse significant associations between words & phrases and stress. As seen above, the word clouds are visualizations of the Pearson correlations of words and phrases with stress scores obtained from the survey. The word clouds were generated by uploading the dataset containing the processed Facebook posts onto Wordle.net, a free online word cloud creator that no longer exists today. The red word cloud represents words commonly used by users with high stress levels while the blue word cloud represents words from Facebook posts of low-stress users. The size of each word indicates the correlation strength while the colour intensity indicates frequency (darker being more frequent).

Based on the researchers' analysis, the language of high stress users is made prevalent either by expressions of perceived lack of control, expressions of a need state or a lack of resources, along with a negative-angry sentiment. Also, high stress language seems to be comorbid with mental health conditions. Indeed, it is intriguing to see how these words reflect the adverse effects that stress can have on health. On the other hand, the language of low stress users has prominent positive affect, which include discussions of meals as well as feelings of social inclusion.

Most of the time, the size of each word in a word cloud corresponds to the relative frequency of that particular word in a corpus. In the case of this visualization, however, the Pearson correlation coefficient between words & phrases and stress score is used as the metric that affects word size in a word cloud. On the other hand, the colour intensity of each word is used to measure frequency instead. In short, the researchers have attempted to address the common limitation of size misinterpretation by adding another metric when generating each word cloud.

Although word clouds are simple to visualize and interpret, there are certain limitations that come with them. For instance, word clouds do not categorise words that have similar or the exact same meaning. In the context of my chosen visualisation, pairs of words such as "depressed" and "depression", as well as "I" and "me", are almost equal in size when compared in their own pairs. Having such synonyms as duplicated could omit out other unique words from appearing in the word cloud, which could affect the researchers' analysis of identifying the words with the highest Pearson correlation values within the processed dataset.

Another limitation of the visualisations discussed is that only Facebook data was used, despite also gaining access to the users' Twitter updates during the data collection phase. To improve the results of the analysis, separate sets of word clouds can be generated using posts from other popular social media platforms, such as Twitter, Instagram, LinkedIn and Reddit. These different sets of word clouds, representing each social media platform, can then be compared against each other for further analysis to evaluate if the words with the highest Pearson correlation values are consistent across all sets. This is especially important since the social media language used in each platform may vary.

To reiterate, word clouds are simple to use and were popularized in the early 2000s, when the photo sharing site Flickr first introduced their usage to display commonly used tags on

its website. Being one of the most widely used forms of information visualization today, critics believe that word clouds can often be misinterpreted by the general public, primarily due to the issue of word size and the lack of context. However, the word clouds discussed above are generally appropriate in my opinion, as it has been made clear that the clouds were generated for academic research purposes to supplement the findings of a mental health study.

References:

1. Guntuku, Sharath Chandra, Anneke Buffone, Kokil Jaidka, Johannes C. Eichstaedt, and Lyle H. Ungar. "Understanding and measuring psychological stress using social media." In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 13, pp. 214-225. 2019.
2. Viégas, F. B., & Wattenberg, M. “Timelines tag clouds and the case for vernacular visualization.” Interactions, 15(4), 49-52. 2008.