EECS: 510 Social Media Mining
Project Report

Social Media Analysis of Stress

By Amrita Das, Golla Pranith Reddy, Pradyoth Hegde

Abstract

Stress can be defined as a state of mental or emotional strain or tension resulting from adverse or very demanding circumstances. While some argue that adequate levels of stress are necessary in order to have improved productivity, if this goes unchecked, it can cause serious issues. In this project we perform a Social Media Analysis of Stress. We work on the real-world dataset, Twitter and gather Tweets for different keywords which relate to stress. We perform Sentiment Analysis and use other Elementary Machine Learning Algorithms to perform some analysis and provide key insights based on this dataset.

Introduction

Stress constitutes a serious hurdle in personal and public health. Tens of millions of people each year suffer from work related and personal stress leading to depression and only a fraction receives adequate treatment. [2] Stress can be defined as a feeling of mental or emotional strain resulting from different situations and circumstances. This can lead to several diseases including but not limited to cardiovascular disease, diabetes, asthma and other chronic illnesses. The goal of the project is to gather sufficient information about the stress of different individuals using tweets posted on Twitter to help better redress this problem.

Among the samples studied earlier, in patients undergoing stressful life events, larger social network size has been previously found to be associated with reduced mood disturbance, while aversive social support was associated with increased mood disturbance. [1]

Related Work

Sentiment Analysis can be an excellent source of information in providing insightful findings while determining market strategy, improve campaign success, product messaging, customer service and product reviews. [2] Amazon is a great example of using sentiment analysis techniques in providing better customer experience in its product category. However, apart from the above advantages, it can greatly benefit the healthcare industry in finding the root cause of most of the mental and physical illness.[3] In order to detect emotions in tweets, supervised learning methods were applied automatically to classify short texts, according to a finer-grained category of the emotions. The process flow of Emotex [4] Twitter messages are collected, labeled and classified according to the emotions the users convey. Twitter messages are gathered, features are selected and classifiers are trained to classify tweets into multiple emotions.

Theoretical Background

Mining Tweets: The first step in this project was to mine Tweets from Twitter. But, before doing so, it would help to have a better understanding about the corpus available along with more information about Tweets. Using [5] as a reference, we can summarize some basic information about Twitter handles and tweets.

A generic profile on Twitter consists of two main components:

- Tweets: These are colloquialisms used by Twitter to describe a status update. It is analogous to the body of an email. These should be 140 characters or less.
- Followers: These are the list of users that will receive a tweet when it has been posted. It is like the “To” field of an email. Information is spread wider if there are a lot of followers.

Besides the body of a tweet, there are a few other parameters which are important

- Mentions: To address a user, ‘@username’ is included in a tweet to refer to them directly. This allows users to quickly identify tweets directed at them.
- Retweets: This is a form of attribution, where ‘RT @username’ or ‘via @username’ denotes that the text originally appeared on another user’s profile.
- Hashtags: In addition to users, tweets can include tags to arbitrary topics by including a hashtag #topic. If sufficient users pick up on that topic, that would appear on the list of ‘trending topics’

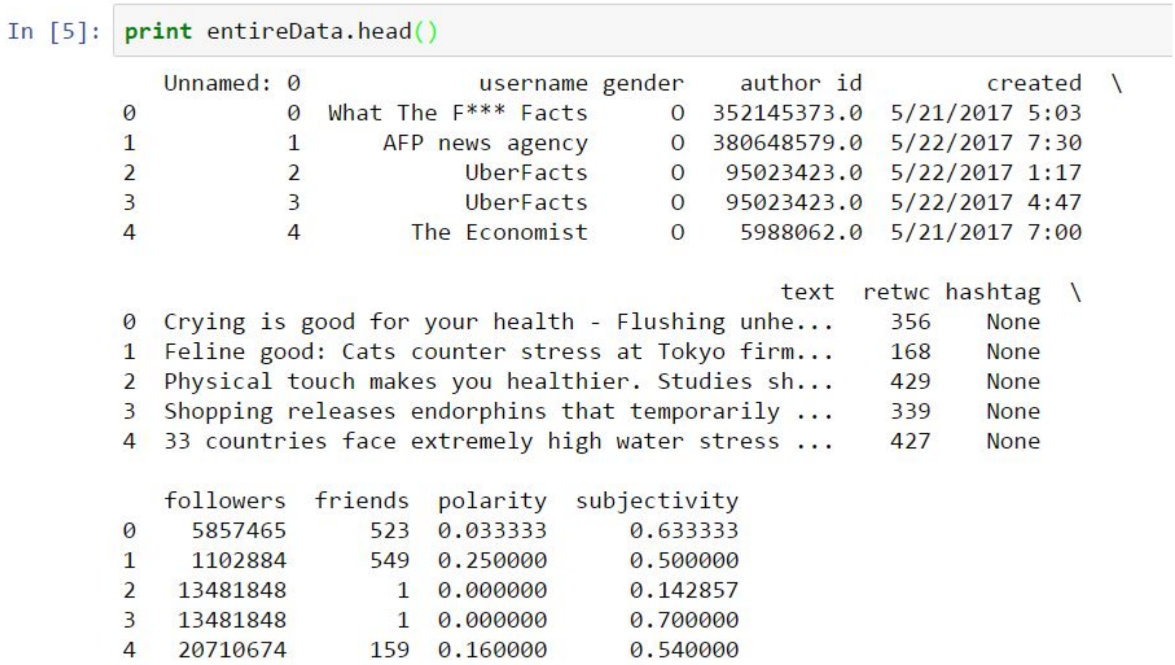
Amongst all the available attributes of a tweet, we use “username”, “authorid”, “when it was created”, “text”, “retweets”, “hashtag”, “followers”, and “friends”. Because the “text” attribute of a tweet has the information which the user is trying to express, we primarily work on this attribute. However, there is a lot of noise associated with the “text” attribute. There are a lot of Twitter users who post spam and other unnecessary content and that is why some filtering needs to be done to weed out the spam. After the tweets are cleaned, we can perform some sort of analysis on it. Unless we have clean data, the results obtained cannot be trusted. That is why data pre-processing is extremely crucial for any form of Analysis. The next sections are going to talk about the techniques we used to obtain the results.

Sentiment Analysis: The most basic form of analysis that can be performed on a given sentence/tweet is to find out if it is a positive or a negative sentiment. This is usually done using a term called “polarity”. There is another basic element in sentiment analysis which tells us whether a given sentence speaks in a broader sense, or is very specific. This is represented using a term called “subjectivity”. Using Python’s “Textblob” library, this was achieved. The “text” of the tweet was sent as an input to obtain the polarity value, which was in the range of -1 (very negative) to +1 (very positive) and a subjectivity value, which was in the range of 0 (very general) to 1 (very specific).

Gender Prediction: While performing Analysis, it would be helpful in order to be able to perform separate analysis for different genders. To predict the gender from username, we wrote a Python script which predicts the gender of a person based on the Username. It provided three distinct classes of outputs - Male, Female or Other. We used this script to predict the gender for each of the users.

Results and Analysis

We mined tweets for different keywords related to stress and summarized them as follows. The different keywords used to mine tweets are: “stress”, “stressed”, “#stress”, “#stressed”, “don’t worry”, “don’t stress”, “too stressed”, “not stressed”. The data folder contains the tweets mined on different days along with the merged files for each keyword. Since we couldn’t accommodate all the graphs that we would have liked to, we have included everything in the project folder of the SSCC. We tried to obtain both positive and negative elements of Stress. We totally had 1152 tweets collected over a span of 10 days. We passed the “text” attribute of the tweet into the Textblob package to obtain Polarity and Subjectivity values for the Text. We then used the “username” attribute of the tweet in order to predict the Gender of the Twitter user. There were three categories in the gender attribute, it could either be Male, Female or Other. Others was typically used to represent organizations who have twitter accounts and constantly post tweets.
Therefore, in order to perform the analysis, we had 12 attributes. We first created a confusion matrix in order to see the relationship between the different attributes for each plot. After that, separated positive, negative and neutral tweets based on the polarity values. We then plotted different polarity and subjectivity plots based on the gender.



```
In [5]: print entireData.head()

   Unnamed: 0            username gender    author id         created  \
0           0  What The F*** Facts      0  352145373.0  5/21/2017 5:03
1           1      AFP news agency      0  380648579.0  5/22/2017 7:30
2           2           UberFacts      0   95023423.0  5/22/2017 1:17
3           3           UberFacts      0   95023423.0  5/22/2017 4:47
4           4        The Economist      0    5988062.0  5/21/2017 7:00

                                               text  retwc hashtag  \
0  Crying is good for your health - Flushing unhe...    356    None
1  Feline good: Cats counter stress at Tokyo firm...    168    None
2  Physical touch makes you healthier. Studies sh...    429    None
3  Shopping releases endorphins that temporarily ...    339    None
4  33 countries face extremely high water stress ...    427    None

   followers  friends  polarity  subjectivity
0    5857465      523  0.033333      0.633333
1    1102884      549  0.250000      0.500000
2   13481848        1  0.000000      0.142857
3   13481848        1  0.000000      0.700000
4   20710674      159  0.160000      0.540000
```

Figure 1: Screenshot of first five entries of our corpus

This screenshot shows the first five entries of our corpus.
Username - Twitter Handle
Gender - Male, Female or Other
Author id - Used to differentiate users with the same name
Created - When the tweet was created
Text - The actual content of the tweet

Retweet - The number of times that particular tweet was retweeted
Hashtag - Any particular hashtags used
Followers - The number of followers, that user has
Friends - The number of friends, that user has
Polarity - A score from +1.0 to -1.0 to indicate the sentiment of a tweet
Subjectivity - A score from 0 to 1.0 indicating how specific or generalized a tweet is .

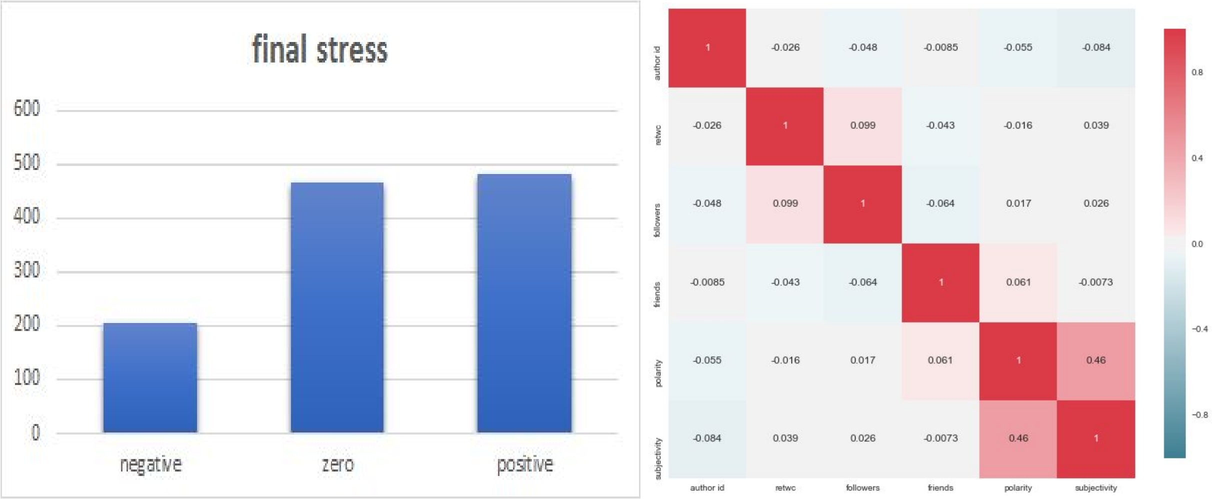


Figure 2: Counts for Positive, Negative and Neutral Tweets.

So, we have different plots to summarize this data.

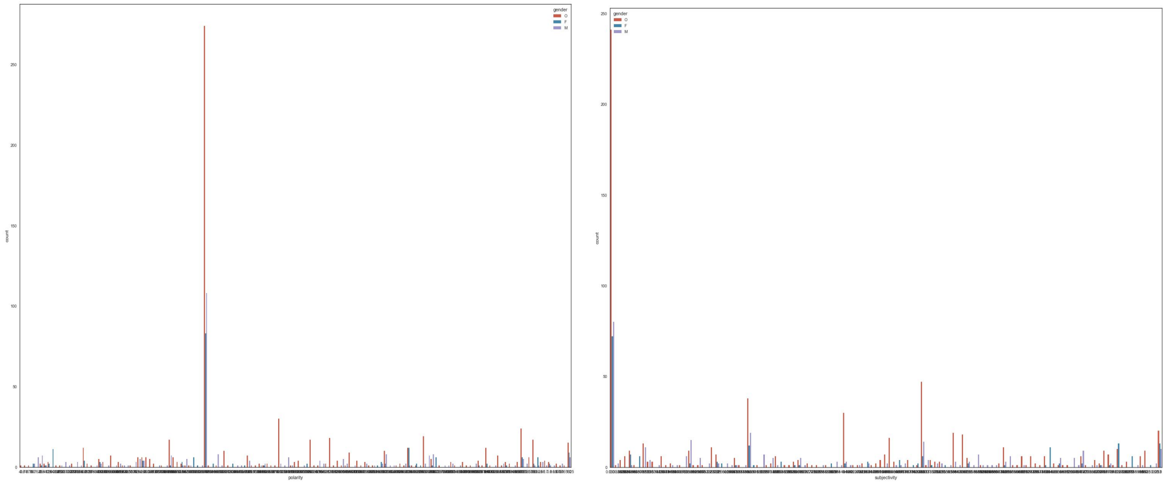


Figure 3: . Count Plots for Polarity and Subjectivity over the different genders

Count Plot for different polarity and subjectivity values across genders. If we look at each keyword individually, we plot the histogram for the Number of Positive, Negative and Neutral Tweets first, then we plot the confusion matrix, followed by the polarity and subjectivity plots across genders:

- "Stress"

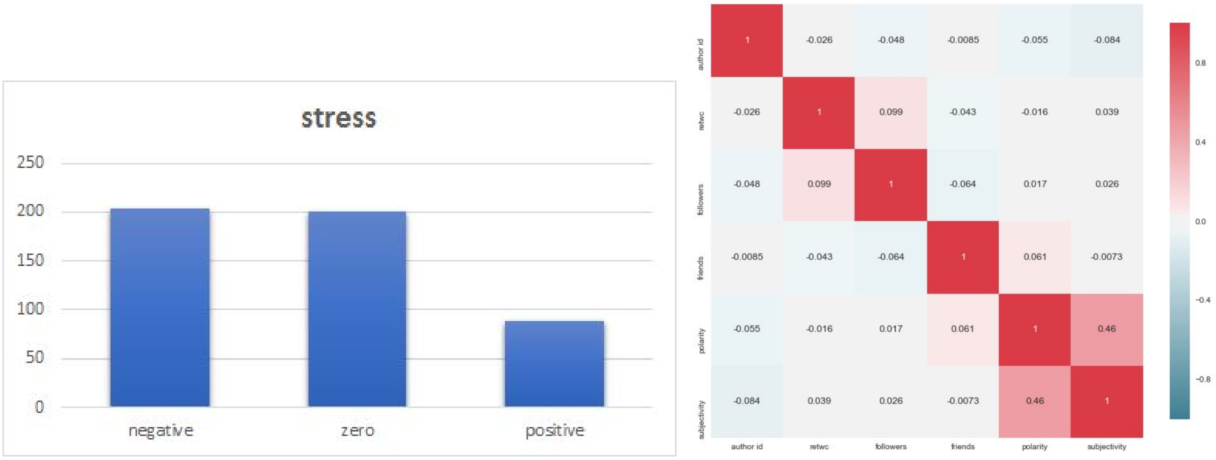


Figure 4: Correlation matrix for the word "Stress".

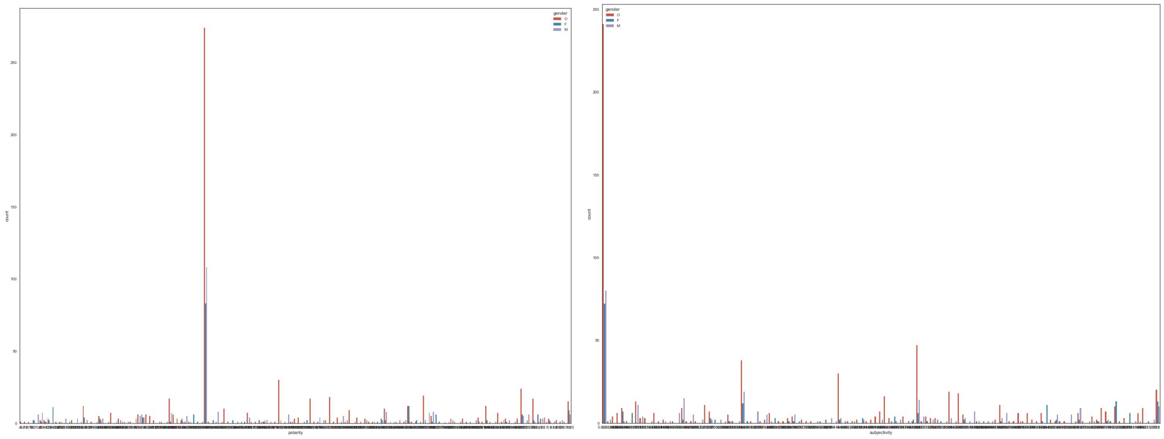


Figure 5:. Count Plots for Polarity and Subjectivity over the different genders

- "Stressed"

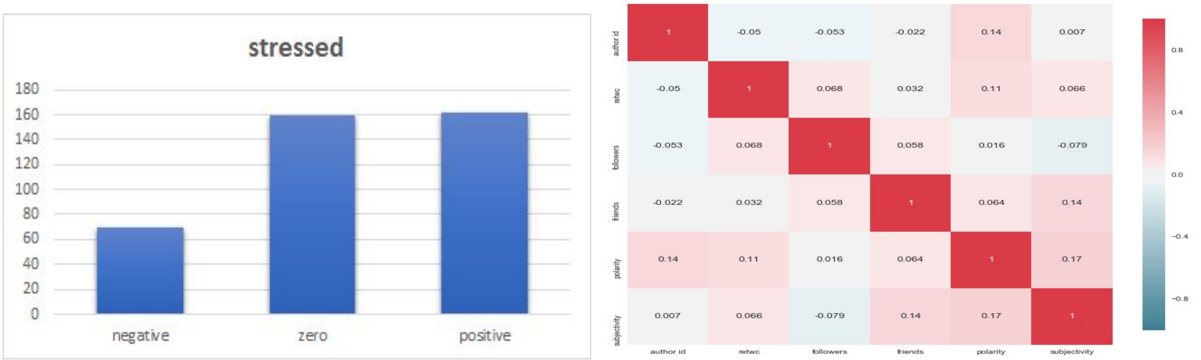


Figure 6: Correlation matrix for the word "Stressed".

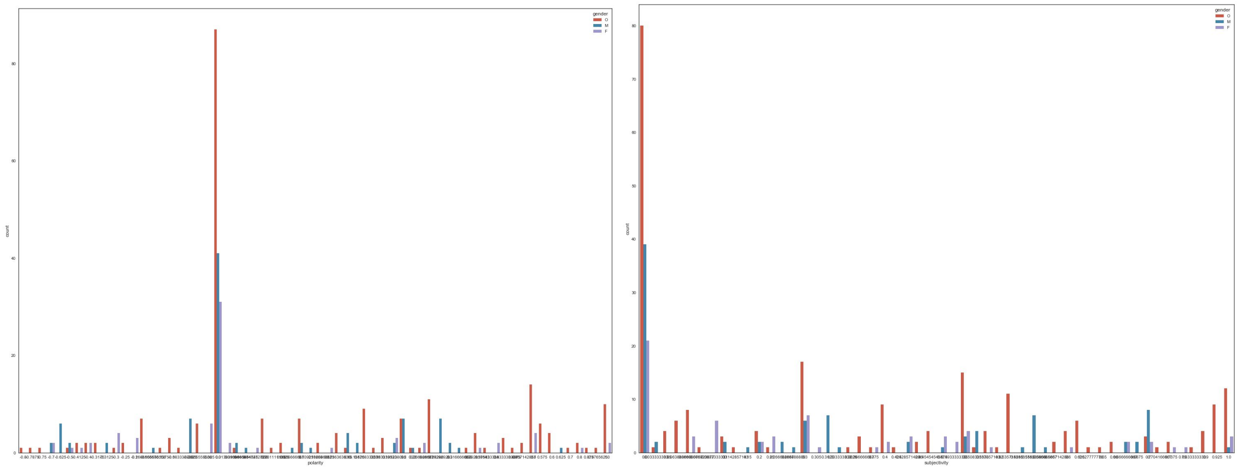


Figure 7: . Count Plots for Polarity and Subjectivity over the different genders

We only showed the graphs for "stress" and "stressed" because these two keywords provided about 900/1152 tweets and these two keywords represent the majority of our data.

Analysis

An examination of the given data shows us that there are more Neutral and Positive Tweets regarding Stress when compared to Negative Sentiments of Stress. This shows that there is a lot of positive reinforcement for individuals who are stressed on Social Media Websites like Twitter.

Some of the most Positive Tweets we had gathered with a polarity of 1.0 are:

- "Feeling stressed? SQUISH IT OUT! Watch my DIY stress ball video... Guys, they were awesome to make https://t.co/OpYmqa1Ka9"
- "Surround yourself with those who bring out the best in you, not the stress in you"
- "Too blessed to be stressed when your mom's the best. #MothersDay https://t.co/EEvZFQztRP"
- "13 of The Best Apps To Manage Stress by @LollyDaskal https://t.co/O5Vz4aOXk5 #Leadership #Stress #Business #HR"

Some of the most negative Tweets we had gathered with a polarity close to-1.0 are:

- "Trump's speech in Saudi Arabia stressed friendship in the face of 'violent extremism' https://t.co/ti3YRgaK8i https://t.co/flg8OnM0hn"
- "STRESSED OUT FROM BEING REAL IN A FAKE WORLD"
- "The economy is bad, people are stressed, suicide is on the rise. Dont pressure your debtors. Don't push them. Corpses can't repay debts."

Some of the most prevalent causes of Stress are:

- President Trump (and his Tweets)
- Current Economy
- Depression

Some of the best ways to alleviate Stress include:

- DIY stress ball
- By surrounding yourself with friends
- By helping someone else in need

We also wanted to check which gender of users contributed more towards the positive and negative twitter messages and the findings are as follows:

For Negative Tweets:

- 171 out of 672 were by Male - 25%
- 122 were by Female - 18%
- 379 by Others - 57%

For Positive Tweets:

- 112 out of 481 were by Male - 23%
- 65 were by Female - 13%
- 304 for Others - 63%

Work and Non-Work Related Stress

When it comes to classifying Work Related and Non-Work Related stress, we checked for tweets with the word "Work" and negative sentiment tweets and found that a mere 3% of tweets were work related stress tweets and the rest of them weren't related to work.

Users and "author id" and with the Most Negative Tweets:

- The guardian (87818409)
- Circa (441389311)
- Chippy (195951022)

Users and "author id" with the Most Positive Tweets:

- UberFacts (95023423)
- Gautam Gambhir (99448420)
- Edinburgh University (23426889)

Word Cloud

The word cloud generated from 2500 tweets consists of the most frequently used words associated with 'Stress' and this shows how different words are interlinked and associated on Twitter. The importance of each tag is shown with font size, with stress, depression and anxiety being important keywords.

Some tweets lack latitude and longitude information. In such a case, we removed them as we wanted to gather information about the geographic regions. We performed this analysis using R, where full address of each tweet location using the google maps API geocode.

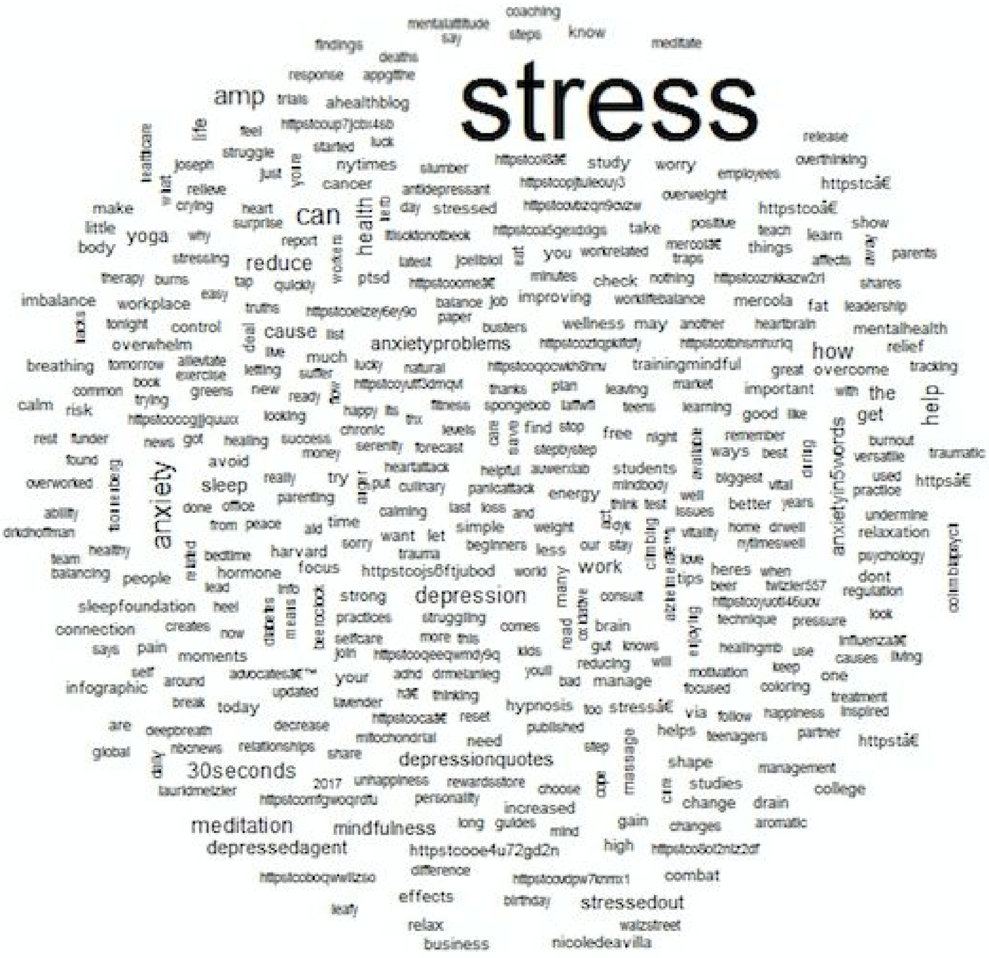


Figure 8: Word Cloud for "Stress".

Geocoding location data

With the Tweets gathered, we looked up user info for each of the users returned by the search. Converted textual user location data to approximate latitude & longitude coordinates with the Google geocoding web-service, geocode().

As shown below in Figure 9, the red dots shows the locations where users having tweeted regarding stress. From the graph below, we can say that people in the US and Europe are tweeting more about "Stress" than that of Asian countries with USA being the highest at 34% and Europe at 21%.

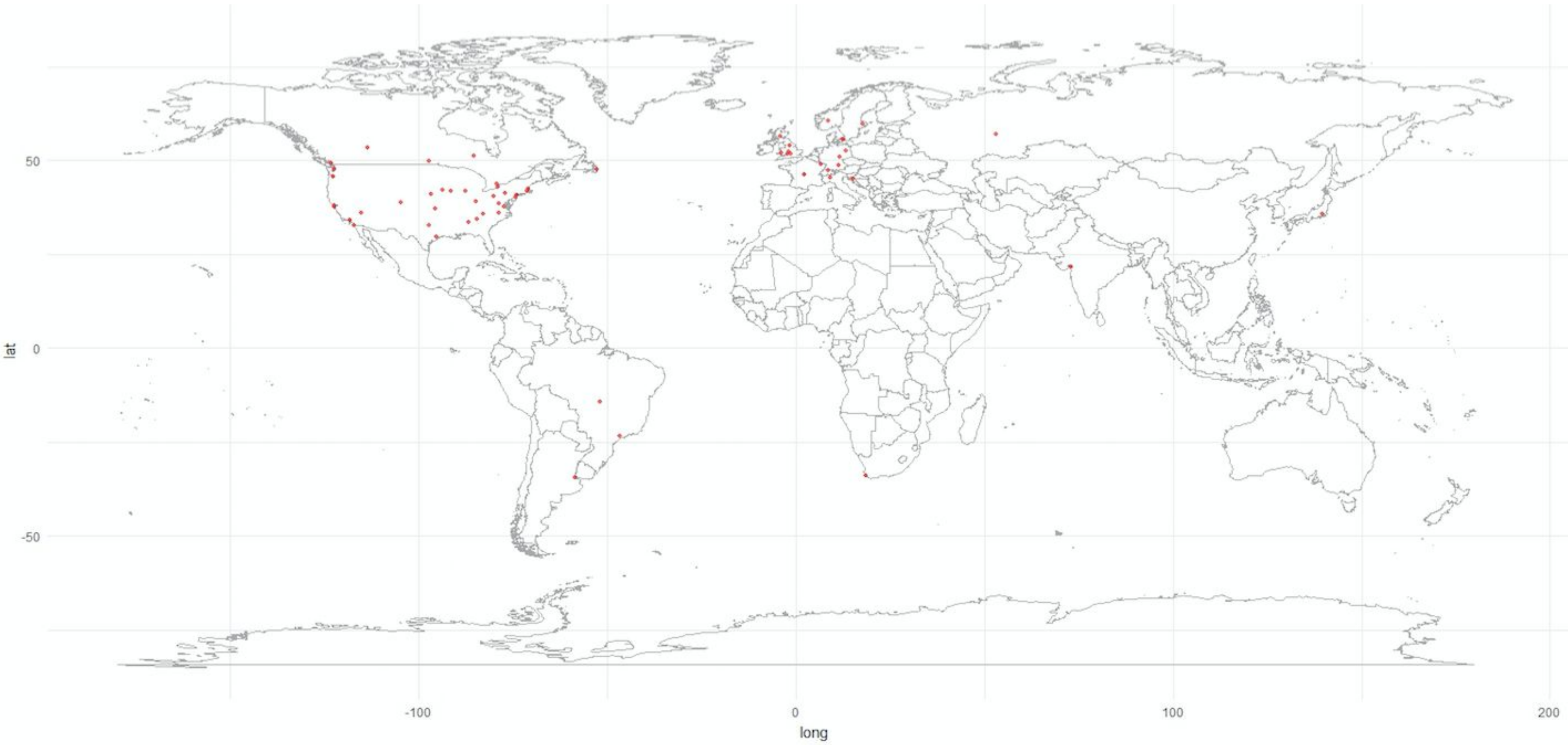


Figure 9: Graph plot showing tweets across different locations in the world.

Limitations

For some cases, it was helpful to check for subjectivity values in order to determine whether the Tweet was talking about a particular form of Stress or being very vague and general. That is why using just the Polarity values in order to understand more about the sentiment of something isn't always effective. Also, the quality of the dataset is extremely important. A twitter organization account only allows us to mine just 1% of the available tweets. In order to have accurate information, we need to know the complete picture, using just 1% of the available data can provide skewed results.

Conclusion and Future Scope

We can conclude based on the above analysis that Twitter users in US and Europe tweet about "Stress" more than the countries in Asia. Lifestyle and work life in the US and Europe could be the reasons. One of the future scope that we aim to achieve is gaining more insights about the number of people in each age group tweeting about "Stress" on social media like Twitter. Further, we aim to implement machine learning techniques in gaining deeper insights like tweets classified based on city, zip code and state from the tweet collection and analyzing on a large tweet collection dataset to help improve accuracy.

References

[1] Julie M Turner Cobb, Cheryl Koopman, Joshua D. Rapinowitz, David Spiegel, “The interaction of social network size and stressful life events predict delayed-type hypersensitivity among women with metastatic breast cancers- https://www.princeton.edu/genomics/rabinowitz/publications/23.pdf

[2] Christine Day,”The Importance of Sentiment Analysis in Social Media Analysis“- https://www.linkedin.com/pulse/importance-sentiment-analysis-social-media-christine-day

[3] Michael Gamon, Scott Counts, Eric Horvitz, Munmun De Choudhury “Predicting Depression via Social Media”

[4] Maryam Hasan, Elke Rundensteiner, Emmanuel Agu “EMOTEX: Detecting Emotions in Twitter Messages“ - http://galaxy.cs.lamar.edu/~kmakki/2014-ASE/2014%20ASE%20Conference%20Stanford%20University%20Proceedings/Regular%20Full%20Paper/submission103.pdf

[5] Grier, C., Thomas, K., Paxson, V., Zhang, M. 2010 “@spam: the underground on 140 characters or less (Links to an external site.)Links to an external site.” In Proceedings of the 17th ACM conference on Computer and communications security, pp. 27-37.