Investigating The Effects Of Social Media Usage On Mental Health

Deepak Subramani Velumani
University of Rochester
dsubrama@ur.rochester.edu

Abstract

Social media has become part of people's daily activities, many of them spend hours each day on social media. In this paper I am analyzing the sentiments of 3000 people on twitter platform and finding co-relation with their sentiments and the frequency of their posts on the social media. To assess the sentiments as either happy or sad I am using the VADER model, which will take the input as tweets of the 3000 people and classify them as happy or sad. After classification of the tweets and diving the users in the 1:1 ratio based on the account activity, the data is analyzed on the basis of 15 dependencies mentioned in the section 3, such as age, gender, location, friend count, parents (having children), smile on profile photo, and etc. After analyzing the sentiments, for having a better visualization I have plotted the all graphs based on attributes to assess the mental health based on happy and sad sentiment.

Introduction

Social media usage is increasing everyday due to the increase in the number of people getting access of internet in the developing countries, according to a survey done by Statista the number of people on the social media in 2020 were around 3.6 billion people and by 2025 the number is projected to be around 4.41 billion. These numbers are very high, considering exclusion of children and old people. In the same survey it has been estimated that the average user spends around 144 minutes per day. If this is the case then mining of the sentiment of social media is important as it can give us important co-relation in the data. In this proposed study, I am doing an Investigation on the effect of mental health using Twitter data. Twitter is a social media platform where a person can share their thoughts and many celebrities are also using this platform to share their views or connect to their fans or followers. Lately, studies have found that using social media platforms can have a detrimental effect on the psychological health of its users [2]. However, the degree to which we use the social media influences the public is yet to be determined. This systematic review can find that how usage of social media can affect the level of happiness and depression in individuals. Especially evaluating if there is any change in the sentiments of the people based on certain attributes like follower count, number of likes, gender, age, marital status and parental status and etc. These relationships between the data will let us know any hidden parameters which are contributing to mental Health of the people

Related work

With the increasing uptake of social media users across the population [1], it has led to the increasing interest in the intersections of social media and mental health. Many sentiment analysis and predictive models are built upon Twitter [2, 3, 4, 5], one of the most commonly used social media platforms. There have been studies to analyze tweets about scholarly articles to develop models to predict tweets sentiments about specific research domains [6]. Researchers Narr et al. [7] analyzed tweets in several languages and used the Naive Bayes classifier on the n-gram features to predict sentiment. In another study, Bae et al. [8] examined the polarity and sentiments of tweets posted by celebrities with over a million followers. Other studies have investigated Twitter in different ways. Zaman et al. [9] (2010) proposed a probabilistic collaborative filtering model that shows the spread of information by predicting future retweets. Furthermore, linguistic analysis of tweets was performed by Pak et al. [10] for their research. They used the Multinomial Naive Bayes classifier algorithm to predict positive, neutral, and negative sentiment in tweets using features extracted from preprocessed tweets. Wang et al. [11] extracted event-based tweets using Semantic Role Labeling (SRL), where they determined the salient topics in the events, using Latent Dirichlet Allocation (LDA). Other than crime and weather, the literature also includes analyses of disease incidence, stock market performance [12, 13], and many other topics. There are studies involving analyzing the expression of loneliness of a user in Twitter [14], in which they found traits such as difficult interpersonal relationships, drug usage, need for change, unhealthy eating and sleeping to be contributing to loneliness. Similarly, Lyu et al. conducted studies where they used user attributes provided by Tweepy API and other features such as religious status, political affiliations and also used LIWC analyzer to for analyzing controversial Covid-19 terms and public opinion on Vaccines respectively. [15, 16]. In this study I am combining the user attributes provided by the Tweepy API along with the LIWC 2015 [17] analysis of user tweets to discover the subtle features that contribute to the mental wellbeing of a user, by using VADER [18] to process the sentiments of the user tweets.

Methodology

In the implementation of this project, I have taken random tweets of 3000 twitter users and using their tweets data as Input I have generated the sentiments of these 3000 people and classified them as happy and sad with value of 0 to 1 using the VADER sentiment analysis tool after the classification of the people. There are many attributes that I have taken into account for the sake of deriving a relationship between mental health and social media usage I will list them out.

The Attributes:

1) Number of Tweets (Frequent vs Infrequent):

To calculate that a person is active or inactive, I have used the 3000 twitter users and calculated the median of tweets they have tweeted and it was around 47 tweets per month, so considering this if a user tweeted more than 47 tweets per month, I have classified them as frequent and if they have tweeted less than 47 then I have classified them as Infrequent.

2) Location - US maps:

To get the location of the user I have used the location specified by the user in the profile information, by using the geopy library in python I have got the co-ordinates of the user on the map, again using zipcode library in python I got the location of the user in the United States, Getting the location helped me in fetching other attributes such as population density, water level, median Income, median home value of that particular location where the user stays.

3) Follower's Count:
This is the number of followers a particular user has. It was fetched by using the tweepy API in python

4) Friends Count:

In this section I have calculated the number of friends (number of people a user is following) a particular user had. This was also fetched by tweepy API in python.

5) Listed Count:

The listed count for a particular user is the number of the communities in which the user is a member. This was fetched using the tweepy API.

6) Duration:

This is the amount of time the user has been using twitter, it is the date at which the account was created. It was fetched using the tweepy API

7) Likes Count:

It is the number of likes the user has provided to the other users in a month. It was also fetched using the tweepy API

8) Verified:

This is the number of people who have the verified status on their account, calculating it there were around 67 people out of 3000 who were having verified account. It was also fetched using the tweepy API

10) Age

This is the age of the user. It was also fetched using the profile picture of the user using the Face++ API in python. The data shows that the average age of the 3000 users was around 32 years.

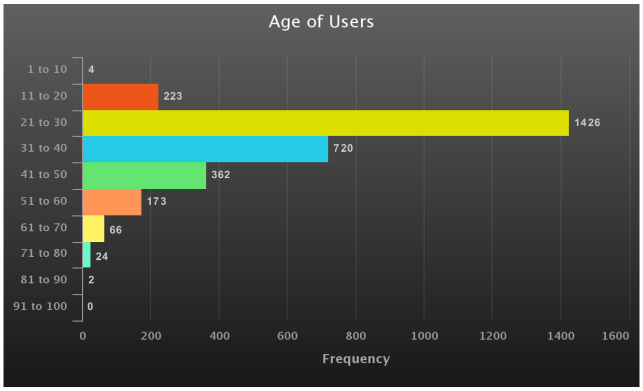

Figure 1: Representation of age of 3000 users on a scale of 1 to 100

11) Gender:

This is the gender of the user. This was fetched using the profile picture using the Face++ API in python. According to the results there were around 1698 males and 1302 Females in the 3000 users.

12) Smile:

The smile attribute has been derived from the profile picture of the user. If the user smiles in his profile picture, he will get the value as one and if he doesn't smile, he will be assigned the value as zero, I took Boolean values so that we classify the user as smiling and not smiling. I also have chosen the 3000 users for analyzing such that all 3000 people are having profile pictures

13) Population Density:

This is the number of people staying at a particular zipcode it was estimated using the geopy API and uszipcode API.

14) Water Level:

This is the water density of a particular place (water area/land area), this was taken into account to estimate if water level of a certain place has an impact on people's mental health whose sentiments are being estimated. It was computed using the geopy and uszipcode API's using the location information of the user.

15) House Income

I am calculating the median household for a particular user. I am doing this through the geopy API and uszipcode API, I am taking the location of the user and calculating the median Income of the are where the user is present. I am doing this to evaluate the user's sentiment based on the area where he is living.

16) Parents

To find if the user is father or mother of a children, I am doing Regular expression search on the tweets they have posted. If they mentioned anything in tweets such as ["my/our (1–20) year old," "my/our … X," and "I have … X" where X represents words that stand for kids, including "boy(s)," "girl(s)," "kid(s)," and "child(ren)."]. This was done using the RE library in python.

17) LIWC:

17.1) Average Words:

I wanted a value for each user from the LIWC attribute table so I took the average of all 3000 tweets for a particular user so that I could send that value in the Linear Regression model to get the contribution of that attribute to the sentiment of the word

17.2) Analytic:

From the LIWC table I am taking the attribute Analytical, which will represent if the user is Analytical (explains concepts thoroughly) or not.

17.3) Confidence:

From the LIWC table I am taking the attribute confidence, which will represent the confidence of the individual.

17.4) Authentic:

From the LIWC table I am taking the attribute Authentic, which will represent if the user is Genuine or not.

17.5) Exclamation:

From the LIWC table, I am checking for the people who use exclamatory remarks.

17.6) Negate:

From the LIWC table, I am inferring people who use Negate a lot (denying regularly).

17.7) Swear words:

From the LIWC table, I am inferring people who use Abusive language.

17.8) Self:

the LIWC table, I am searching for people who are using terms such as (I, me, my).

17.9) Group:

In the LIWC table, I am finding people who are using terms such as (we, us, our).

17.10) Power:

In the LIWC table, I am finding people who show superiority or bully others or being dominant.

17.11) Risk:
From the LIWC table, I am finding people who show them as risk takers.

17.12) Home:

In the LIWC table, I am finding people who are using many home related topics.

17.13) Work:

In the LIWC table, I am inferring people who are using work related topics in their posts.

17.14) Money:

In the LIWC table, I am inferring people who use money related topics to posts on the twitter.

17.15) Sexual:

In the LIWC table, these are people who use Relationships and sexual language while posting on twitter.

17.16) Reasoning:

In the LIWC table, these are the people who use reasoning in their posts (cause, because, etc.).

17.17) Certainty:

In the LIWC table, these are the people who are very clear and certain in whatever they post on the twitter.

Experiments and results

Below are the attributes which were evaluated based on the sentiments provided by the VADER sentiment analysis tool. I have evaluated the graphs and plotted them based on the results that I have received.

The Attributes:

1) Number of Tweets (Frequent vs Infrequent):

In the below Histogram we can infer the sentiments of the frequent and infrequent users of Twitter, here we can clearly see that as even thought the frequency of users is more it does not have much impact on the sentiment of the users. After doing the Linear Regression, the contribution of the frequency of the tweets on the sentiment was around negative 5%, this means that the increase in the frequency of the number of tweets had a negative impact on the user's mind.

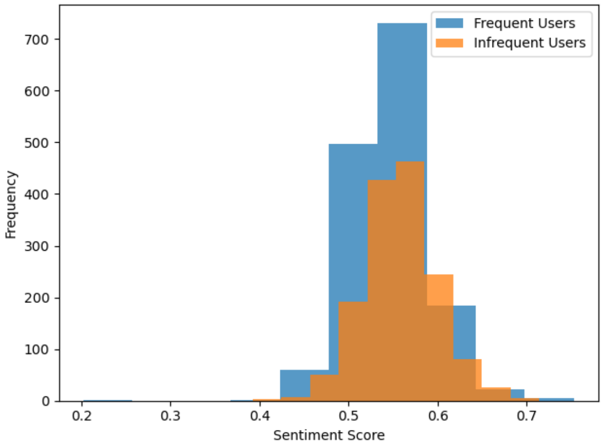

Figure 2: Frequency of Tweets vs Sentiment Score

After doing the T-test the T-value and P-value obtained were (Tvalue=7.3, Pvalue=1.78e-13).

2) Location – US maps:

The following graph shows the Twitter Users Distribution across the Unites States:

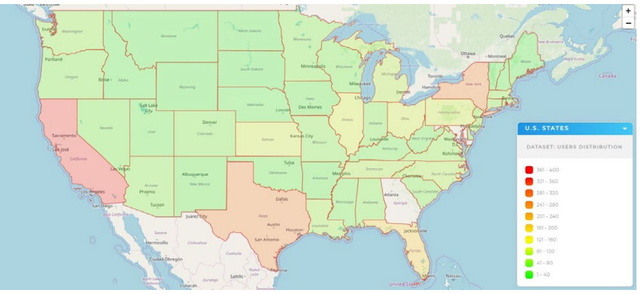

Figure 3: Distribution of Users across United States

It was observed that the urban states such as California, New York, Texas, Philadelphia and Georgia fall in the red spectrum which denotes high density of Twitter Users. States like Florida, Illinois, Ohio, Kansas and Washington fell in the yellow spectrum which denotes a medium user distribution. The remaining states such as Nebraska, Montana, Utah, Alabama, etc. were green in color, which showed a lower density of users.

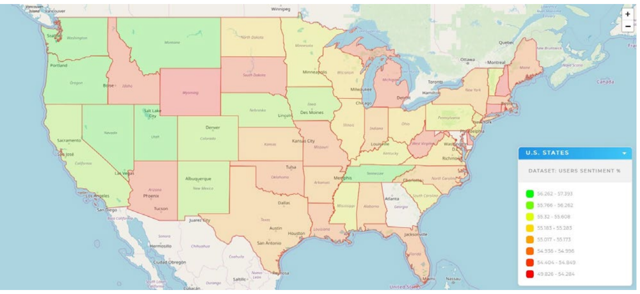

Figure 4: Sentiment Score of Users across States

The above graphs show the sentiment distribution of users across the states. States in green denotes users with a positive mental health and red states denotes users with a lower average mental health sentiment score. We found out that people in semi-urban states such as Nevada, Oregon, Nebraska, etc. had the best mental health compared to urban and rural states. States such as South Dakota, Wyoming, Idaho, Arizona and Boston were among the states which showed the lowest emotional health of the users. This perhaps shows that people with lesser access to technologies and the people who had all the access but lived in overpopulated areas, had a worser mental health compared to the people living in semi-urban areas.

3) Followers Count:

In the below histogram we can see that the with the increase in the number of followers a user had, made the user happier. After doing the Linear Regression, the contribution of the number of followers on the sentiment was around positive 15%, this means that the increase in the number of followers had a positive impact on the user's mind. After doing the T-test the T-value and P-value that I received were (Tvalue= -2, Pvalue=0.0455).

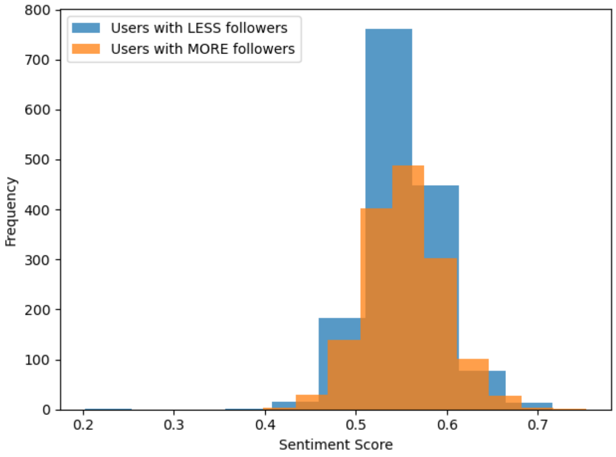

Figure 5: Followers Count vs Sentiment Score

4) Friends Count:

In the below histogram, although the two groups tend to be closer, we can see that the users who followed more people, tended to be more depressed than otherwise. This could possibly denote that people who spend more time thinking or analyzing other lives, tend to be less happy.
The linear regression showed that, the contribution of the number of people I follow to the sentiment was around negative 6%, this means that if I follow a lot of people then I will have a negative impact on my mind. After doing the T-test the T-value and P-value obtained were (Tvalue= -4.31, Pvalue=1.67e-05).

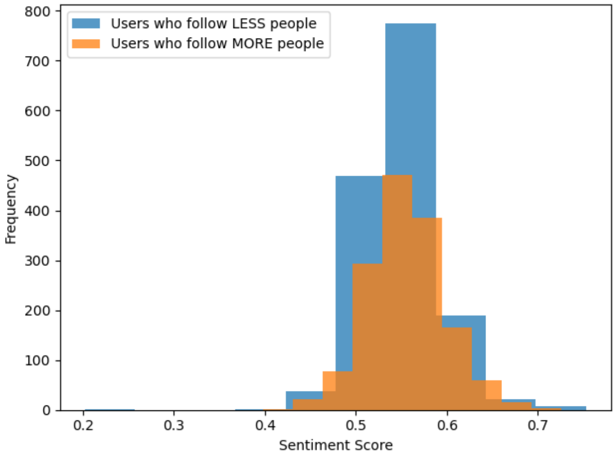

Figure 6: Following Count vs Sentiment Score

5) Listed Count:

In the below graph we can see that the with the more memberships the users are a part of, they tend to be sad instead of happy.

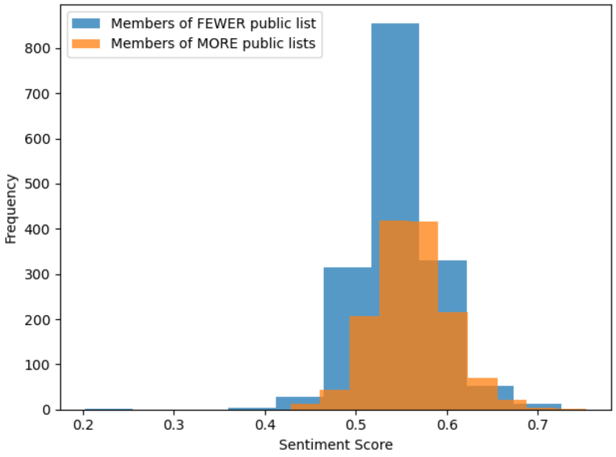

Figure 7: Memberships Count vs Sentiment Score

Linear Regression showed that, the contribution of the number of people I follow to the sentiment was around negative 8%, this means that if I follow a lot of online communities or groups then I will have a negative impact on my mind. After doing the T-test the T-value and P-value that I receive were (Tvalue= -10.61, Pvalue=7.25e-26).

6) Duration:

In the below graph we can see that the with the increase in duration in which users have had an account on twitter, they tend to be happier than those who created an account recently.

Linear Regression showed that, the contribution of the amount of time of since account created on twitter was around positive 4%, this means that if I follow a lot of online communities or groups then I will have a positive impact on my mind. After doing the T-test the T-value and P-value that obtained were (Tvalue= -7.33, Pvalue=2.88e-13).

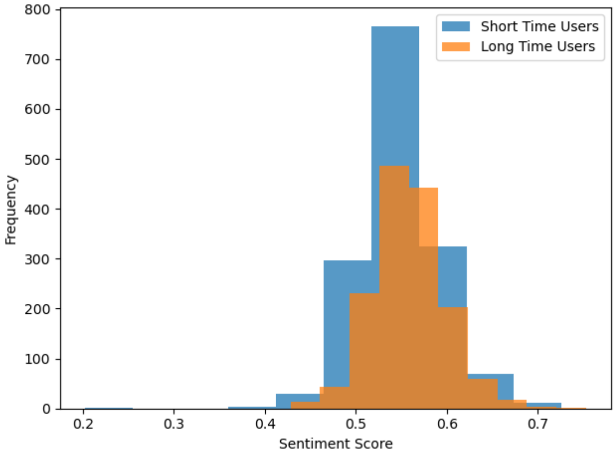

Figure 8: Account Age vs Sentiment Score

7) Likes count:

In the below graph we can see that the users who liked more of other users posts and tweets were more happy then those who liked lesser posts.
After doing the Linear Regression, the contribution of the number of likes on twitter was found to be positive 3%, this means that if I follow a lot of online communities or groups then I will have a positive impact on my mind. After doing the T-test the T-value and P-value that I received were (Tvalue= 4.9, Pvalue=6.39e-07).

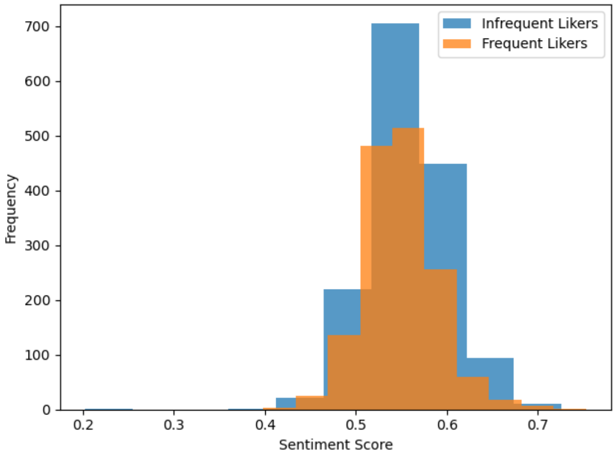

Figure 9: Posts Liked vs Sentiment Score

8) Verified:

The following graph shows that the people who were Verified on Twitter were pretty less in count, and that they tended to exhibit more happiness compared to the users who were not verified and possibly not famous.
The Linear Regression showed that, the verified status had a 1.4% contribution to the positive mental health of the twitter user, and the T-test results were (Tvalue=-4.36, Pvalue=1.28e-05).

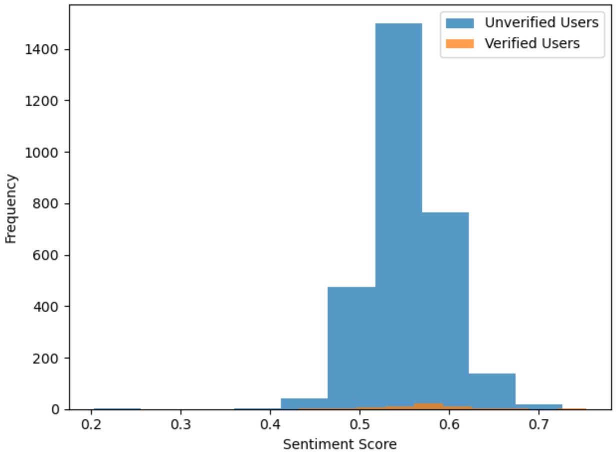

Figure 10: Verified Status vs Sentiment Score

9) Age:

The following histogram showed that older twitter users were relatively happier than the younger users.
Linear Regression showed that, the contribution of the age on twitter was found to be positive 4%. After doing the T-test the T-value and P-value that I received were (Tvalue=-7, Pvalue=2.76e-12).

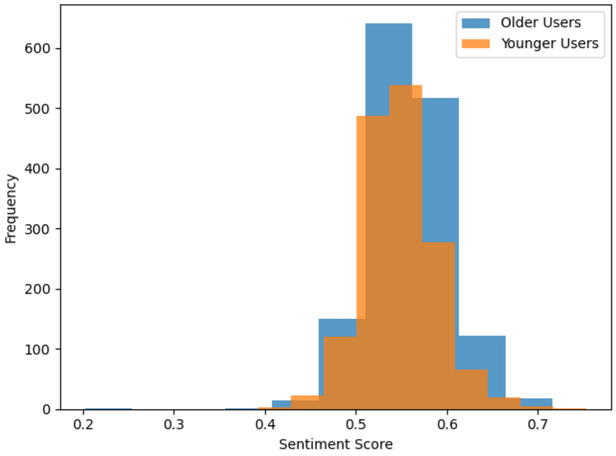

Figure 11: Age of Users vs Sentiment Score

10) Gender:

In the below graph we can see that the males on twitter are slightly happy than female users.

After doing the Linear Regression, the contribution of the gender on twitter was around negative 0.8%, which means that male users were marginally happier than female users in this aspect. After doing the T-test the T-value and P-value that I received were (Tvalue=-7.01, Pvalue=2.76e-12).

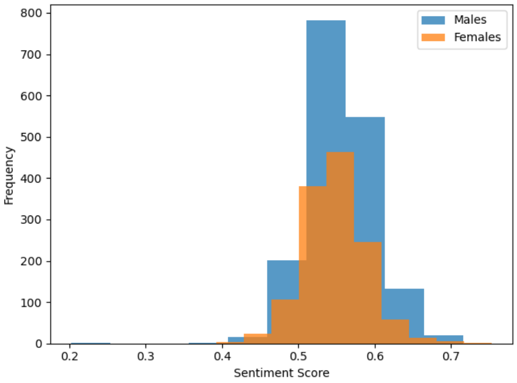

Figure 12: Gender vs Sentiment Score

11) Smile:

In the below graph we can see that the with the smile on the profile picture of the people the user seems to be slightly happy.

After doing the Linear Regression, the contribution of the smile on profile picture was around positive 0.8%, After doing the T-test the T-value and P-value that I received were (Tvalue=-9.22, Pvalue=4.98e-20).

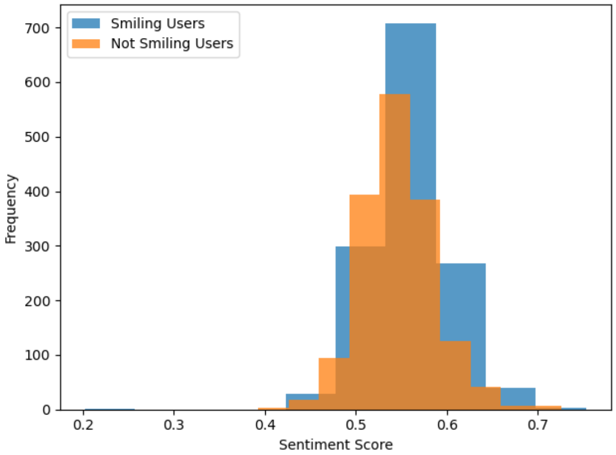

Figure 13: Smiling in Profile Picture vs Sentiment Score

12) Population Density:

In the below histogram we can see that the with the increase in the population density users are happier instead of sad.

After doing the Linear Regression, the contribution of the population density was around positive 0.8%, this means that users who are in urban areas are happier than people who are in rural areas. After doing the T-test the T-value and P-value that I received were (Tvalue=-0.88, Pvalue=0.37).

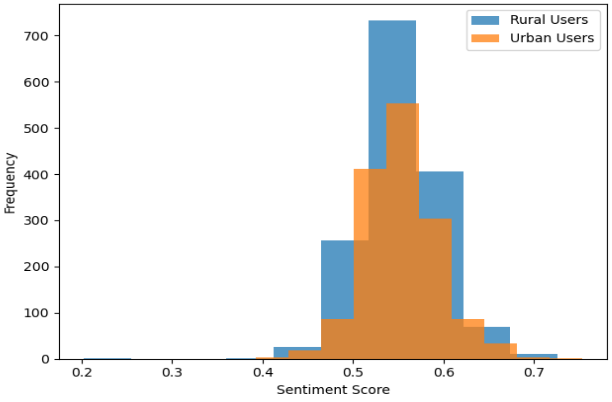

Figure 14: Population Density at User's place vs Sentiment Score

13) Water Level:

The following graph shows that, the Twitter Users were not much affected from the amount of water present in their location.

Linear Regression showed that the contribution of the water level was around negative 0.3% which is very minimal to make a conclusion. The T-test showed that the T-value and P-value were (Tvalue=-0.0056, Pvalue=0.99) which clearly shows that the correlation is completely random and water level does not have an influence on the mental health of users in the United States.

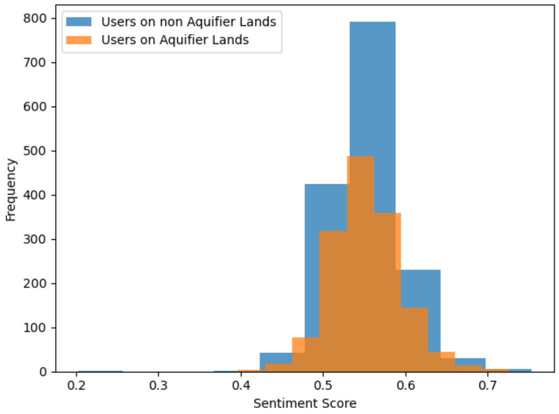

Figure 15: Water Level at User's location vs Sentiment Score

14) Household Income:

The following graph shows that there wasn't much correlation between the median income of the places where the users lived in and their sentiment score
After doing the Linear Regression, the contribution of the income was around negative 0.4%, which was very low to make any conclusion. After doing the T-test the T-value and P-value that I received were (Tvalue=0.55, Pvalue=0.57), which denotes the correlation is random.

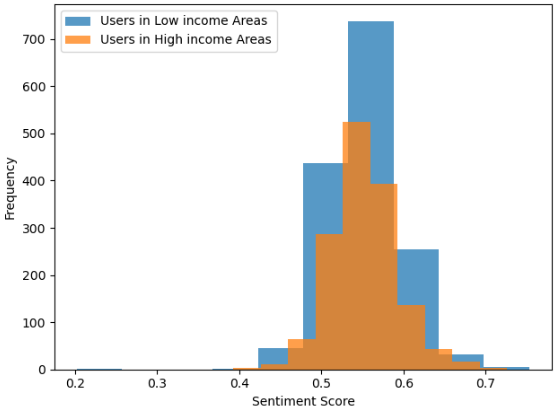

Figure 16: Median Income at user's location vs Sentiment Score

15) Parents

In the below graph we can see that the parents were slightly happier.
After doing the Linear Regression, the contribution of the parents on twitter was around positive 4%, this means that users who generally refer to their children or post about having kids are slightly happy. After doing the T-test the T-value and P-value that I received were (Tvalue=-2.74, Pvalue=0.006).

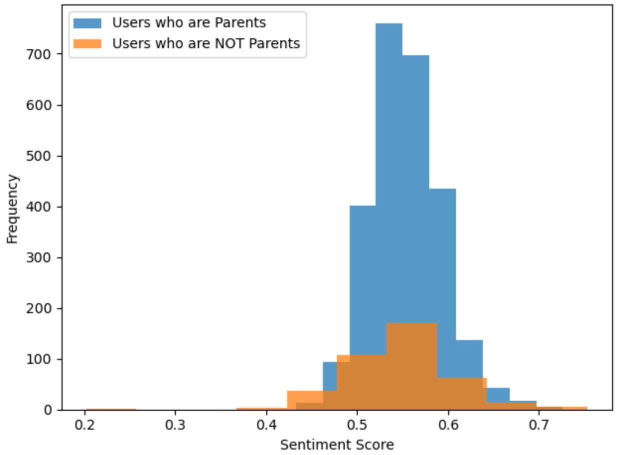

Figure 17: Parental Status vs Sentiment Score

16) Religion

The follow histogram shows that the religious status of a user did not have much influence on their sentiment score.
Linear Regression showed that, the sentiment of the religious users was around negative 0.1%, which is very low to make any conclusion After doing the T-test the T-value and P-value that I received were (Tvalue=1.29, Pvalue=0.19), which shows that the correlation is random, and a conclusion cannot be made.

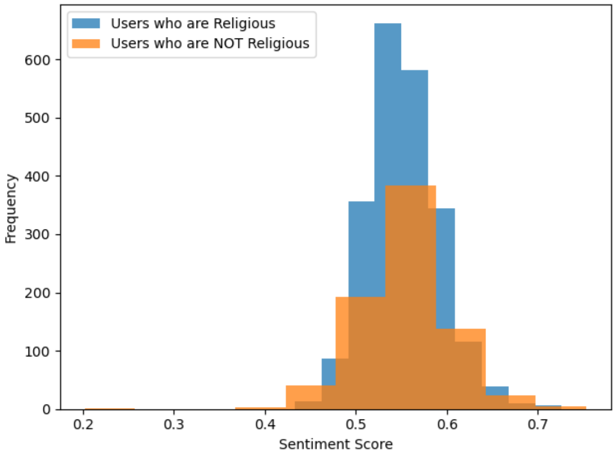

Figure 18: Religious Status vs Sentimental Score

Conclusion

I have presented the correlation of the user attributes and the LIWC features with the emotional sentiment of the user we found out that people who exhibited a lot confidence, analytical ability and those who talked about themselves were found to have the happiest sentiments.

From the user attributes we found that the people who had the most followers and those who liked the most number of post tended to have a better mental health. Users who have created their account much earlier and those who had more age tended to show better mental health online than the younger users. people who had a confident authentic tone and expressed more certainty and reasoning in their tweets showed a positive mental trait than others. It was surprising to see that the people who talked about money a lot showed better mental health, than the people who talked about their home and work. On the negative side a major discovery was users who frequently posted sexual and relationship contents were the most depressed. It was also noteworthy that the people who use the greater number of words in their tweets, the people who followed many people and followed many communities or forums showed a worser mental health than their counterpart People who showed more power and domination, negated other opinions and often used swear words showed a bad mental health, it was worth noting that users who talked about risk taking mostly ended up with a bad mental health than the others.

The remaining attributes showed lesser contribution to the mental health of the user than the rest, I hereby present the chart by showing the influence of each attribute on the overall mental health of the user.

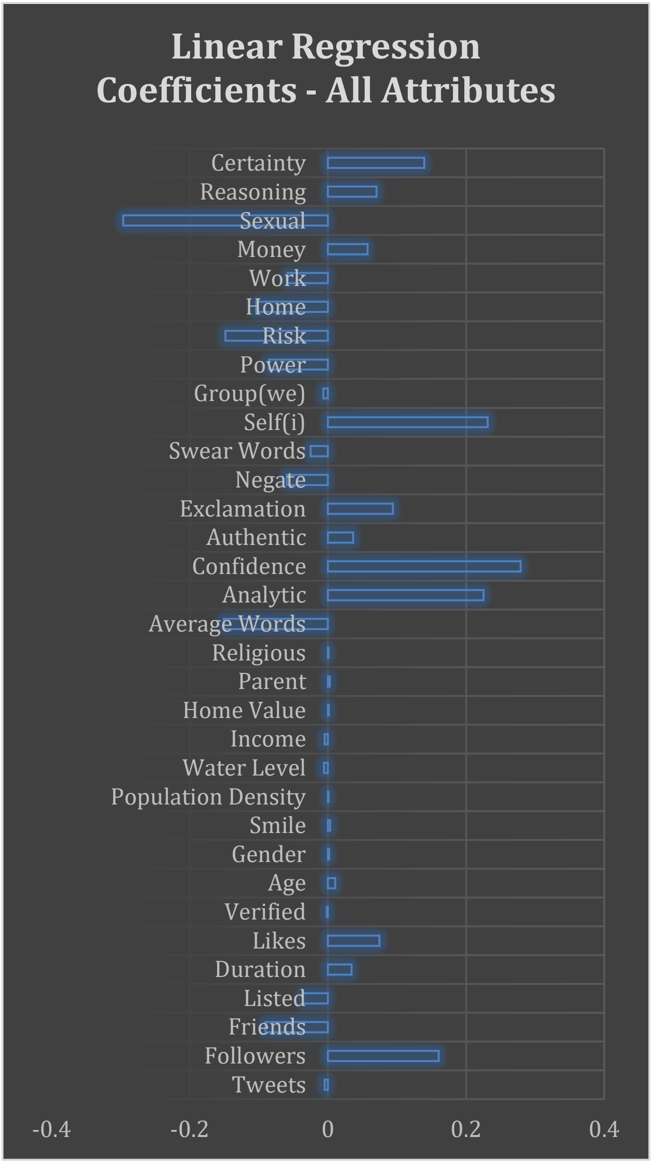

Figure 19: Linear Regression Coefficients

References

[1] Perrin, Andrew. "Social media usage." Pew research center 125 (2015): 52-68.

[2] Didegah, F., Mejlgaard, N., & SÃ¸ rensen, M.P. (2018). Investigating the quality of interactions and public engagement around scientific papers on twitter. Journal of Informetrics, 12(3), 960–971.

[3] Haunschild, R., Leydesdorff, L., & Bornmann, L. (2020). Library and information science papers discussed on twitter: A new network-based approach for measuring public attention. Journal of Data and Information Science, 5(3), 5–17.

[4] Ibrahim, N.F., & Wang, X. (2019). Decoding the sentiment dynamics of online retailing customers: Time series analysis of social media. Computers in Human Behavior, 96, 32–45

[5] Jaidka, K., Guntuku, S.C., Lee, J.H., Luo, Z., Buffone, A., & Ungar, L.H. (2021). The rural–urban stress divide: Obtaining geographical insights through twitter. Computers in Human Behavior, 114, 106544.

[6] Hassan, S.-U., Saleem, A., Soroya, S.H., Safder, I., Iqbal, S., Jamil, S., Bukhari, F., Aljohani, N.R., & Nawaz, R. (2020). Sentiment analysis of tweets through altmetrics: A machine learning approach. Journal of Information Science, 0165551520930917.

[7] Narr, S., Hulfenhaus, M., & Albayrak, S. (2012). Language-independent twitter sentiment analysis. Knowledge discovery and machine learning (KDML), LWA, pp. 12–14.

[8] Bae, Y., & Lee, H. (2012). Sentiment analysis of twitter audiences: Measuring the positive or negative influence of popular twitterers. Journal of the American Society for Information Science and technology, 63(12), 2521–2535.

[9] Zaman, T.R., Herbrich, R., Van Gael, J., & Stern, D. (2010). Predicting information spreading in twitter. In Workshop on computational social science and the wisdom of crowds, nips, volume 104, pp. 17599–601. Citeseer

[10] Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In Proceedings of European Language Resource Association, volume 10, pp. 1320–1326.

[11] Wang, X., Gerber, M.S., & Brown, D.E. (2012). Automatic crime prediction using events extracted from twitter posts. In International conference on social computing, behavioral-cultural modeling, and prediction, pp. 231–238. Springer.

[12] Achrekar, H., Gandhe, A., Lazarus, R., Yu, S.-H., & Liu, B.Y. (2011). Predicting flu trends using twitter data. In 2011 IEEE conference on computer communications workshops (INFOCOM WKSHPS), pp. 702–707. IEEE.

[13] Mittal, A., & Goel, A. (2012). Stock prediction using twitter sentiment analysis. Standford University, CS229. Retrieved from http://cs229.stanford.edu/proj2011/GoelMittal-StockMarket PredictionUsingTwitterSentimentAnalysis.pdf

[14] Guntuku, Sharath Chandra, et al. "Studying expressions of loneliness in individuals using twitter: an observational study." BMJ open 9.11 (2019): e030355.

[15] Lyu, Hanjia, et al. "Social media study of public opinions on potential COVID-19 vaccines: informing dissent, disparities, and dissemination." Intelligent medicine (2021).

[16] Chen, Long, et al. "Fine-grained analysis of the use of neutral and controversial terms for COVID-19 on social media." International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation. Springer, Cham, 2021.

[17] Pennebaker, James W., et al. The development and psychomet-ric properties of LIWC2015. 2015.

[18] Hutto, Clayton, and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." Proceedings of the International AAAI Conference on Web and Social Media. Vol. 8. No. 1. 2014.