EDA Mini Project

Topic: Patterns of Social Media Usage and Emotions

Objectives

This project aims to find the correlations of one's social media usage patterns, together with their emotional states, so as to figure out whether our stereotypes about social media are valid. To achieve this goal, the following objectives are established:

1. To search for dataset(s) that captures the information mentioned above
2. To preprocess the dataset(s) such that it is more comprehensible
3. To conduct EDA and compare the results with our assumptions

Assumptions

Below are the assumptions we have had:

1. Younger people are more invested in and driven by social media
2. People who spend more time on social media are more unhappy
3. Some social media are more addictive than others
4. Some social media cater for certain gender(s) only
5. Using different social media can give rise to different emotions
6. Girls and non-binary people spend more time on social media than boys
7. Number of likes have nothing to do with happiness

Let's go ahead and see if our assumptions are right!

Dataset Description

We have found a dataset on Kaggle which encompasses variables that fit well with our project theme. To specify, it consists of age and gender of the users / interviewees; social media platforms they usually use; daily usage time (daily time they spend on that particular platform); number of messages sent, posts made, likes and comments received per day, as well as their dominant emotions (users' everyday dominate emotional states including happiness, neutral, boredom, anxiety, sadness and anger).

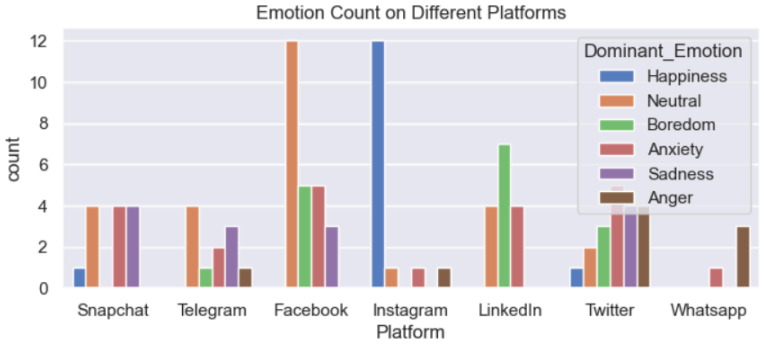
Data Preprocessing

Before digging into the analyzing part, it is important for us to modify the dataset in order to facilitate the analysis process. First of all, we have discovered some flaws in the dataset which could hamper our work, so we need to get rid of them. Specifically, there is a mistake in row 46, where the age '27' and the gender 'male' are misplaced from each other; and another one where a 'Marie' tag is located in the gender column. We decide to keep the first mistaken datum (row) by simply swapping the tags and discard the second one, since we do not know what exactly 'Marie' implies. Besides 'Marie', we have also dropped the 'User_ID' column, believing it has nothing to do with our project. Finally, we have added two columns in the dataset, namely 'Daily_Usage_Range', converting the values in 'Daily_usage_Time (minutes)' into ranges separated by 20 minutes; and 'Emotion_Positivity', in which we have scored different emotions according to their positivity in hopes of quantifying the emotional state to yield more detailed analysis. ('Happiness : 1, 'Neutral' : 0, 'Boredom' : -0.5, 'Anxiety' & 'Sadness' & 'Anger' : -1)

Analysis

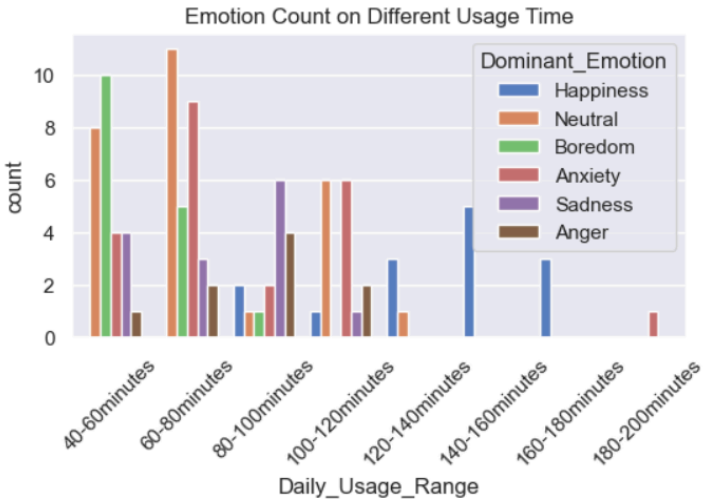
Here show the findings of what we have analyzed through the dataset.

a) Emotion Count on Different Platforms



First of all, we have plotted a bar chart to see what kinds of emotions different social media platforms tend to yield. Interestingly, Instagram is seemingly the “happiest” platform, where 12/15 (80%) of its users feel “happy” in their everyday lives. On the other hand, Twitter and Whatsapp are the most negative platforms, with 13/19 (68%) and 4/4 (100%) respectively of their users feeling either anxious, sad or angry. Meanwhile, boredom of LinkedIn’s users is spotted, accounting for 7/15 (47%) of the count.

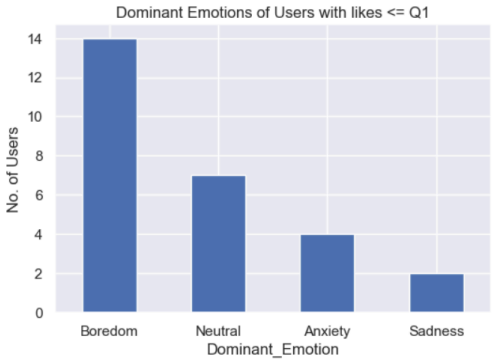
b) Emotion Count on Different Usage Time

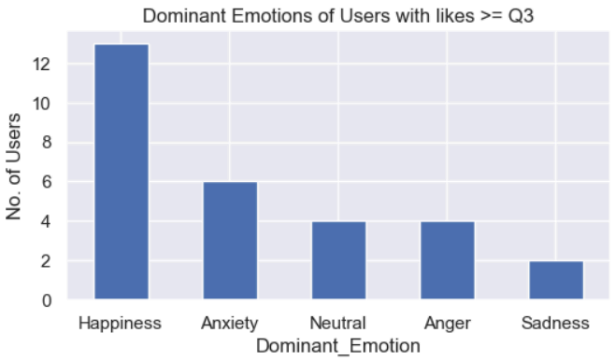


The next thing we are interested to know is about how the time spent on social media affects our emotions (or vice versa). From what we see on this bar chart, people who spend more time tend to be happier: out of the 13 people who spend more than 2 hours daily, 11 of them actually feel happy, who make up for 85% in this group. However, it is noted that the very individual in this dataset who spends more than 3 hours daily is dominated by anxiety.

On the other hand, those with less time spent are way more negative: among all users within 40 to 80 minutes, none of them are happy, while the tallest bars of boredom and anxiety are found in this group as well, which account for 10/16 (63%) and 13/22 (59%) of the total counts of these two emotions respectively. That being said, quite a few of them possess a neutral mind, which may imply one is less driven by social media if they spend less than 80 minutes on it.

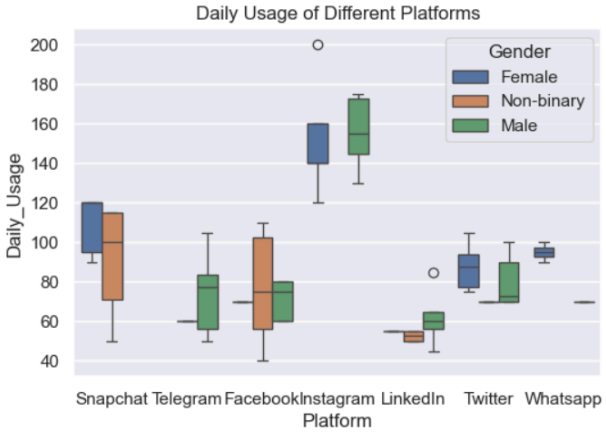
c) Relations of Like Received and Emotions





We then proceed to learn the relation between number of likes and dominant emotions. The bar chart on the left represents users within 25% percentile in terms of the number of likes received (i.e. not more than 15 likes), whereas the one on the left is those within 75% (i.e. not fewer than 40 likes). By comparing the two, a big difference that immediately came into sight is that there is no count of happiness on the left, suggesting none in this 25% percentile group feels happy in daily life. Meanwhile, nearly half (13/29) of the 75% percentile group claim to be happy. In contrast, none in this group consider themselves bored, while more than half (14/27) of the 25% percentile group do. Although people with more likes seem to be happier, we are not neglecting the fact that they are more likely to feel low as well. As we can see, the count for rather negative feelings (anxiety, sadness and anger) from more-like users is 12, which is twice the size of that from fewer-like users.

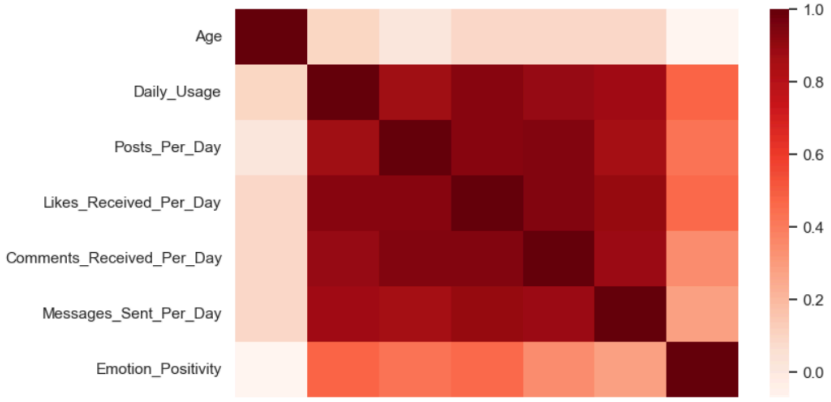
d) Daily Usage of Different Platforms



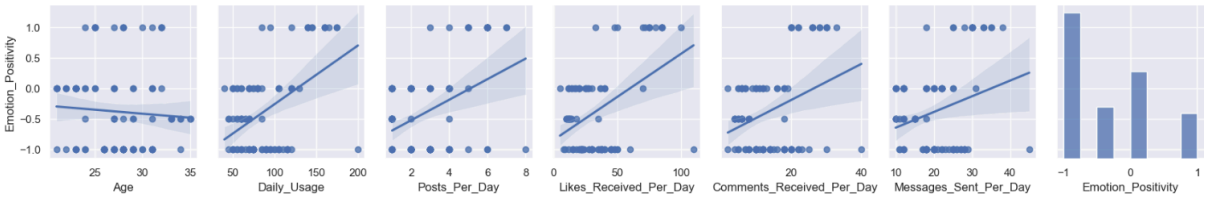
The fourth piece of information we are about to look at is the box plots depicting the spans of daily usage in different social media. We have also divided the plots into different genders in the interest of the way they are involved. The first thing catching our eyes is how Instagram prevails over all other platforms in terms of the usage time, of which the minimum time spent here in this dataset, 120 minutes, is already equal to the maximum from the other platforms. On the other hand, People tend to spend less time on LinkedIn, whereas its very outlier on the high end from the male plot has not even surpassed 6 medians from the other plots. Another thing that we have observed is certain platforms are apparently not that popular among specific gender groups, say, there are no male, female and non-binary participants in Snapchat, Telegram, and Instagram and Whatsapp respectively. Lastly, the spans of usage time for different gender groups are pretty much aligned with one another in a single platform, suggesting no key distinction of the time spent on social media regarding different

genders. As for the distributions of the box plots, we are not able to obtain any useful information due to the constraint of sample size.

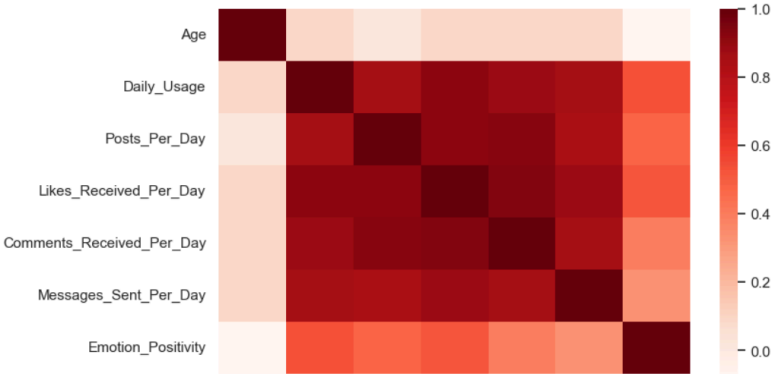
e) Correlations of Features



Finally, we will look into the correlations among different quantitative features in this dataset. We can see the middle area is filled with the deep color, indicating that daily usage, posts, messages sent, likes and comments received are highly and positively correlated. On the contrary, the color of age is quite light when put against other features, which means age has basically nothing to do with them. As for emotion positivity, its correlation coefficients with other features vary within 0.2 to 0.5, showing rather weak or moderate correlations. Let's see if we can grasp a better idea from its paired regression plots with other features.



When we look carefully into the 5 graphs in the middle, it's not hard to realize that there has always been an outlier at the bottom right corner, indicating a negative emotion shared by an individual who has the highest daily usage and greatest numbers of posts, messages sent, likes and comments received per day. Let's remove it for now and see how the results differ:



We can see the colors for emotion positivity have slightly grown deeper, despite not much. At the end of the day, we have decided to reserve this outlier, because it is a vital figure

representing the negative emotion from the individual being invested the most in social media in this dataset. Therefore, it is determined that the positivity of emotion has rather weak correlations with other features.

Results and Conclusion

After analyzing the data, we can now check whether our earlier assumptions are correct. First, our heatmap has proven the irrelevancy between ages and level of involvement with social media, so younger people are not more invested in and driven by social media. Second, according to our bar chart of emotion count on different usage time, people who spend more time on social media are generally happier, while our pair plot hasn't shown a clear correlation between these two. But either way, the second assumption is false. What's next is whether some social media are more addictive than others and cater for certain genders only, and the answer is both true, according to the box plots. It is also true that people who use different social media have different emotions, but we can not prove the causation relationship. How about boys spending less time on social media than other genders? The answer is no, since we cannot see such a trend on the box plots of any platforms. Finally, we are not sure whether likes have anything to do with happiness, as the heatmap does not suggest a tight correlation while tha bar chart shows that nearly half of the users who receive 40 likes or more daily are happy. By and large, most of our stereotypical views on social media are proven incorrect.

Limitations and Future Works

There are two main limitations faced during the project. The first is regarding the inadequacy of data in the dataset we have chosen, which refers to both 1) the small sample size, which contains only about 100 rows of data, and 2) the limited sample species, say the age from the pool is only ranged from 21 to 35 years old, and there are only 7 platforms available. These limitations have given rise to the pretty unconvincing results we have come up with, and even certain not fully functional graphs, such as the box plots we have illustrated. We have tried to search for other datasets with more rows and more abundant data species for merging or even replacement, but they do not blend well with our topic, and eventually we have given up the search due to the time limit. The second challenge is about how to interpret the correlations between the dominant emotions and other quantitative factors in a reasonable way. In this project, we have resorted to scoring different emotions with either 1, 0, -0.5 and -1 points, despite knowing it is not quite a rigorous way to handle this, and, as expected, the correlations we have drawn based on this model is likely to be inaccurate, since it contradicts with the result we have acquired by qualitative means. But still, we have to use it as we are not familiar with any professional tricks in statistics which can pull it off. For future work, we would like to replace the current dataset with a more robust one, which contains at least 300 sets of data, involves wider age groups and choices of social media, as well as some sorts of quantitative indicators for evaluating emotions, such as stress level from 1 to 10, or laughing frequency in daily life etc.

Reference

https://www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being