Course Number: CMPS 620 *Instructor: Professor Tweneboah*

Social Media Usage and its Relationship with Emotional Well-Being

Maxwell Bilyk[1]*

[1] Masters Student; mbilyk@ramapo.edu

Abstract: "67% of adolescents report feeling worse about their own lives as a result of their social media use." Research like this and hints towards some type of relationship between social media and emotional well-being. Using data from AI researcher Emirhan Bulut, an in depth analysis was used to explore this relationship. The analysis contains exploratory data analysis, a classification model to predict dominant behavior based on social media usage and a clustering algorithm to determine if clusters existed within our data. From the analysis, it was determined that people who used instagram logged more time spent on their devices, LinkedIn users with a lot time spent on the app are almost always reporting their emotion as angry, we can very accurately predict someone's emotion based on their social media usage with features like age, time spent on the platform and likes received as our most important predictors and some distinct clusters do exist within our data.

Keywords: Classification 1; Random Forest 2; K-Means Clustering 3

1. Introduction

Social media has become the most popular technology of our generation, allowing users to create, share and interact with information from people all over the world. As of 2023, an estimated 4.9 billion people use social media globally and these numbers are only expected to increase. With so many people using these platforms, it is important to research their effect on our well being. 39% of social media users report that they are addicted to social media while 67% of adolescents report feeling worse about their own lives as a result of their social media use.[1] These statistics emphasize some sort of relationship between social media usage and emotional well-being. To help study this relationship, AI Inventor Emirhan Bulut prepared a survey based dataset. It captures valuable information on social media usage and the dominant emotional state of users based on their activities. The dataset holds information like age, gender, platform, time spent on the platform and the surveyees dominant emotion throughout the day.

2. Materials and Methods

Data sources included social media usage data from platforms such as Facebook, Twitter, Instagram, and Reddit, and emotional well-being data collected via self-reported surveys. The social media data, spanning from January 2023 to June 2023, included metrics like daily usage time, number of posts, likes, comments, age and platform used. Python (version 3.9) was used for the analysis, leveraging libraries such as pandas and numpy for data manipulation, matplotlib and seaborn for data visualization, Scipy and statsmodels for statistical analysis, and scikit-learn for machine learning. Jupyter Notebook served as the integrated development environment.

The dataset came from a kaggle repository [2] that collected survey data which asked people about their age, gender, daily usage time on apps, likes received, posts

posted, comments received and messages sent and on which platforms and lastly their dominant emotion throughout the day. The preprocessing stage involved cleaning data by imputing missing values using median and mode, and identifying and removing outliers using the interquartile range (IQR) method. Social media usage data and emotional well-being survey responses were merged based on unique participant identifiers and normalized to account for variations in usage patterns across different platforms.

3. Results

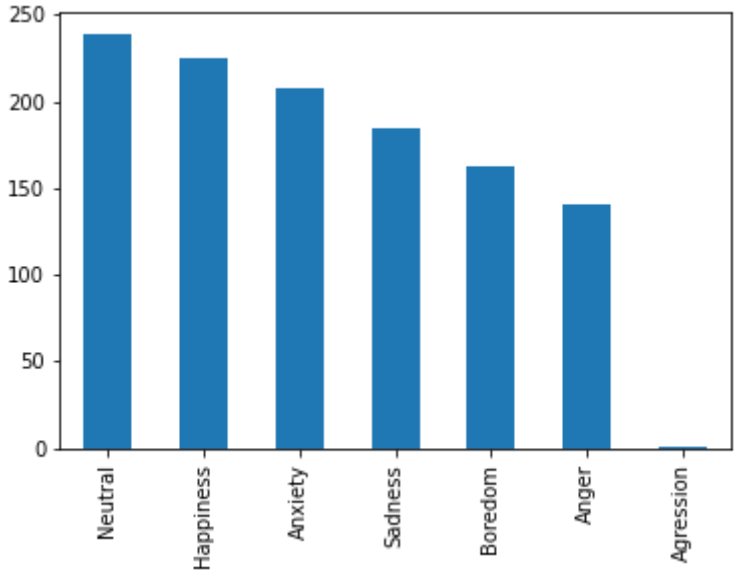

Figure 1. This figure shows the distribution of the target variable (Dominant Emotion) from the dataset.

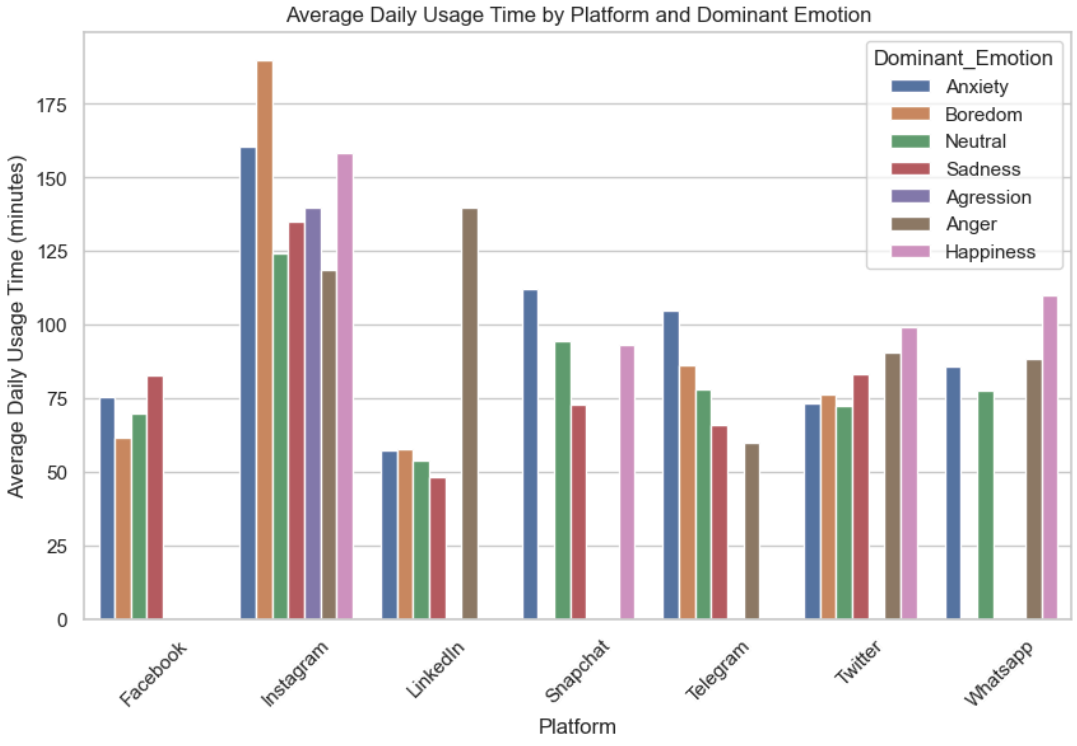

Figure 2. This figure is a grouped bar plot that is grouped by social media platform, colored by dominant emotion and scaled on average daily usage time in minutes.

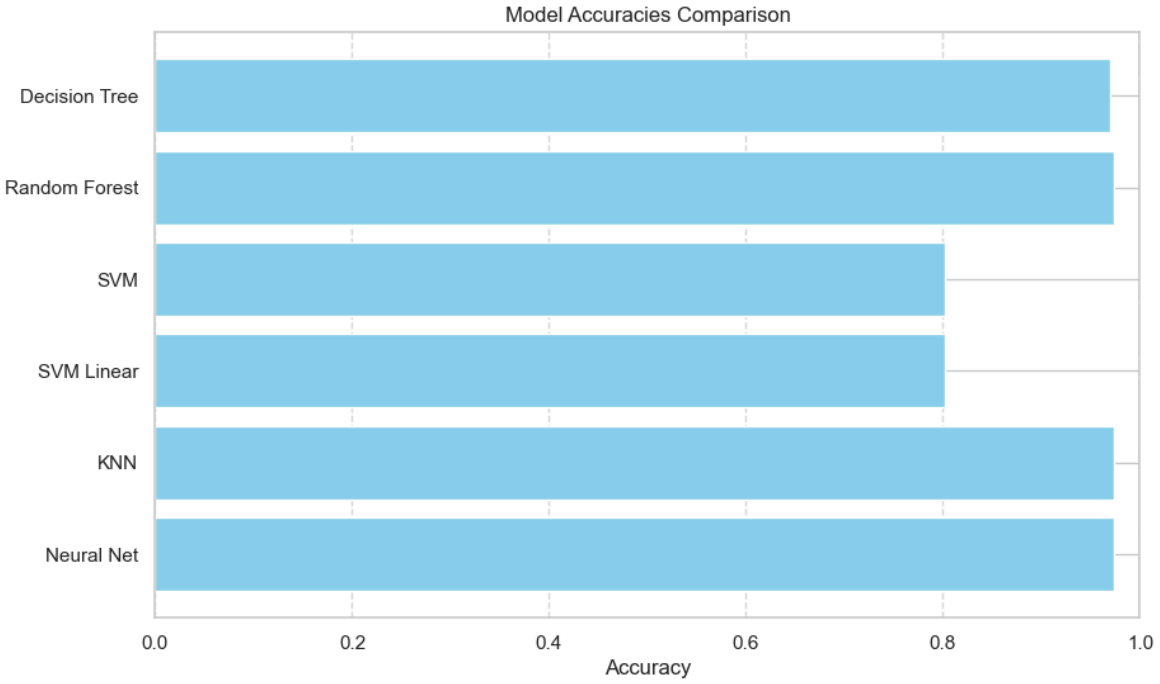

Figure 3. This figure shows a barplot comparing the accuracy score of different classification models.

Table 1. This table shows the different performance metrics from a random forest classification model on the different classes from the target variable.

Emotion	Precision	Recall	F1-Score	Support
Anger	0.90	0.96	0.93	27
Anxiety	1.00	0.91	0.96	35
Boredom	1.00	1.00	1.00	38
Happiness	0.96	0.98	0.97	50
Neutral	1.00	1.00	1.00	51
Sadness	0.97	0.97	0.97	31

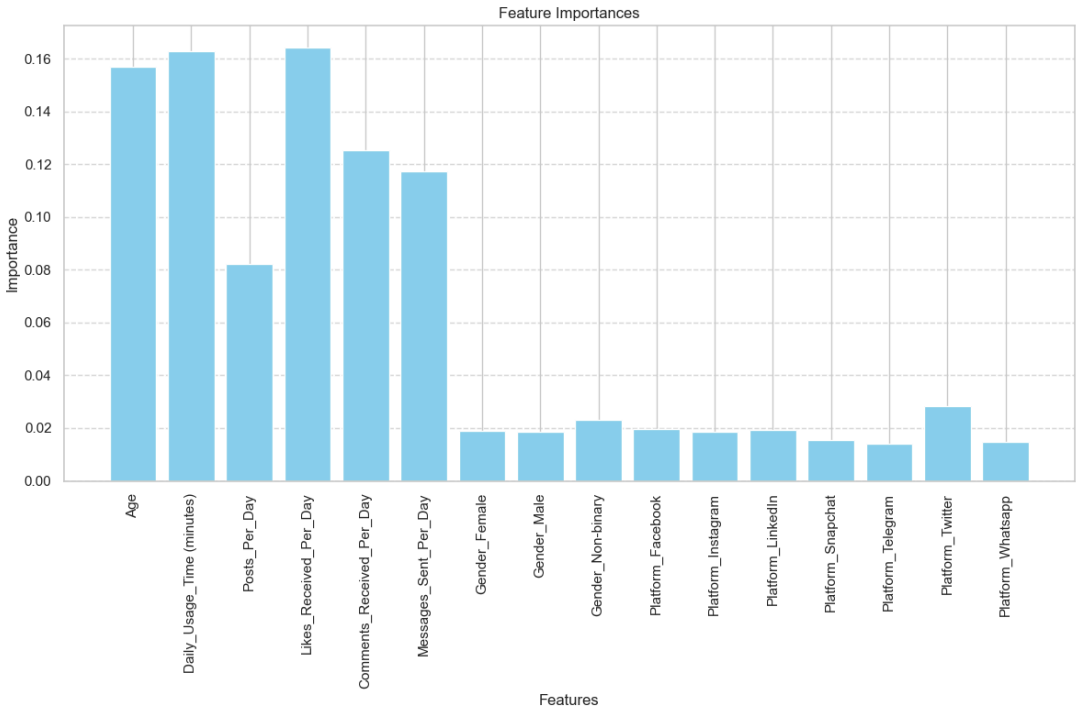

Figure 4. This figure is a feature importance plot which compares different features from the dataset and their relative importance score in terms of predictive power added to the classification model.

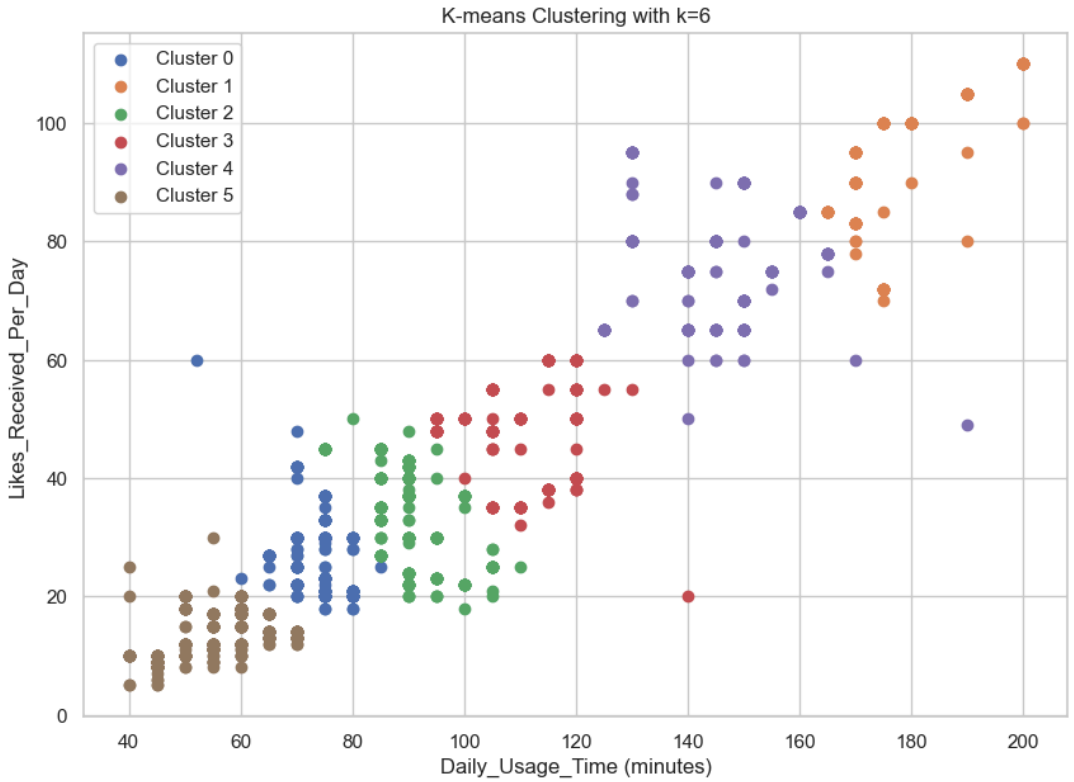

Figure 5. This figure shows a scatterplot of points comparing daily usage time and likes received per day, colored by clusters determined by K-Means clustering algorithm.

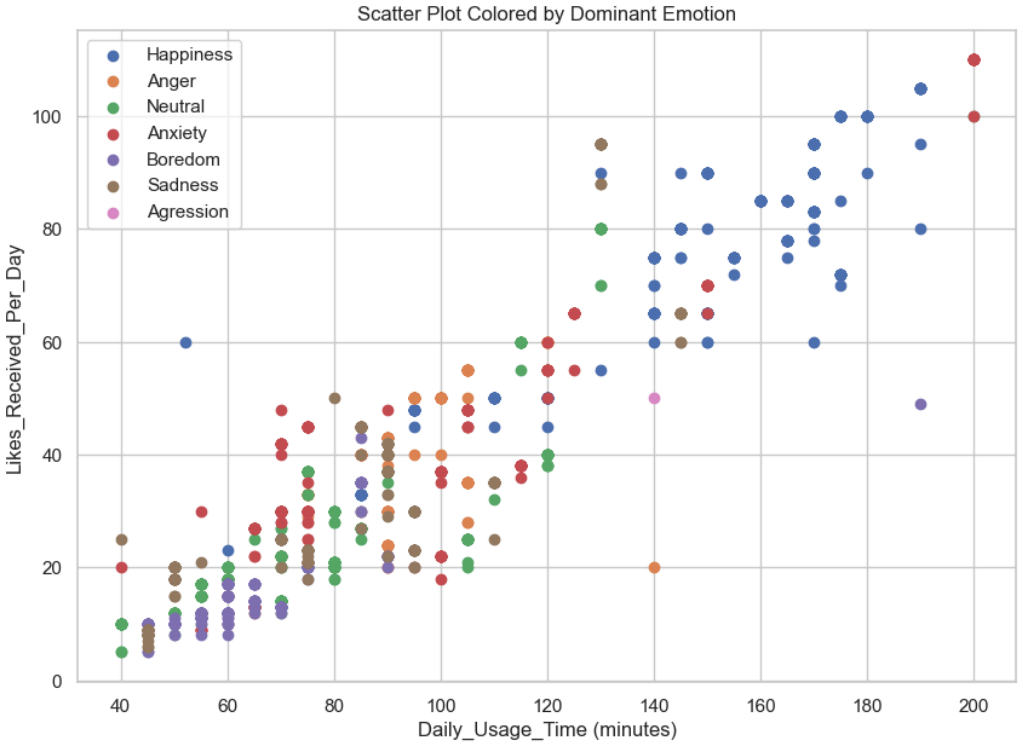

Figure 6. This figure shows a scatterplot comparing daily usage time and likes received per day and is colored by dominant emotion throughout the day.

4. Discussion

From **Figure 2.** we can see that every platform has users which reported their dominant emotion as anxiety. further investigating the same figure we can all see that instagram users tend to stay on longer in comparison to other platforms. Furthermore, LinkedIn users that have a lot of time spent on the app (past 60 minutes) almost always report anger. This is likely because this group of individuals is on LinkedIn looking for jobs, and the longer they are looking for a job they will grow angrier and angrier from rejection and lack of opportunity. Instagram was the only app to have a surveyee report their emotion as aggressive. Snapchat users did not report being bored at all which makes sense since it is more of a messenger app so people have no reason to be on it without messages to answer, and if they are answering messages, they are talking to friends and not being bored.

Observing **Figure 1.** we can see that the distribution of our target variable is not exactly uniform but also not super imbalanced to any specific group, however there is only one entry of aggression in our dataset so that class is severely underrepresented.

Looking at **Figure 3.** We can see that a lot of classification models predicting dominant behavior perform very well on this dataset. However the support vector machine models did not perform well. The final model picked was Random Forest model with a maximum depth of 10 leafs because it was right up there at the top in terms of accuracy, but unlike some of its competitors, it has the ability to explain some of its reasoning for prediction through its important feature.

Table 1. shows the performance of the random forest model on all the different classes from the target variable. The model ended up performing very well on all classes, so we don't have to worry about any unproportionally inaccurate class predictions, minus the aggression class which did not have enough instances to train. Furthemore, this random forest model also went through k-fold cross validation with 5 splits and

reported an average accuracy of 0.972 between the different iterations, which gives a lot more confidence that the model is actually performing well on unseen data.

Figure 4. shows the feature importance from the random forest model. The most important features in terms of predicting someone's behavior are daily usage time,how many likes they received and their age. This is very interesting. With the two most important features in terms of predicting their daily usage time and likes received, we can conclude that there is a lot of correlation between social media usage and the effect on their overall emotion throughout the day. With age also being an important factor, we can conclude that people with similar social media usage can have very different emotions based on their age. It is very interesting to see that a person's gender or the app they were using really matters in comparison to the other features. This implies that no specific apps are overly influencing anyone's behavior, but it is more based on how they are interacting with their platforms of choice.

Figure 5. and **Figure 6.** both show the positive linear relationship between daily usage time and likes received per day. However **Figure 5.** is colored by clusters made from a k-means clustering algorithm. When taking the same graph and instead coloring the points by dominant emotion like in **Figure 6.** we can see that some clusters are very similar because of their emotion reported. For example when cross referencing both graphs you can deduce that cluster 1 and 4 are clusters of our happy surveyees, while cluster 5 seems to contain more bored and neutral reporting individuals.

5. Conclusions

To conclude, the above research indicates that we can confidently state there is a distinct relationship between social media usage and emotional well-being. The exact relationship does not have a general explanation, but rather on an individual to individual basis, we can predict a person's dominant emotion throughout the day by knowing some of their demographic information in tandem with their social media usage.

Some key insights stated were, every app had reports of surveyees with anxiety. LinkedIn users that were on the platform for more than an hour a day were almost always angry, distinct clusters do exist within our data and they seemed to be clustered by their dominant emotions respectively.

Data Availability Statement: Data set for this project can be accessed here: https://www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being?select=train.csv

References

1. Agrawal, Sumeet Kumar. "Metrics to Evaluate Your Classification Model to Take the Right Decisions." Analytics Vidhya, 5 June 2024, www.analyticsvidhya.com/blog/2021/07/metrics-to-evaluate-your-classification-model-to-take-the-right-decisions/#:~:text=Classification%20Metrics%20like%20accuracy%2C%20precision,in%20evaluating%20the%20model%20performance.
2. Bulut, Emirhan. "Social Media Usage and Emotional Well-Being." Kaggle, 19 May 2024, www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being?select=train.csv.
3. "Pandas.Pivot_table#." Pandas.Pivot_table - Pandas 2.2.2 Documentation, pandas.pydata.org/docs/reference/api/pandas.pivot_table.html. Accessed 5 July 2024.
4. Raj, Shivam. "Effects of Multi-Collinearity in Logistic Regression, SVM, RF." Medium, Medium, 18 June 2020, medium.com/@raj5287/effects-of-multi-collinearity-in-logistic-regression-svm-rf-af6766d91f1b#:~:text=Random%20Forest%20uses%20bootstrap%20sampling,different%20set%20of%20data%20points.
5. Wong, Belle. "Top Social Media Statistics and Trends of 2024." Forbes, Forbes Magazine, 12 Apr. 2024, www.forbes.com/advisor/business/social-media-statistics/.