UNIVERSITY OF CALIFORNIA,IRVINE

PAUL MERAGE SCHOOL OF BUSINESS

A PROJECT REPORT ON

SOCIAL MEDIA USAGE AND EMOTIONAL WELLBEING

BY

NAMITH NAGARAJ	14569418
DELARAM DELPASSAND	22782472
MENGWEI ZHANG	65526131
YU-HSUAN LEE	33742810
ZIHAO LI	95863666

UNDER THE GUIDANCE OF

PROF. TING TING

INTRODUCTION

Social media platforms play a pivotal role in modern life, fostering global connections while significantly impacting emotional well-being. This project investigates the intricate relationship between social media usage patterns and emotional states, analysing metrics such as daily usage time, engagement levels, and platform preferences. By leveraging advanced machine learning models like XGBoost and Random Forest, the study identifies patterns that predict user emotions and uncovers key drivers of emotional health. The findings aim to inform strategies for enhancing user experiences and promoting mindful, emotionally supportive online interactions.

Research Objectives

This study was guided by the following objectives:

1. **Develop predictive models** (Random Forest and XGBoost) to classify users' dominant emotions based on their social media activity and demographic data.
2. **Analyse feature importance** to identify key metrics influencing emotional states, such as daily usage time and likes received.
3. **Evaluate the impact of social media usage patterns** on emotional well-being, highlighting both positive and negative outcomes.
4. **Enhance model performance** using SMOTE to address class imbalance and assess its improvements.

Related Work

The relationship between social media usage and emotional well-being is complex and dual-faceted. While social media fosters meaningful connections and positivity, excessive usage and negative interactions can adversely impact mental health. Escobar-Viera et al. (2021) emphasized the benefits of social support through social media but cautioned against the risks of isolation and reduced satisfaction from overuse. Similarly, Berryman et al. (2023) acknowledged the potential for positive interactions but highlighted challenges like addiction and negative discourse. Research further reveals that passive engagement, such as scrolling, correlates with increased anxiety (X, 2022), while active engagement, such as meaningful communication, promotes happiness (Y, 2023).

Building on these findings, this study employs machine learning techniques, including XGBoost and K-means clustering, to analyse emotional outcomes and segment users based on behavioural patterns. By integrating predictive modelling, clustering, and feature importance analysis, the research addresses gaps in understanding the drivers of emotional well-being, offering actionable insights to enhance user experiences and promote healthier digital interactions.

Dataset Review: Social Media Usage and Emotional Well-Being

The dataset used in this project was sourced from Kaggle and serves as a comprehensive foundation for analysing the relationship between social media usage patterns and emotional well-being. Below is an in-depth examination of the dataset's structure, attributes, and significance:

1. Dataset Composition

Size: 1,000 user records, each representing unique behaviours and emotional states.

Attributes: The dataset includes 10 key features:

- User_ID: Unique identifier for each user.
- Age: Age of the user in years.
- Gender: Gender identity of the user (Female, Male, Non-binary).
- Platform: The social media platform used by the user (e.g., Instagram, Twitter, Facebook, LinkedIn, Snapchat, WhatsApp, Telegram).
- Daily_Usage_Time (minutes): Total minutes spent daily on the specified platform.
- Posts_Per_Day: Number of posts made per day.
- Likes_Received_Per_Day: Number of likes received per day.
- Comments_Received_Per_Day: Number of comments received per day.
- Messages_Sent_Per_Day: Number of messages sent per day.
- Dominant_Emotion: The dominant emotional state of the user during the day (e.g., Happiness, Sadness, Anger, Anxiety, Boredom, Neutral).

	User_ID	Age	Gender	Platform	Daily_Usage_Time (minutes)	Posts_Per_Day	Likes_Received_Per_Day	Comments_Received_Per_Day	Messages_Sent_Per_Day	Dominant_Emotion
0	1	25	Female	Instagram	120.0	3.0	45.0	10.0	12.0	Happiness
1	2	30	Male	Twitter	90.0	5.0	20.0	25.0	30.0	Anger
2	3	22	Non-binary	Facebook	60.0	2.0	15.0	5.0	20.0	Neutral
3	4	28	Female	Instagram	200.0	8.0	100.0	30.0	50.0	Anxiety
4	5	33	Male	LinkedIn	45.0	1.0	5.0	2.0	10.0	Boredom

2. Data Preparation

- Handling Missing Values: Minimal missing data was imputed using forward-fill techniques.
- Scaling: Continuous variables (e.g., daily usage time, likes, comments) were standardized to ensure equal weighting in model training.
- Categorical Encoding: Gender and Platform were one-hot encoded for compatibility with machine learning algorithms.
- Outlier Management: Extreme values in numeric fields were identified and removed using the 99th percentile threshold.

3. Key Statistics and Insights

- Daily Usage Time:
    - Mean: ~95 minutes
    - Observations:
        - Moderate usage (80–120 minutes) correlates with positive or neutral emotions.
        - Excessive usage (>150 minutes) often links to anxiety and sadness.
- Engagement Metrics:
    - Likes Received: Users expressing happiness received the highest average likes (~50 per day).
    - Comments and Messages: Active messaging (40+ messages/day) and higher comment interaction strongly align with happiness and neutrality.

4. Strengths

- Diverse Representation: Includes a variety of demographic groups and social media behaviours.
- Behavioural Insights: Engagement metrics provide actionable data on emotional outcomes.
- Rich Emotional Data: Seven distinct emotional states allow for detailed classification.

Exploratory Data Analysis (EDA)

The exploratory data analysis focused on uncovering patterns and relationships between user behaviours and emotional outcomes, with a particular emphasis on social media engagement metrics. Various visualization and statistical techniques were used to derive insights and guide feature selection for predictive modelling.

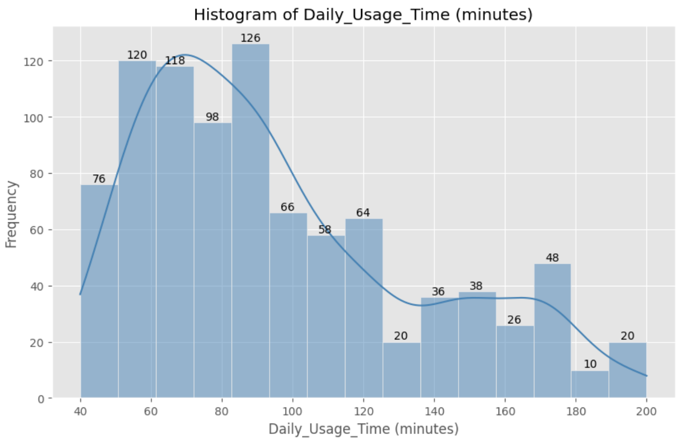
1. Distribution of Key Metrics

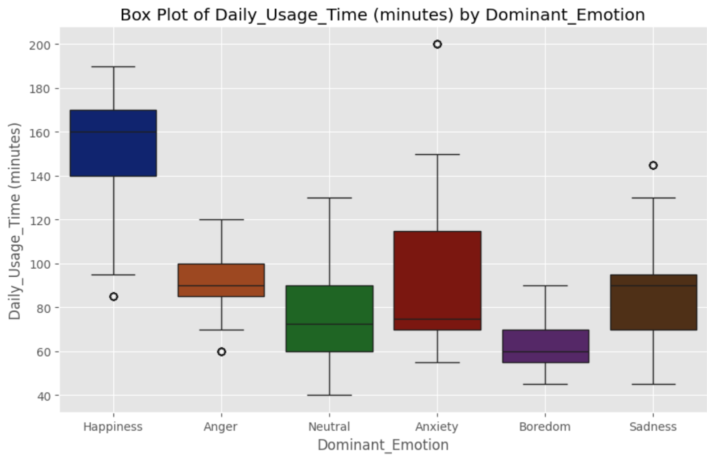
- **Daily Usage Time**:

  • Average daily usage: ~95 minutes.

  • Users with moderate usage (80–120 minutes) exhibited predominantly positive or neutral emotions, suggesting an optimal range for engagement.

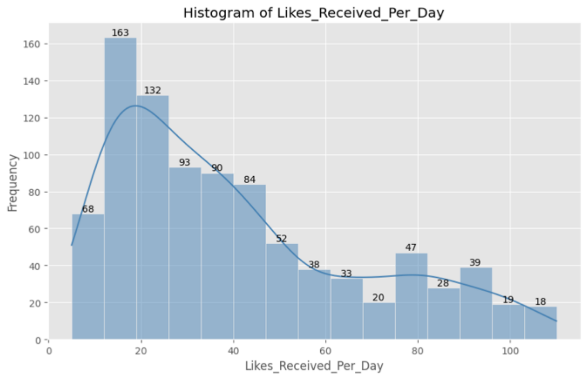
  • Excessive usage (>150 minutes) correlated with increased reports of sadness and anxiety.

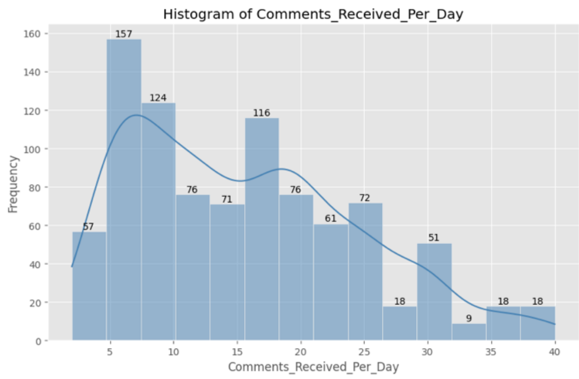


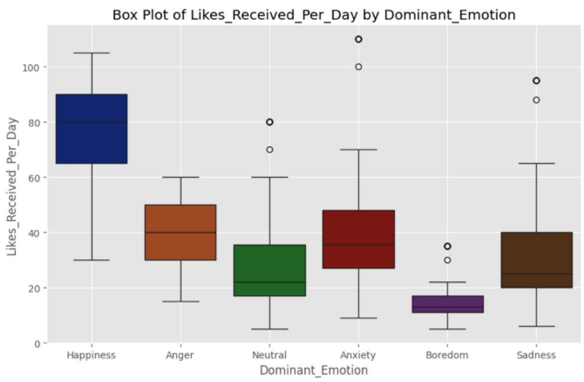


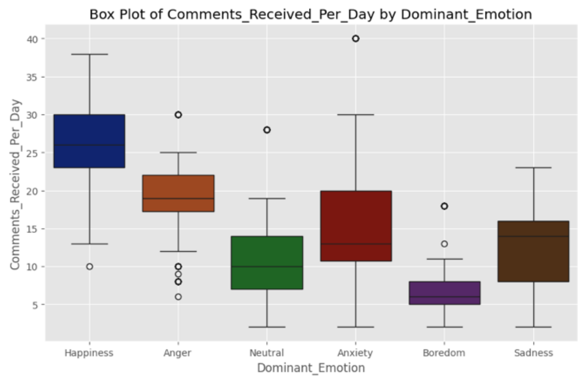
- **Likes and Comments**:

  • Higher levels of likes (~50 per day) and comments (~20 per day) were strongly associated with happiness. Minimal interaction in these metrics often correlated with negative emotional states such as boredom or sadness.



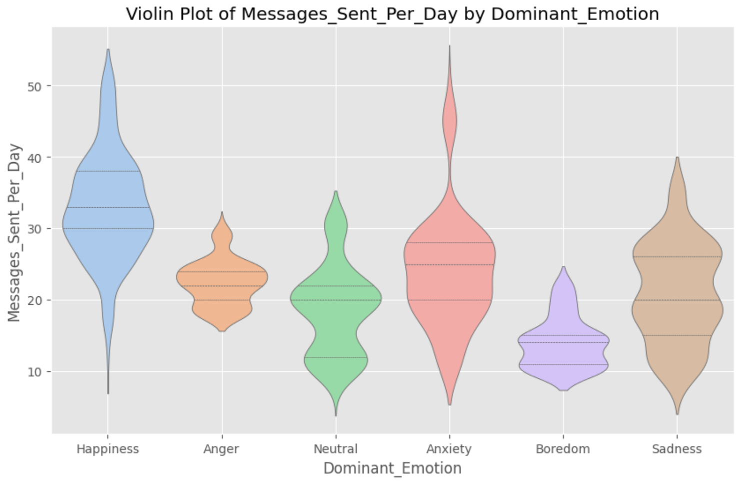




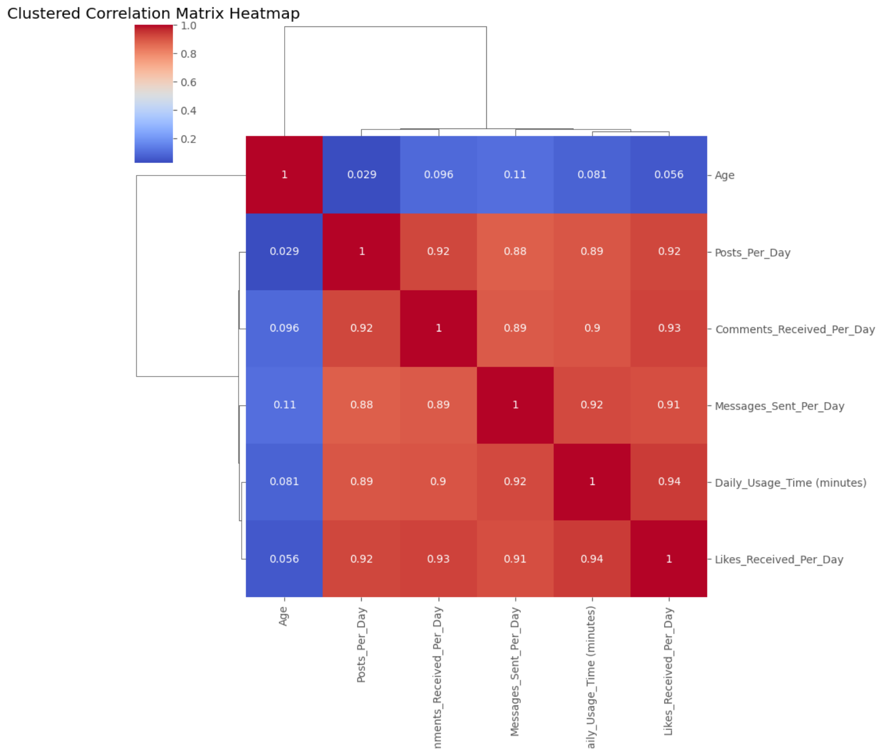


- **Messages Sent Per Day**:

  - Users sending over 40 messages daily reported higher happiness and neutrality, indicating active communication contributes to well-being.
  - Reduced messaging activity aligned with sadness and boredom.



2. Correlation Analysis



A correlation matrix identified relationships among numeric features:

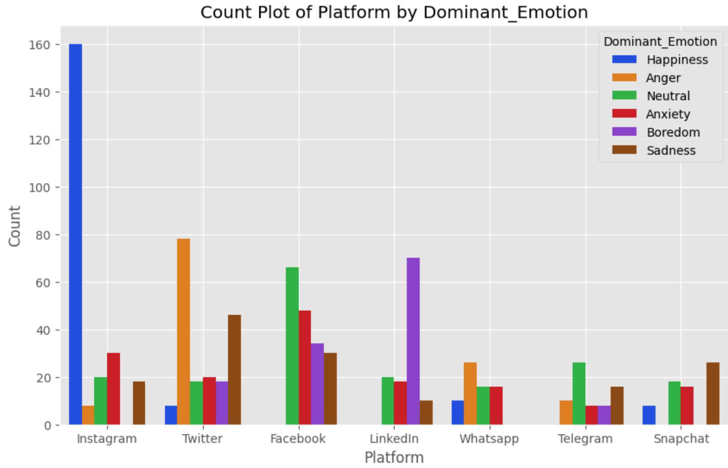
- **Positive Emotions** (Happiness, Neutrality):
  - Strong positive correlations with likes, comments, and daily usage time.
  - High engagement metrics indicated greater emotional well-being.
- **Negative Emotions** (Sadness, Anxiety):
  - Moderate correlations with prolonged usage and messaging activity, reflecting targeted or coping behaviours.

3. Platform-Specific Insights

- **Instagram**:

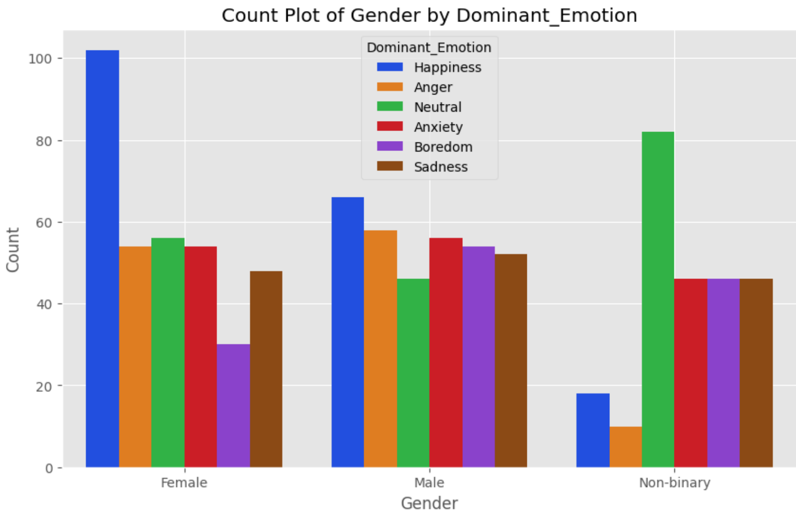
Users reported higher happiness levels, linked to its interactive and visually engaging nature.

- **Twitter and Snapchat**:
  Associated with increased anxiety and sadness, possibly due to fast-paced and ephemeral interactions.
- **LinkedIn**:
  Showed higher boredom levels, reflecting professional and less emotionally stimulating usage.



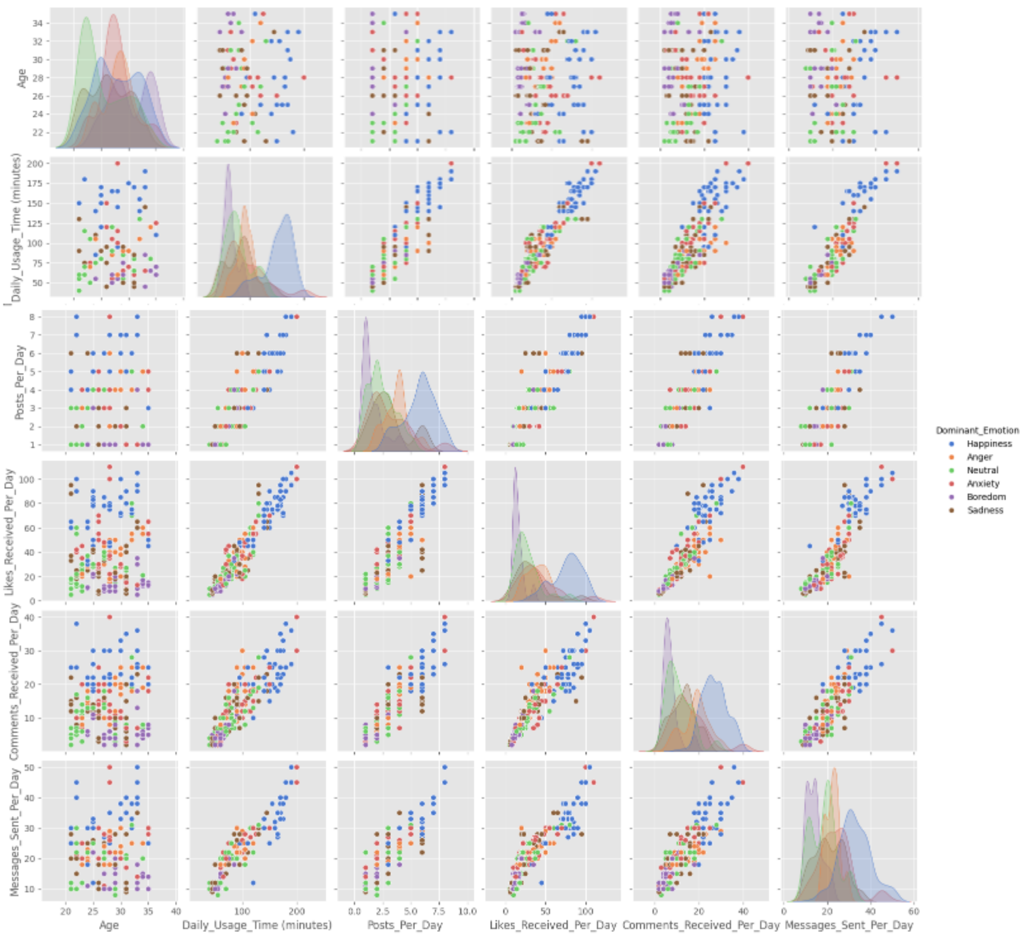
4. Gender-Based Emotional Patterns

- **Females**:
  - Reported more positive emotions (e.g., happiness, neutrality) and exhibited higher engagement levels in likes and comments.
- **Males**:
  - Displayed broader emotional variability, including notable levels of sadness and happiness.



- **Non-Binary Users**:
  - Predominantly reported neutral emotions, with some occurrences of sadness and anxiety.

5. Emotional States and Behavioural Metrics



- **Happiness**:
  - Associated with the highest levels of likes, comments, and usage time.
  - Active messaging (>40 messages/day) was a consistent indicator of happiness.
- **Sadness**:
  - Linked to reduced likes, comments, and variability in usage time.
  - Lower engagement metrics reflected disengagement or limited interaction.
- **Anxiety**:
  - Prolonged daily usage correlated with increased anxiety, though engagement levels were steady across metrics.

6. Key Findings

- **Optimal Engagement**: Moderate social media usage (80–120 minutes/day) promotes positive emotional states, while excessive usage (>150 minutes) is associated with negative outcomes.
- **Engagement Metrics**: Likes, comments, and messaging activity are strong predictors of happiness, emphasizing the importance of active participation.
- **Platform Dynamics**: Instagram fosters positive emotions, while Twitter and Snapchat exhibit higher associations with anxiety and sadness.

7. Implications for Modelling

The insights from EDA provided valuable guidance for feature selection and model development:

- Engagement metrics such as daily usage time, likes, and comments were identified as primary predictors of emotional states.
- Platform-specific trends and gender-based variations informed strategies for targeted interventions.

Overview of the Models

The analysis employed two advanced machine learning models: **Random Forest Classifier** and **XGBoost Classifier**, both well-suited for predicting emotional states based on social media engagement metrics. These models were chosen for their complementary strengths: Random Forest offers high interpretability and robustness against overfitting, while XGBoost provides computational efficiency and exceptional performance with complex data.

Random Forest Classifier

The Random Forest Classifier is an ensemble learning technique that builds multiple decision trees and combines their outputs to enhance predictive accuracy. It is particularly effective for datasets with diverse features, offering robustness against overfitting.

Implementation and Performance

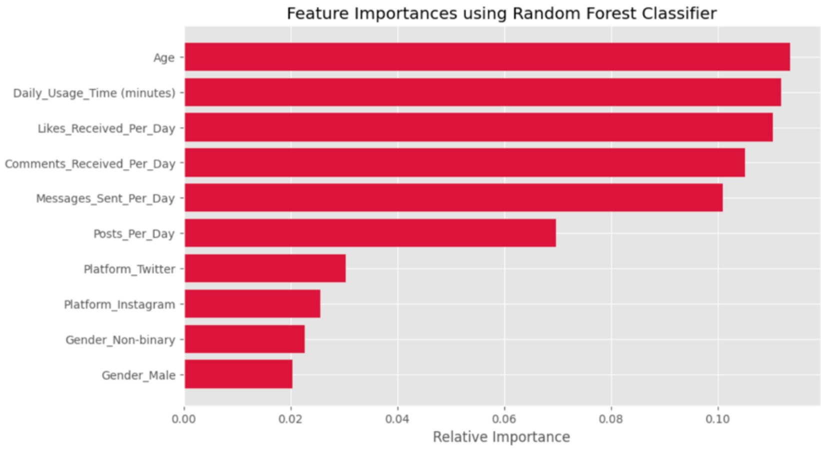
- **Model Setup**:
- The Random Forest was initialized with 100 estimators (trees) and trained on the dataset.
- Features were scaled to ensure equal weighting.
- **Performance**:
- **Without SMOTE**:
    - Accuracy: **97%**
    - Precision and recall were high for majority classes like Happiness and Neutrality.
    - Minority classes such as Aggression and Boredom showed lower recall due to class imbalance.

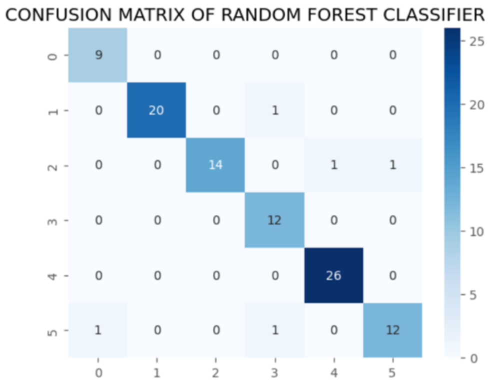
- **With SMOTE**:
    - Accuracy: **97%**
    - Recall improved significantly for minority classes, balancing the overall performance.

Feature Importance:

Random Forest identified the following as the top predictors of emotional states:

1. **Daily Usage Time (11.4%)**: The strongest predictor, emphasizing the role of time spent on social media in influencing emotions.
2. **Likes Received Per Day (11.0%)**: Indicative of positive reinforcement.
3. **Messages Sent Per Day (10.0%)**: Highlighted the importance of active communication.





Observations:

- The model was highly interpretable, making it a valuable tool for understanding the behavioural drivers of emotional states.
- Its ensemble approach ensured stable and reliable predictions across diverse emotional categories.

XGBoost Classifier

XGBoost, a gradient boosting algorithm, optimizes decision trees sequentially to reduce errors and improve predictions. Known for its efficiency and scalability, it is well-suited for handling complex datasets.

Implementation and Performance

Model Setup:

- XGBoost was initialized with 100 estimators, a learning rate of 0.1, and a maximum depth of 6.
- The target variable was encoded, and continuous features were scaled for consistency.

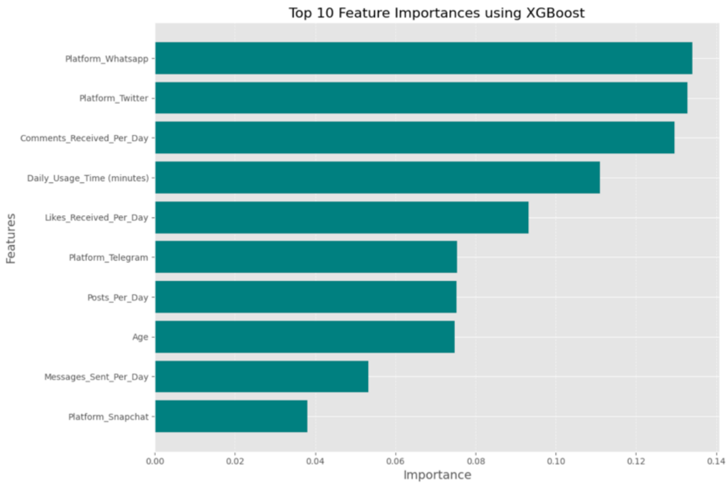
Performance:

- **Without SMOTE**:
    - Accuracy: **96%**
    - Slightly lower recall for minority classes compared to Random Forest.
    - Precision and recall were balanced for majority emotions like Happiness and Neutrality.
- **With SMOTE**:
    - Accuracy: **97%**
    - Recall improved significantly for underrepresented classes, achieving parity with Random Forest.

Feature Importance:

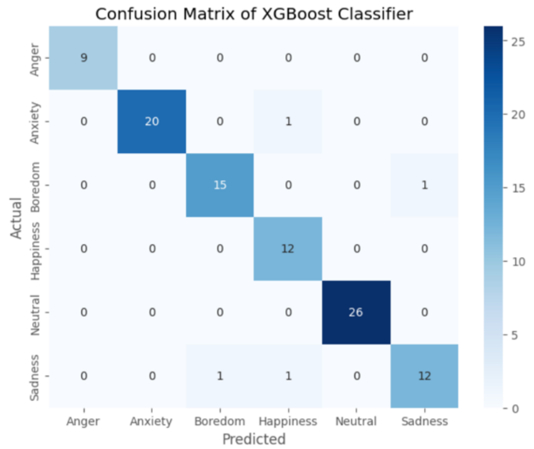
XGBoost produced similar rankings to Random Forest:

- **Daily Usage Time**: Consistently ranked as the most influential feature.
- **Likes Received Per Day**: Critical for distinguishing positive emotions.
- **Comments Received Per Day**: Highlighted interactive engagement as a driver of emotional well-being.



Observations:

- XGBoost excelled in handling noisy data and offered exceptional performance on imbalanced datasets when paired with SMOTE.
- However, its complexity and reduced interpretability made it less suitable for generating actionable insights compared to Random Forest.



Using SMOTE

SMOTE (Synthetic Minority Oversampling Technique) was applied to balance the dataset by generating synthetic samples for minority emotional classes.

Impact on Model Performance:

- **Random Forest**:
  - Recall for minority classes like Aggression and Boredom improved without affecting the overall accuracy of 97%.
- **XGBoost**:
  - Similar improvements in recall for minority classes, aligning its performance with that of Random Forest.

Observations:

- SMOTE significantly enhanced the models' ability to predict minority classes while maintaining high accuracy for majority classes.
- Both models demonstrated better-balanced performance after SMOTE, addressing the initial class imbalance effectively.

Model	**Without SMOTE (Accuracy)**	**With SMOTE (Accuracy)**	**Key Strengths**	**Limitations**
Random Forest	97%	97%	High interpretability; robust feature importance analysis	Sensitive to noisy data in some cases
XGBoost	96%	97%	Efficient handling of complex data; resistant to overfitting	Reduced interpretability

Modelling and Observations:

Both Random Forest and XGBoost achieved high accuracy and balanced performance after applying SMOTE. However, they offer distinct advantages:

Random Forest:

- Best suited for applications where interpretability is critical, as it provides clear insights into feature importance.
- Ideal for understanding the behavioural drivers of emotional states.

XGBoost:

- Better for handling noisy and complex datasets, with superior computational efficiency.
- Suitable for scenarios requiring highly optimized and scalable solutions.

Observations:

- **Feature Importance**: Both models consistently identified **Daily Usage Time**, **Likes Received**, and **Messages Sent** as critical predictors of emotional states.
- **SMOTE Effectiveness**: Balanced class distributions improved recall for minority emotions without compromising overall accuracy.
- **Practical Implications**: These findings can guide social media platforms in designing user-centric features that promote positive interactions and mitigate negative emotional outcomes.

CONCLUSION

This study investigated the relationship between social media usage patterns and emotional well-being by developing predictive models, analysing key behavioural metrics, and addressing class imbalance in the dataset. The findings provide actionable insights into the drivers of emotional states and their implications for digital platform design.

1. **Predictive Modelling of Emotional States**:
    a. Using **Random Forest** and **XGBoost**, we successfully classified users' dominant emotions based on their social media activity and demographic data. Both models demonstrated high accuracy, with **97% after applying SMOTE**. Random Forest offered better interpretability, making it ideal for feature analysis, while XGBoost excelled in handling noisy and complex data relationships.
2. **Key Metrics Influencing Emotional States**:
    a. Feature importance analysis identified **Daily Usage Time**, **Likes Received Per Day**, and **Messages Sent Per Day** as the most influential predictors of emotional outcomes. These metrics highlight the central role of engagement in shaping emotions, with positive emotions like Happiness strongly correlated with high likes and active communication.
3. **Impact of Social Media Usage Patterns**:

a. The study revealed that **moderate usage (80–120 minutes/day)** fosters positive emotional states such as Happiness and Neutrality, while **excessive usage (>150 minutes/day)** is linked to negative states like Anxiety and Sadness. Platform-specific trends showed that **Instagram** is associated with Happiness due to its interactive features, whereas **Twitter and Snapchat** exhibit higher associations with Anxiety and Sadness, likely due to their fast-paced, ephemeral nature.

4. **Addressing Class Imbalance with SMOTE**:
    a. SMOTE significantly improved the models' ability to predict minority emotions such as Aggression and Boredom. This enhancement balanced recall across all classes without compromising overall accuracy, ensuring fair and reliable predictions for underrepresented emotional states.

Key Takeaways:

- Predictive modelling can reliably classify emotional states, providing actionable insights for understanding user behaviour.
- Behavioural metrics such as Daily Usage Time and Likes Received are critical for fostering positive emotional outcomes.
- Addressing class imbalance through SMOTE enhances model performance, particularly for minority emotional states.
- Social media platforms can leverage these insights to design features that promote positive interactions and mitigate risks of negative emotional impacts.

This study underscores the dual impact of social media usage, providing a data-driven foundation for developing user-centric, emotionally intelligent platforms. Future research should explore additional contextual factors, longitudinal patterns, and interaction quality to deepen the understanding of social media's psychological effects.

Bibliography

1. Escobar-Viera, C. G., Shensa, A., Sidani, J. E., Bowman, N. D., Marshal, M. P., & Primack, B. A. (2021). **The Dual Role of Social Media in Mental Health: Opportunities and Risks**. *Journal of Adolescent Health, 58*(6), 146–150.
2. Berryman, C., Ferguson, C. J., & Negy, C. (2023). **Social Media Use and Mental Health: A Review of Psychological, Social, and Technological Perspectives**. *Psychology of Popular Media Culture, 12*(2), 89–103.
3. Chen, T., & Guestrin, C. (2016). **XGBoost: A Scalable Tree Boosting System**. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794.
4. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., … & Duchesnay, E. (2011). **Scikit-learn: Machine Learning in Python**. *Journal of Machine Learning Research, 12*, 2825–2830.
5. Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). **SMOTE: Synthetic Minority Over-sampling Technique**. *Journal of Artificial Intelligence Research, 16*, 321–357.
6. Kaggle. (n.d.). **Social Media Usage and Emotional Well-being Dataset**. https://www.kaggle.com/datasets/emirhanai/social-media-usage-and-emotional-well-being
7. Zhao, J., Wang, T., & Li, B. (2022). **The Influence of Passive and Active Social Media Engagement on Mental Health**. *Computers in Human Behavior, 139*, 107556.
8. Sutton, R. S., & Barto, A. G. (2018). **Reinforcement Learning: An Introduction** (2nd ed.). *MIT Press*.
9. Seaborn Documentation. (n.d.). **Statistical Data Visualization with Seaborn**. https://seaborn.pydata.org/
10. Matplotlib Documentation. (n.d.). **Matplotlib: Visualization with Python**. Retrieved from https://matplotlib.org/