Pre-Analysis Plan: Understanding the Relationship Between Social Media User Interactions and Emotional States

Yuina Barzdukas, Brian Stoss, Max Chang
DS 3001
3/31/25

Definition of an Observation

Each observation in this study represents a single user's interaction on a given day. If a user interacts multiple times in a day, these interactions are already aggregated to form a single data point per user per day. In other words, the data points represent a single user's dominant emotion during the day.

Supervised or Unsupervised Learning

This study primarily employs supervised learning to classify users' dominant emotional states based on their interactions with social media platforms. Additionally, some exploratory unsupervised learning techniques, such as clustering, may be used to uncover latent patterns in the data.

Models & Algorithms

We will utilize classification models to predict the dominant emotional state of a user based on their interaction data. The primary models include:

- Logistic Regression: A baseline model for classification.
- Decision Trees: To capture nonlinear relationships between features.
- Random Forest: An ensemble method to improve classification accuracy.

For unsupervised learning, if necessary, we will employ:

- K-Means Clustering: To explore potential groupings of users based on interaction features.

Success Metrics

To evaluate model performance, we will use:

- Accuracy: The proportion of correctly classified emotional states.
- Precision & Recall: To assess the reliability of predictions, especially for less frequent emotions [1].
- F1-Score: A balance between precision and recall [1].
- ROC-AUC Score: To evaluate the model's discriminatory power [1].

For exploratory analysis, in which we have already somewhat achieved in our wrangling and EDA, we will assess:

- Correlation Coefficients: To quantify relationships between key interaction features and emotional states.
- Chi-Square Tests: To determine statistical significance in categorical variables.

Anticipated Weaknesses & Mitigation Strategies

- Data Quality Issues: Missing values will be handled through imputation techniques such as mean/mode imputation for numerical variables and k-nearest neighbors (KNN) for categorical variables.
- Selection Bias: We will assess representativeness and, if necessary, apply weighting adjustments.
- Subjectivity in Emotional State Reporting: We will compare with other sources (e.g., sentiment analysis from text interactions) to ensure robustness due to self-reported emotions.
- Feature Engineering Challenges: We will perform feature selection techniques such as Principal Component Analysis (PCA) to reduce dimensionality and improve model interpretability [2].

By implementing these methodologies, we aim to accurately model and interpret the relationship between user interactions and emotional states while mitigating potential biases and data limitations.

References

[1] Olumide, Shittu. “A Complete Guide to Model Evaluation Metrics.” *Statology*, 23 Jan. 2025, www.statology.org/complete-guide-model-evaluation-metrics/.

[2] Frost, Jim. “Principal Component Analysis Guide & Example.” *Statistics By Jim*, 29 Jan. 2023, statisticsbyjim.com/basics/principal-component-analysis/.