The rapid development of the modern world has increased the use of social media. One of the most popular social media platforms in Indonesia is X, where users can share tweets to express themselves. Activities that occur on X can reveal aspects of a user's personality. This research focuses on using the Logistic Regression algorithm, incorporating various data preprocessing scenarios, data balancing techniques, data splitting ratios, and feature extraction methods, to classify the personalities of X users based on their tweets. The data for this study were collected using a widely distributed Big Five Personality questionnaire and further enriched by crawling tweets from 257 Indonesian X-user respondents. The data was classified into five personality traits, namely openness, conscientiousness, extraversion, agreeableness, and neuroticism. Data preprocessing techniques were applied in two scenarios to make the dataset more suitable for model building. Dataset imbalance was addressed by applying two data balancing techniques, namely Random Removal Augmentation and Synthetic Minority Oversampling Technique (SMOTE). Hyperparameter tuning is performed using grid search to identify the optimal model parameters. Experiments were conducted with several combinations of feature extraction such as TF-IDF, TF-IDF with BOW, and TF-IDf with N-gram. The highest accuracy achieved is 71%.

Index Terms—Big Five Personality, Logistic Regression, Personality Classification