Assignment 2: Content Analysis and Regression Report

**Research Background**

Literary characters are frequently portrayed as networks of personality qualities, which is essential for capturing readers' emotions and drawing them into the story. According to Docherty, character is conveyed through a variety of methods, including appearance, speech, and deeds, and it offers insights into people's attitudes and habits. Real-life characteristics are combined with fan interpretations to create characters in the Real Person Fiction (RPF) genre that reflect current sociological and cultural trends.

Fanfiction abounds on sites like AO3, thanks to the global appeal of the Korean boy band BTS. Academics have observed how male K-pop idols exemplify "manufactured versatile masculinity" by fusing traditional norms of hegemonic masculinity with softer, more feminine appearances or actions. This tendency, known as "liminal masculinity," transcends gender boundaries and captures the androgynous essence of K-pop idols.

**Aim**

This assignment aims to analyze the impact of soft masculinity on the success of BTS fanfiction, using a corpus of 100 fanfiction.

**Data**

The dataset comprises 100 BTS-related fanfictions from the GOLEM project, containing story ID, publication year, word count, kudos, comments, and story content in English (ranging from 1,000 to 1,200 words).

**Experimental Setup**

Data Collection: A dataset containing textual features such as word count, publication year, masculine power score, and lexical richness was collected from an online platform.

Feature Engineering: Two additional features, namely masculine power score and lexical richness, were engineered from the text data using NLP techniques. Masculine power score was calculated based on the frequency of male pronouns, while lexical richness was determined by the diversity of vocabulary used in the text.

Data Preprocessing: Missing values in the dataset were handled by replacing them with zeros to ensure completeness of the data. Descriptive statistics were computed to understand the distribution of data and identify potential outliers.

Regression Analysis: An Ordinary Least Squares (OLS) multiple regression model was fitted to predict the number of kudos received based on the engineered features. The model included masculine power score, publication year, and word count as independent variables.

Assumption Testing: Various assumptions of the regression model, including normality, homoscedasticity, and multicollinearity, were tested to ensure the reliability of the regression analysis.

**Results and Discussion**

Regression Analysis Results: The regression analysis revealed a statistically significant model with an R-squared value of 0.164, indicating that approximately 16.4% of the variability in kudos can be explained by the model. However, the p-values associated with the coefficients of masculine power score and word count were not statistically significant, suggesting that these variables may not have a significant impact on kudos.

Assumption Testing Results:

1. Normality: The Shapiro-Wilk test indicated a significant departure from normality in the residuals (p < 0.001), implying potential issues with the assumption of normality.
2. Homoscedasticity: The scatterplot of residuals versus predicted values exhibited a widening spread of residuals with increasing predicted values, indicating potential heteroscedasticity and violation of the homoscedasticity assumption.
3. Multicollinearity: Variance Inflation Factor (VIF) values suggested low multicollinearity among the independent variables, except for the constant term, which exhibited numerical issues due to its nature.

**Conclusion**

The experimental analysis provided valuable insights into the factors influencing audience engagement with textual content. However, several limitations and areas for improvement were identified, including the need for addressing issues related to model assumptions, such as normality and homoscedasticity, and exploring alternative modeling techniques to improve predictive performance. Further research is warranted to refine the model and enhance its reliability in predicting audience engagement metrics.