The Influence of Main and Minor Character Lines from The Office on an Episode's Average Sentiment Score

AUTHOR
Ashley Raj

Abstract

Problem

This study investigates whether we can predict the average sentiment score of episodes from the popular sitcom called "The Office" based on the number of main and minor character lines spoken in the episode.

Approach

After performing data preparation and cleaning, a box plot of the sentiment score by the top 10 characters in the show, including both minor and major character, was created. The box plot represented the difference in sentiment scores between the minor and major characters over the seasons, where the major characters had the most variability in their lines due to the outliers and their high IQR ranges. A graph of their average sentiment scores of each season was created, where the average sentiment scores were higher for seasons with a larger number of main characters present. An anova test was conducted to determine the effect of the number of main and minor character lines on the average sentiment score. A linear regression model was used to test out the hypothesis, and several versions were made. After conducting anova tests on these models, the variables that were relevant were: "main_character_lines", "minor_character_lines", "season_factor", and "interaction_factor". The accuracy was determined by calculating the R^2 value and the Mean Squared Error.

Results

After running several anova tests, the most successful linear regression model was "lr_model_w_out_imdb_ratings_squares" with an R^2 value of 0.2342289 and a Mean Squared Error of 0.001250424. Before determining which variables to keep, running an anova test on the linear regression model called "lr_model_interaction" revealed that the IMDB ratings and the minor and main character lines squared did

not help the model due to their p-value being greater than 0.05. Removing those variables, resulted in the "lr_model_w_out_imdb_ratings_squares" able to explain 23.42% of the variance present in the data.

Conclusion

IMDB ratings per episode, the squared number of main character lines, and the squared number of minor character lines are not good indicators for predicting the average sentiment score for each episode. The number of minor and main character lines, the interaction variable (multiplying the major and minor character lines per episode), and the season the episode was from were suitable variables to predict the average sentiment score for each episode. Based on the data from the final linear regression model, the null hypothesis can be rejected, showing that the number of main and minor character lines do impact the average sentiment score of an episode from The Office.

Background

The Office is a popular sitcom due to its wide variety of characters and its different situations. The dataset used included the following variables: "season", "episode", "episode_name", "director", "writer", "character", "text", "text_w_direction", "imdb_rating", "total_votes", "air_date", "sentiment_analysis_score", "sentimentr_score", and "syuzhet_score". The sentiment scores measure the emotional tone and feeling expressed in a line of dialogue. Through data analysis, a trend observed was that seasons with a low number of main characters had a lower average sentiment score and seasons with all main characters present would have higher average sentiment scores. Additionally, through a box plot, major characters had the most variability of sentiment scores in their lines due to the outliers and their high IQR ranges.

This study investigates the effect that main and minor characters have on the emotional tone and attitude expressed in the episodes. The research question that was being investigated is "Can we predict the overall sentiment score of episodes based on how many main versus minor characters are in it?". The null hypothesis was that the number of minor and main character lines would not have an effect on the average sentiment score while the alternative hypothesis would be that the number of minor and main characters would have a significant effect on the average sentiment score per episode.

Results

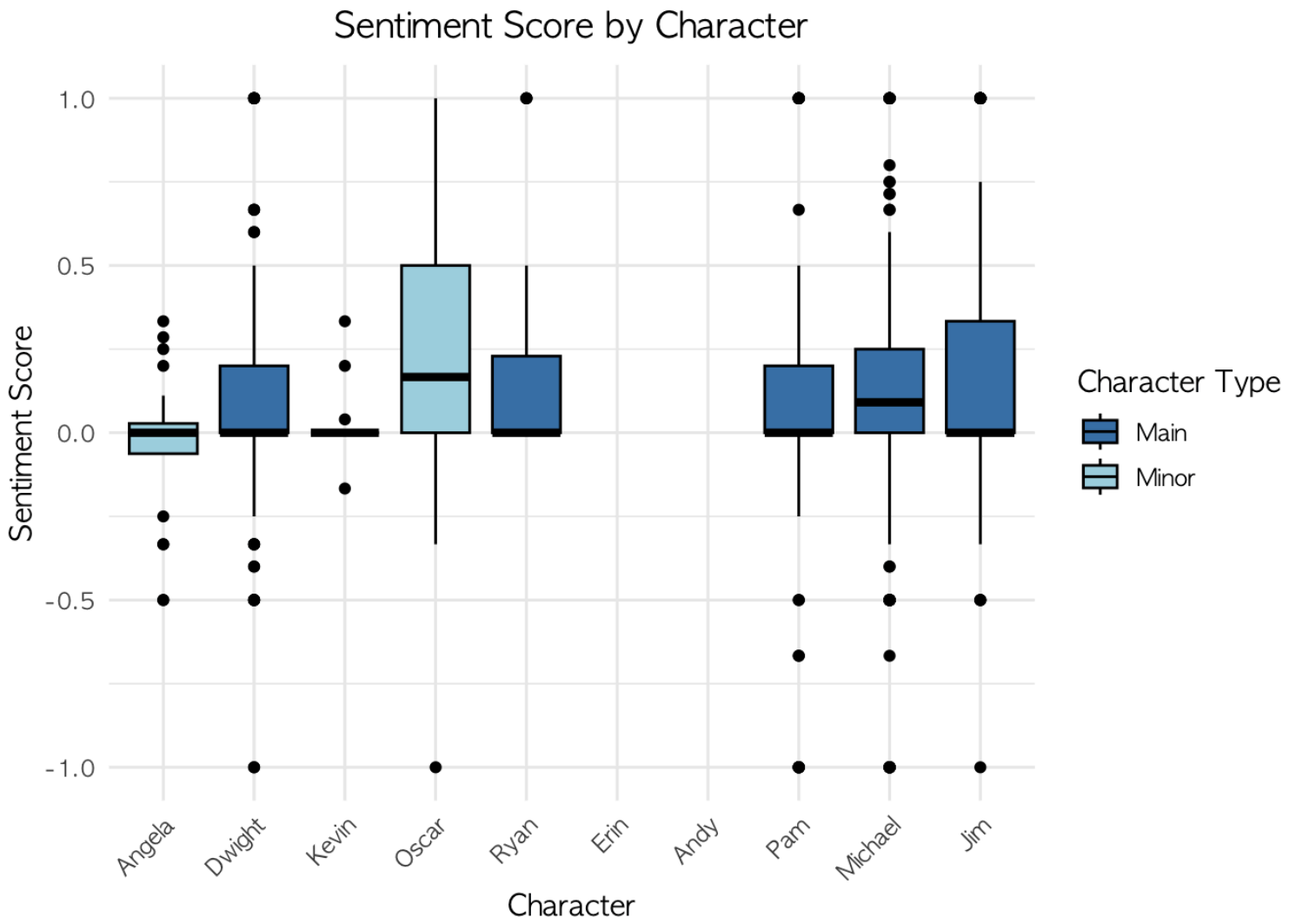


Figure 1: Sentiment Score by Character

This plot shows the average sentiment score per main vs minor character. The characters in dark blue are main characters and the characters in light blue are the minor characters. The character lapse by their appearances per season, which is why certain characters like Erin and Andy appear later on the graph and why Micheal does not appear later on. This box-plot shows that main characters tend to have more extreme outliers and larger IQR ranges like Micheal and Jim.

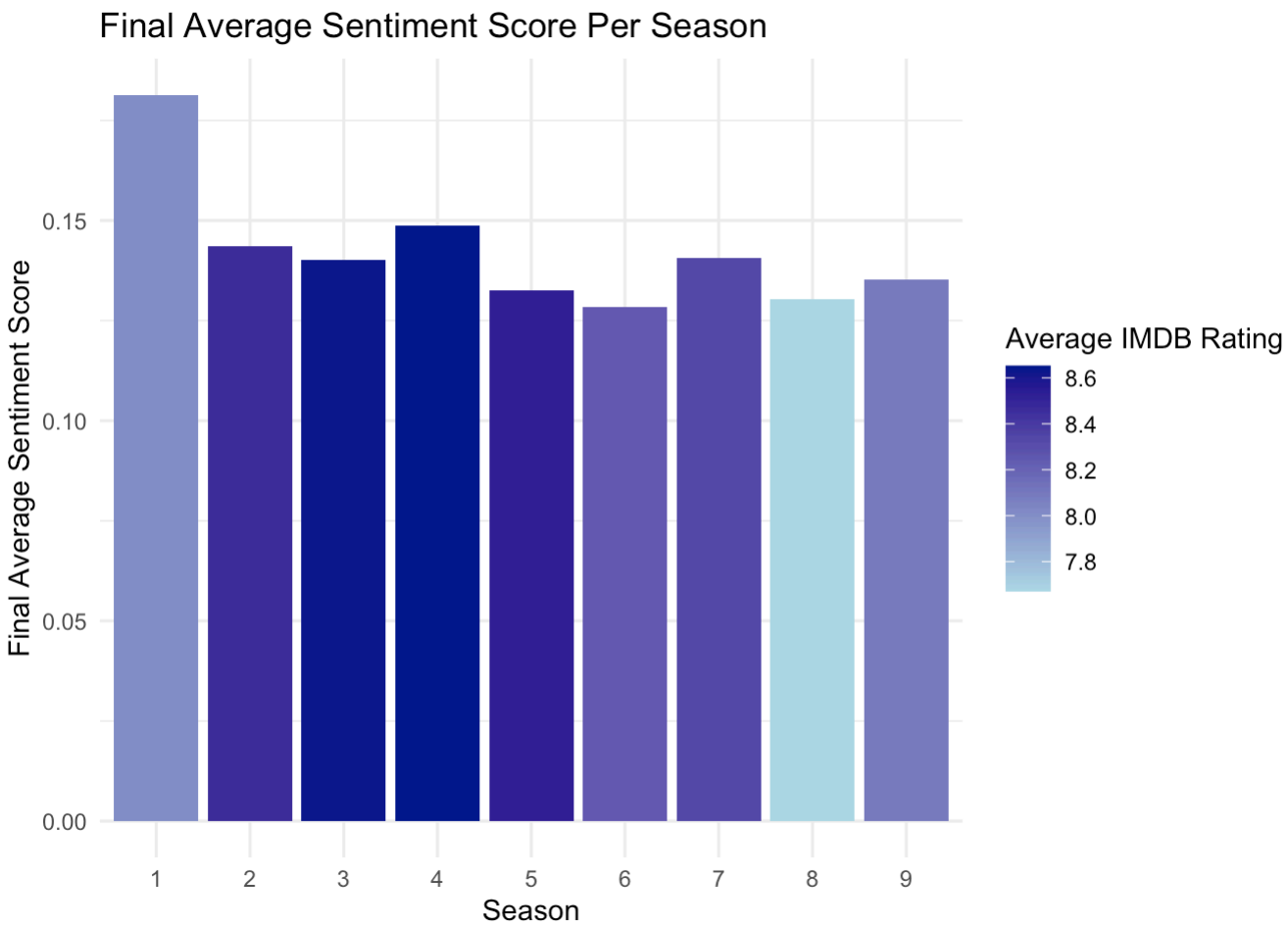
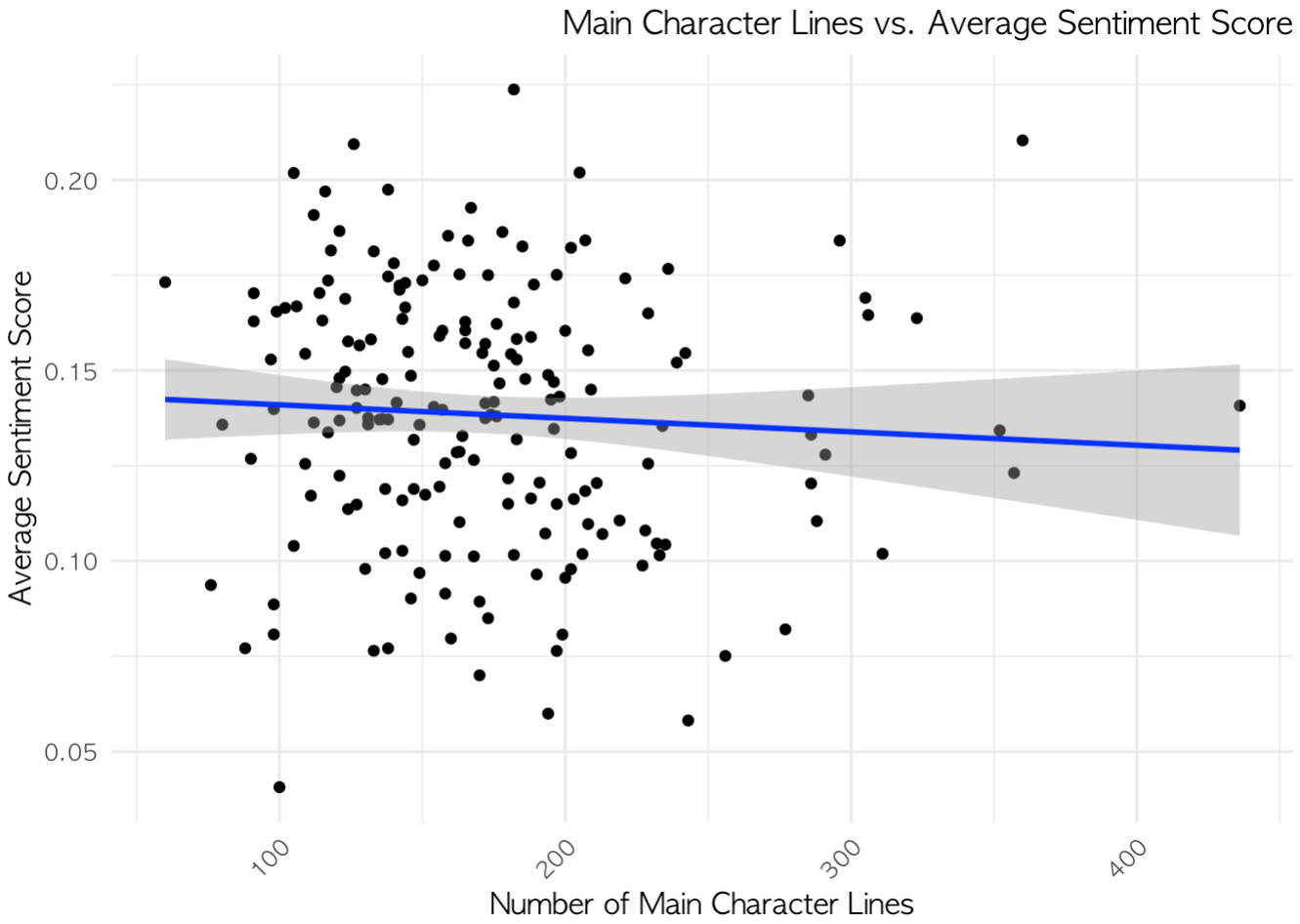


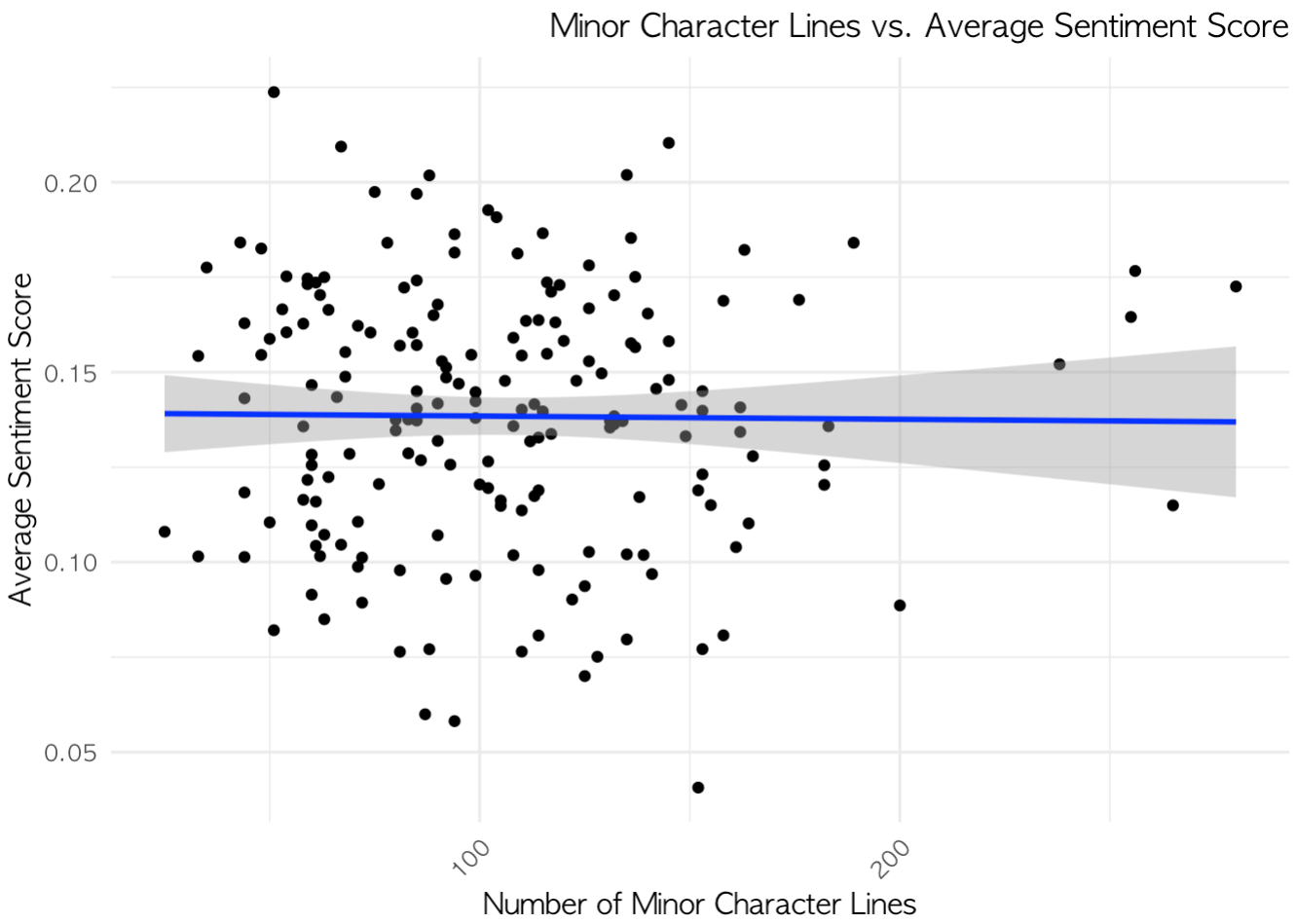
Figure 2: Final Average Sentiment Score Per Season

This plot shows the average sentiment score (using sentiment_analysis_score, sentimentr_score, and syuzhet_score values) per season, along with the average IMDB ratings. A darker purple color represents a high IMDB rating around 8.6 while a lighter blue color represents a lower IMDB rating around 7.8. Season 4 has a large number of main characters, which is why it has a higher average sentiment score and IMDB rating. Seasons 8 and 9 have a lower amount of main characters, which is why they have lower average sentiment scores and IMDB ratings.

```
`geom_smooth()` using formula = 'y ~ x'
```




```
`geom_smooth()` using formula = 'y ~ x'
```



Figures 3 and 4: Main Character Lines vs. Average Sentiment Scores and Minor Character Lines vs. Average Sentiment Scores. These graphs show the distribution of main vs minor character lines and the corresponding sentiment scores they received. The line of best fit can be seen as a straight horizontal line, showing the initial correlation to be weak.

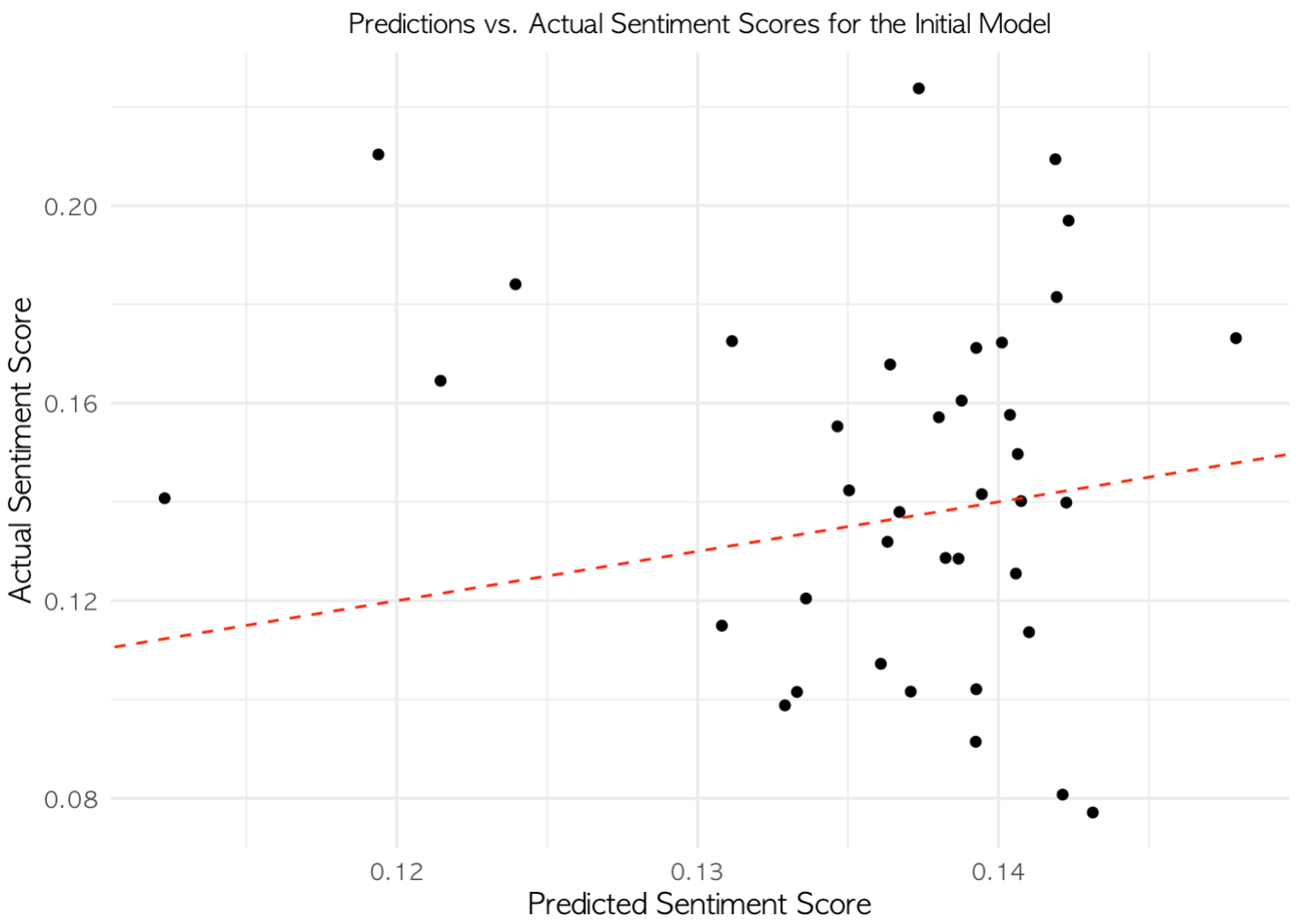


Figure 5: Predictions vs. Actual Sentiment Scores for the Initial Model. The factors I used to predict the average sentiment scores in this model was the number of main and minor character lines. The r-squared value was around 0.0218, showing that only 2.18% of the variance can be explained by the model. The initial model shows there is a low correlation, as the data points do not show an upward trend.

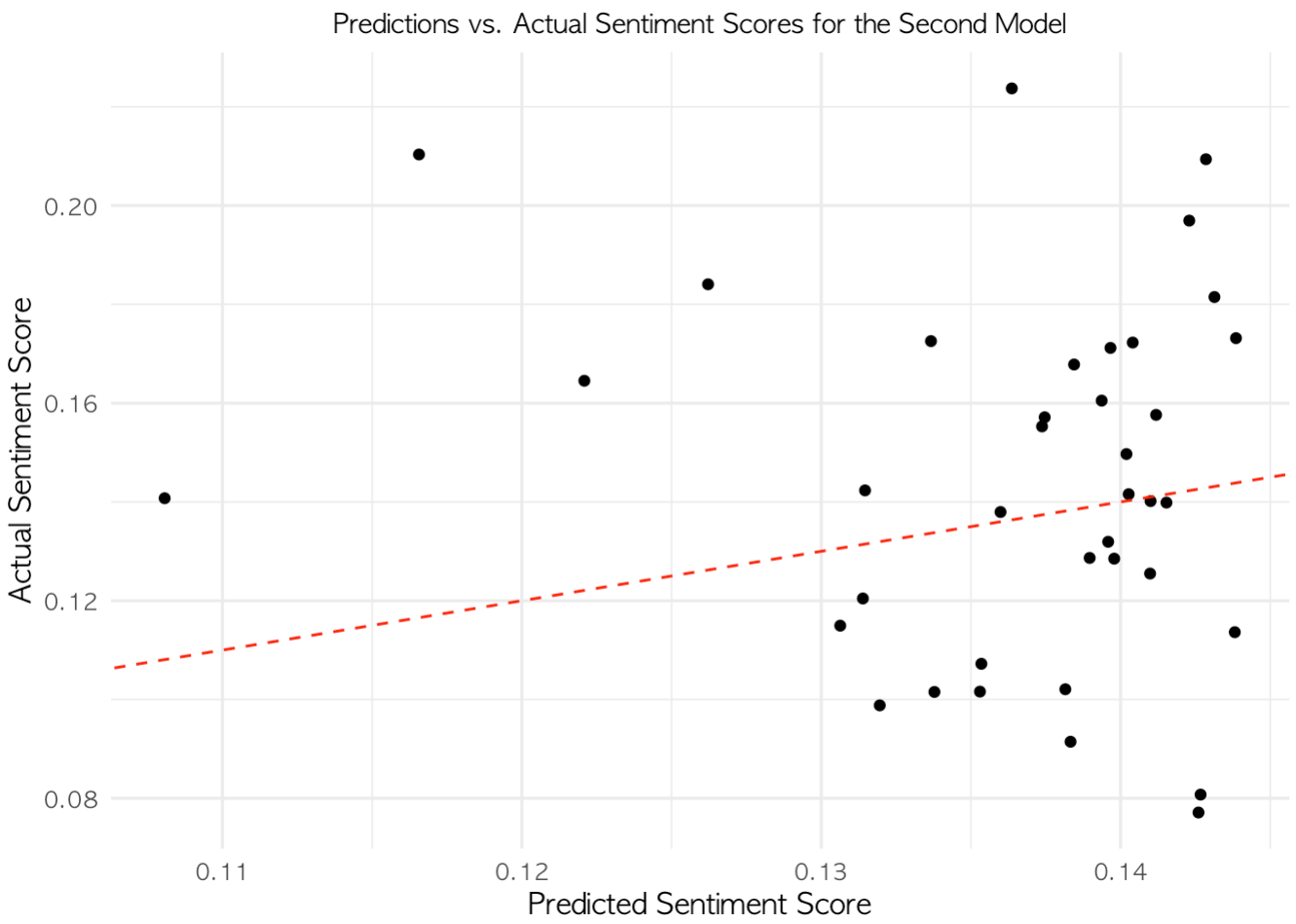


Figure 6: Predictions vs Actual Sentiment Score for the Second Model. The factors used in this model were the number of main character lines, the number of minor character lines, and the IMDB ratings per episode. Based on the results from the anova test, the results from this model had improved but not significantly.

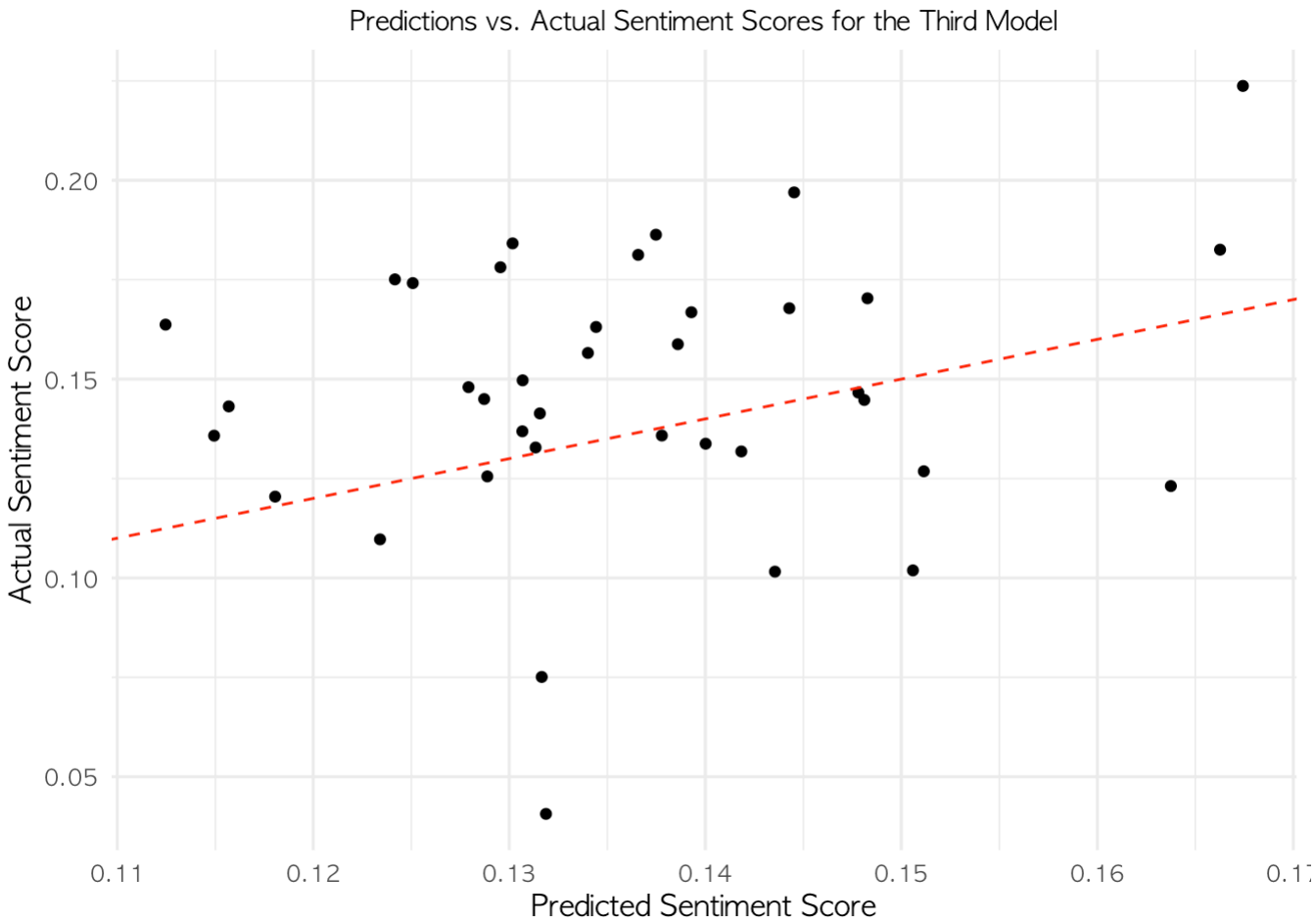


Figure 7: Predictions vs Actual Sentiment Score for the Third Model. The factors from this model include the number of main character lines, the number of minor character lines, the number of main character lines squared, the number of minor character lines squared, the IMDB ratings of each episode, an interaction factor (the number of minor and main character lines multiplied with each other), and the season the episode was in. Based on the anova test conducted on this model, it performed significantly well from the previous two models. However, when performing a summary on this model, the factors that were negatively affecting the model were the number of minor and major character lines squared and the IMDB rating.

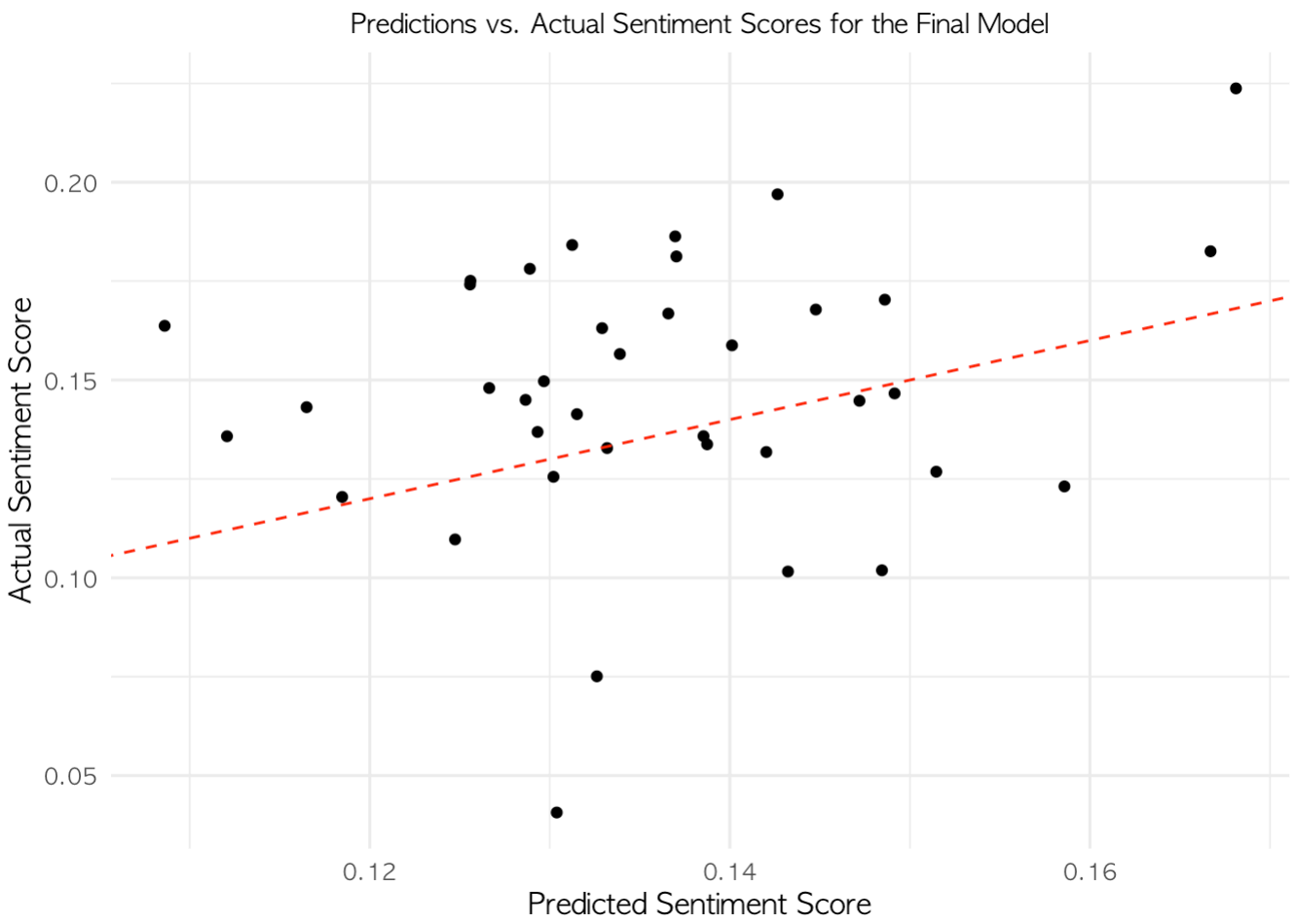


Figure 8: Predictions vs Actual Sentiment Score for the Final Model. The final model contained the remaining factors that did not negatively effect the third model which was the number of main character lines, the number of minor character lines, the interaction factor, and the season the episode was in. This model significantly improved from the initial model, as its R-value had shown that 23.42% of the model's variance can be explained by the model.

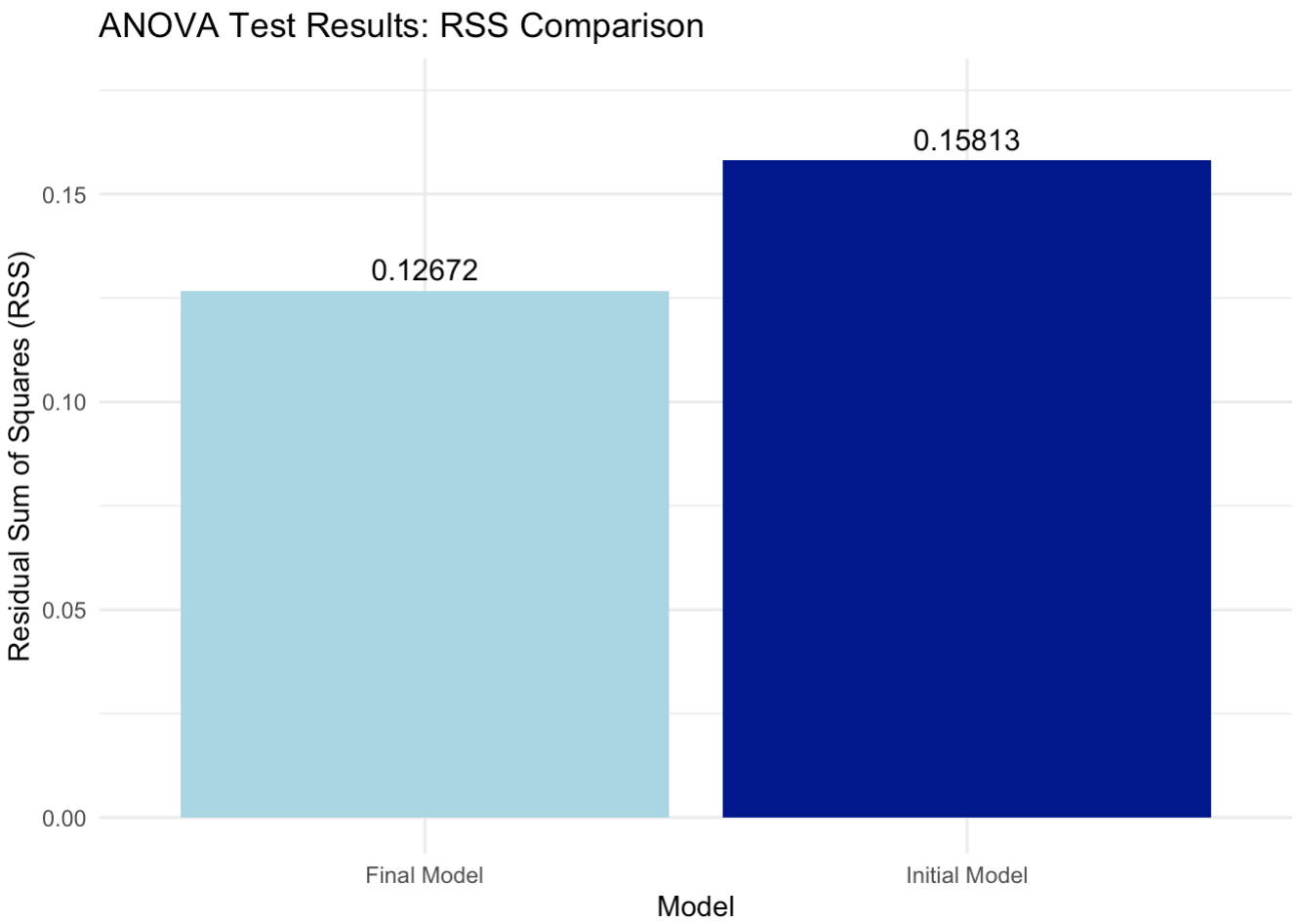


Figure 9: RSS Value Comparisons between the Initial Model vs the Final Model. This plot shows that there were significant improvements from the initial and final model. The RSS value of the final model was significantly smaller than the initial model showing that there was a decrease in the likelihood of errors in the final model.

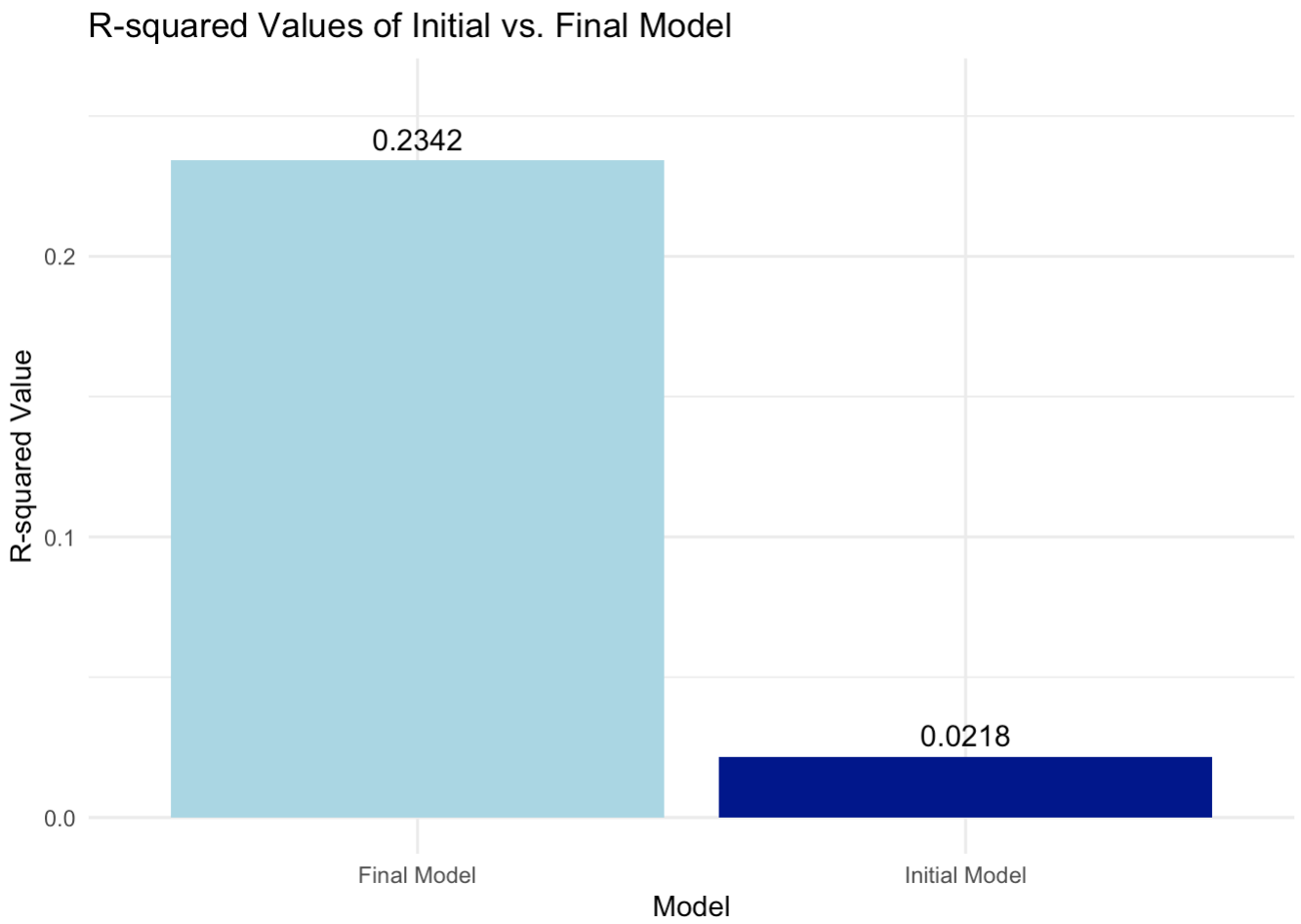


Figure 10: R-squared Value Comparisons between the Initial Model vs the Final Model. This plot shows a significant improvement between the initial and final model. There was a significant increase in the R-squared value between the initial and final model. The R-squared value shows that 23.42% of the variance can be explained by the final model.

Discussion

Based on my results, I can reject the null hypothesis, thereby accepting the alternative hypothesis which is that the number of main and minor character lines do have an effect on the average sentiment score of an episode from The Office. While my R-squared value is not as preferably high, this is due to the fact that the dataset

contains real-world data containing details about The Office. Through the creation of data plots and linear regression models, it can be concluded that the number of main and minor character lines have a significant effect on average sentiment score per episode.

This conclusion can show that understanding the roles different characters play and their dynamics can change content creation and viewer engagement. The Office was able to successfully use this idea to increase the numbers of viewers they had overtime by building upon main characters throughout the episodes. In the future, it can be investigated whether the increase in main character lines also leads to an increase in IMDB ratings, through similar data analysis and linear regression models. Ultimately, sentiment analysis can be applied to other fields including marketing, customer service, and healthcare, not only the entertainment industry to improve the standard of different products and forms of media.

Code and Data Availability

Here is a link to my GitHub repository:

https://github.com/the-codingschool/DSRP-2024-Derek/tree/main/AshleyProject

The data used was called 'office_sentiment.csv', which comes from a tabular format.

Acknowledgements

I would like to thank The Coding School for giving me this wonderful opportunity to learn more about data science and delve deeper into this field. I would also like to thank the instructor Sarah Parker for teaching and explaining various concepts throughout the course. I would also like to thank my mentor Derek for introducing me to the dataset and offering me suggestions on how I can improve my outputs and results. I would also like to thank my teaching assistant Renate for answering any questions I had and offering my useful suggestions.