Harry Potter Series Text Analysis using R

Stats 133 - Introduction to Text Mining Using R

Donggyu Kim

1. Introduction

The Harry Potter series stands out as a representative of modern fantasy series, attracting readers and viewers with stories of magic, friendship, and challenges. The Harry Potter book series, written by J.K. Rowling, has sold more than 500 million copies worldwide, making it the best-selling book series in history. The film has earned more than $7.7 billion worldwide, ranking as the third highest film series of all time. This success shows that the Harry Potter series has a huge popularity and cultural influence in both literature and film industry. The series, which includes seven books and eight films, not only won the hearts of fans around the world, but also offers an opportunity to study through statistical text analysis. This report begins an exploration of the language of the Harry Potter universe using NLP to reveal details hidden within words, whether written or spoken in the movie. Our study is divided into three main parts, starting with how the mood of the movie changes, reflecting the series' transformation into deeper, darker themes as the character grows and faces greater threats. We aim to investigate whether the movies' dialogues contributed to changing the mood of the movie and analyze how the sentiments of each movie change as the series progresses.

The second part of our study analyzes the differences between the original Harry Potter books and their movie versions. Many fans and critics say that the books provide a deeper and more engaging experience compared to the movies. This section is to measure these differences and discuss whether the core of the story changes when it's turned into a film.

Lastly, we examine the characters based on their Hogwarts houses, Gryffindor and Slytherin, looking at their dialogue to find any patterns or differences in dialogues that reflect their house characteristics. This analysis gives us an interesting look into how the language used by characters shows the values and personalities of the different houses.

2. Data Used

There are four types of data used for the Harry Potter analysis. The first dataset is a script from the Harry Potter movie series sourced from Kaggle in csv format. There are a total of three columns, each is the character who spoke, the chapter, and the actual dialogue. Each movie contains an average of approximately 900 rows, with movie 8, "Deathly Hallows Part 2," having the fewest 712 rows, and movie 5, "Order of the Phoenix," providing the most data with 1157 rows. The following data set are text files of the full Harry Potter series, which were obtained from GitHub. Book 1, "Harry Potter and the Philosopher's Stone," had the smallest size and book 5, "The Order of the Phoenix," had the largest. The remaining two datasets, used for Part 3 of the analysis were the character information csv file with the houses they were from and the spell data.

3. Movie Script Analysis

Before I started analyzing the movie script, there were some hypotheses that I had. First, as mentioned in the introduction, since the movies get darker as the series progresses, I assumed that the text analysis will show us the increasing trend of the negative sentiments as the movies progresses. Also, as the character's age increases as the series advances, analyzing the readability index of the script will also show a pattern that will continue to increase.

In order to test these assumptions, I looked for the readability index and average word per sentence for all eight movie series. However, as shown in Figure 1-1, no correlation was found between the series progression and the readability index. In fact, except for the movie 5, "Order of the Phoenix," it shows the decreasing trend of the readability index.

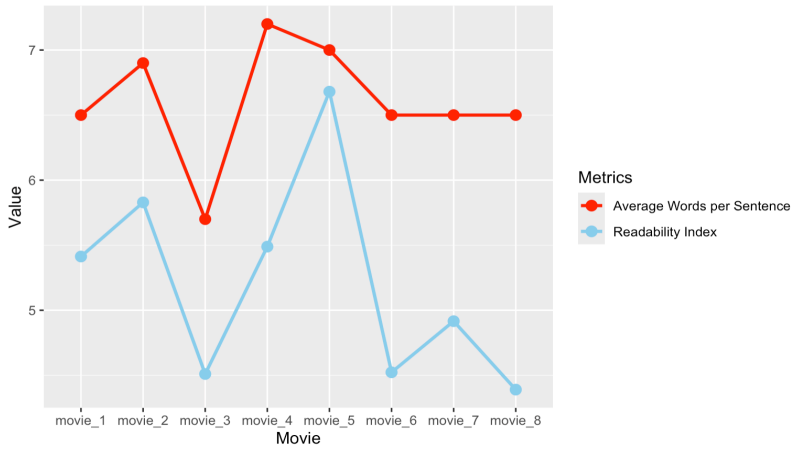


Figure 1-1. Readability index & Average words per sentence of Movie Series

I performed tokenization on the third column of each movie's CSV file, which is dialogue, and performed some pre-processing of the text data to prepare for the text analysis. Since this data came from movie scripts, many unnecessary symbols such as "..." were included, and these symbols were removed separately. I looked for the document term matrix with the corpus obtained and analyzed the words that appeared the most in each movie. As a result, it was confirmed that the word "harry" was used the most in all except for movie 6, "The Half-Blood Prince." Except for the word “harry,” the majority of the words in the most frequent words were common words such as 'now', 'come', 'know', and 'think'.

Then, I conducted an analysis on the bigrams and compiled the 10 most used bigrams in each movie. The result is depicted as shown in Figure 1-2. Excluding the bigram "harry potter," which appeared across all movies, the rankings were mostly occupied by different bigrams for each film. For

example, in the first movie, "Philosopher's Stone" was the second most frequently used bigram after "harry potter," and the "Philosopher's Stone" was rarely used in other movies. From this, I found that, unlike the single word, bigrams can serve as unique characteristics and distinct features of each movie.

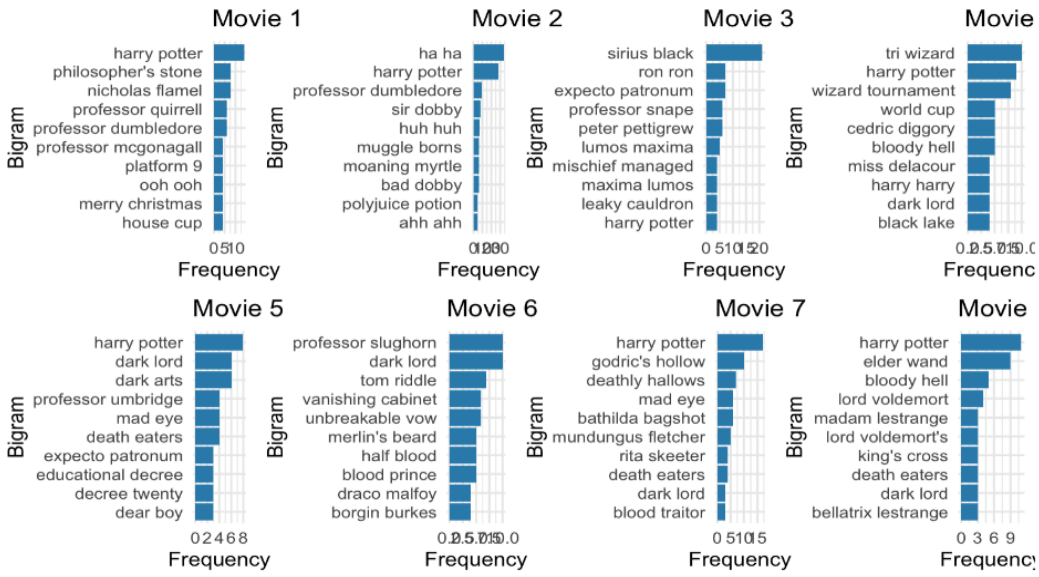


Figure 1-2. Bigrams for for each Movie

Since the movie scripts are categorized by the lines spoken by each character, I was able to easily compare the changes in the number of lines by each character. As expected, Harry had the most lines in all the films. However, Ron and Hermione did not always follow after that, and the characters that were heavily addressed in each film were different. For example, the first film, "Harry Potter and the Philosopher's Stone," features lots of descriptions of Harry's background and family, so all the Dursley family members ranked in the top 10. Most interestingly, Albus Dumbledore did not have many lines during the first three films. However, as the series progressed, the number of lines he had significantly increased, indicating his crucial role in the latter half of the series. In the last film, even though he died in the previous movie, Dumbledore was the fourth person to have the most lines. Another interesting thing is that although Voldemort or Tom Riddle was the main antagonist, he did not have many lines. He was the sixth most frequent speaker throughout the series, and the difference between him and Severus Snape, who was seventh, was not even significant.

In conclusion to the movie analysis, I compared the sentiment of the movie dialogues. As I mentioned in the introduction, I expected the sentiment analysis to reflect the mood of the series that the movie starts with a bright atmosphere and then gets darker toward the second half. However, as shown in

Figure 1-3, there was no significant relationship between the progress of the series and the change in the sentiment. The negative sentiment dominates throughout the whole series. In contrast to what was anticipated, the negative sentiment of the first two films were the highest in the whole series. Only Movie 4, "Goblet of Fire," and Movie 6, "The Half-Blood Prince," showed a predominance of positive sentiment. In fact, except for "The Half-Blood Prince," it can be said that negative sentiment decreases rather than increases as the movie progresses. Beyond positive and negative sentiment, I also compared the remaining eight nrc sentiments as shown in Figure 1.4. Similar to what was found earlier, with the exception of "The Half-Blood Prince," negative sentiments such as anger, fear, and sadness were ahead, while positive sentiments such as joy, trust, and anticipation did not show a clear trend.

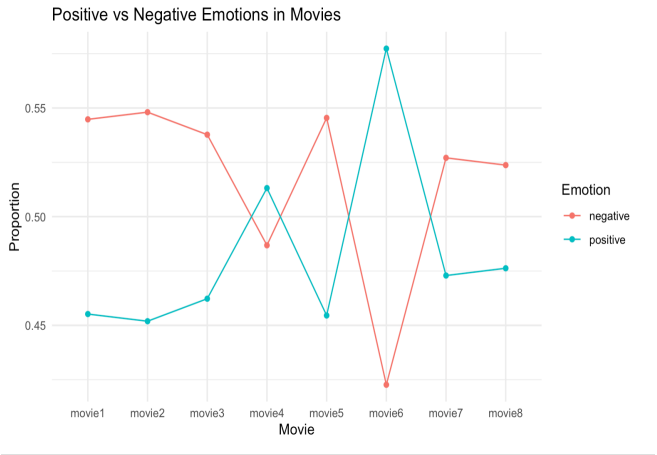


Figure 1-3. Positive and Negative Sentiments Trends

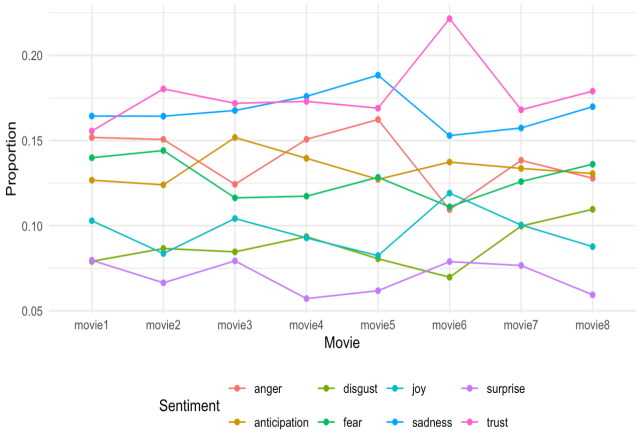


Figure 1-4. Other Sentiments Trends

As a result of analyzing positive and negative sentiments by chapter, Figure 1-5 shows Movies 4 and 6 had the least negative sentiment just like we confirmed in Figure 1-3. In particular, Movie 6 showed a noticeable decreasing trend into negative sentiment. In addition, Movie 2 was identified as the film with the most variation in sentiment.

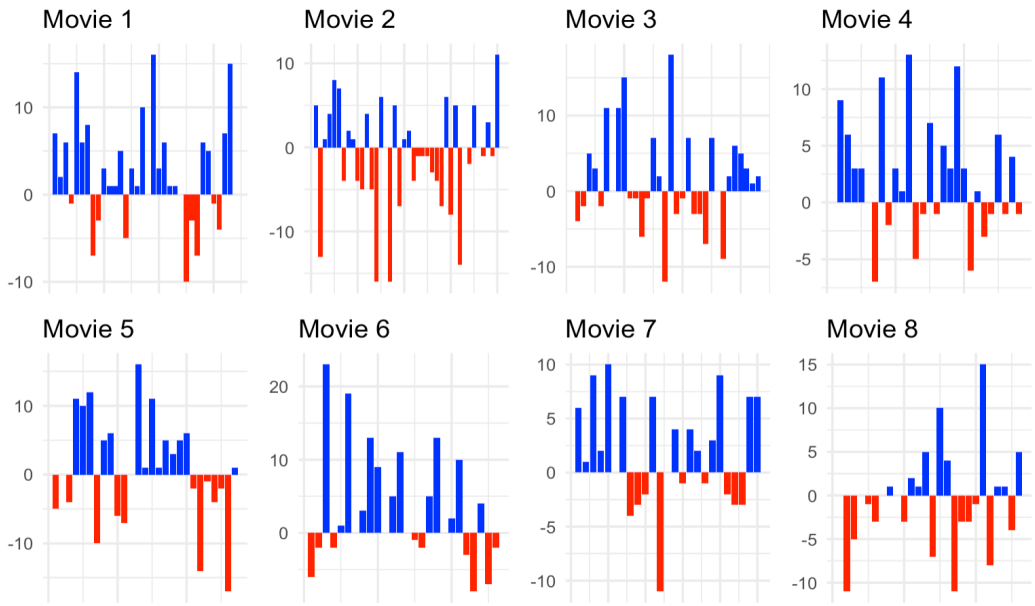


Figure 1-5. Sentiments Trends on Each Movie

4. Comparison between Movie and Book

The next part of the text analysis is a comparison of Harry Potter books and movies. Prior to the text analysis, I looked for the readability index of the book as I did for the movies. As a result, as depicted in Figure 2-1, the readability index increased towards the later books in the series, which could be interpreted as a reflection of the main characters aging and the overall content of the books becoming more serious.

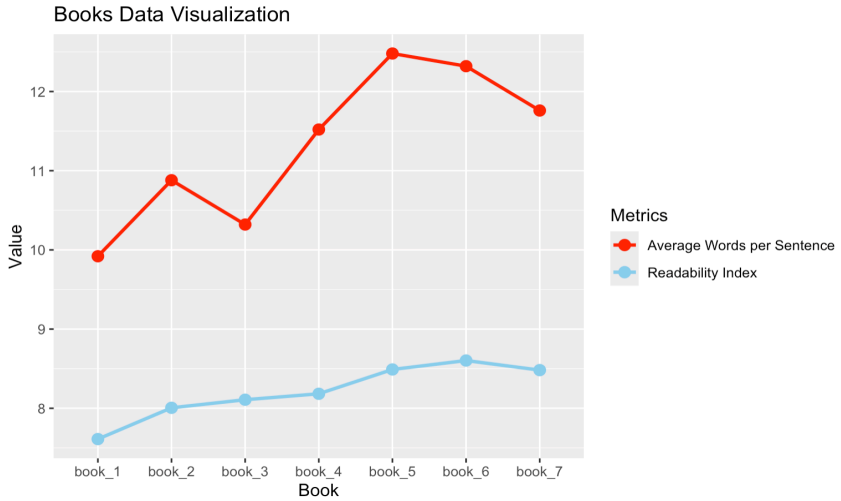
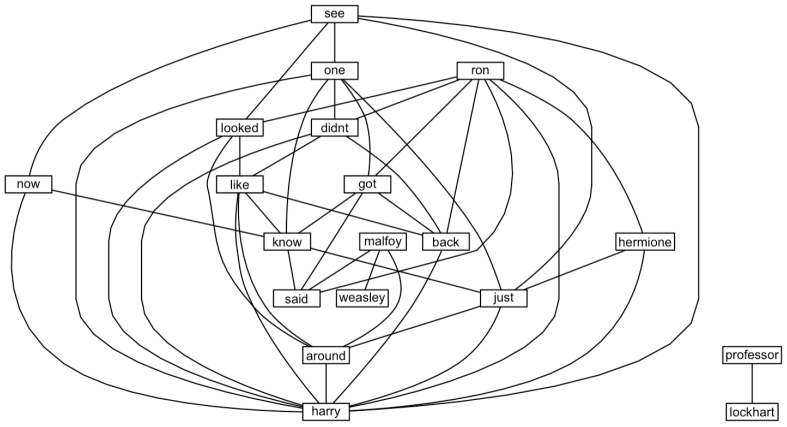


Figure 2-1. Readability Index and Average Words per Sentence

Similar to the movie analysis, I performed vectorization on the text of the book and performed some pre-processing and data cleaning of the text data to prepare for the text analysis. Upon creating and sorting the document term matrix, the results were similar to those obtained from the movie scripts: "harry," "ron," and "hermione" were the most frequently appearing words across all books, with "said," "back," "one," and similar. Words that could specify certain movies, such as "Lupin," "Sirius," and "wand," did appear, but they were not prominent.

I also created correlation plots that can check the relationship between words. When I created a correlation plot for movie analysis, I couldn't find any significant features, perhaps due to the relatively small amount of data. When using book data to create correlation plots, there were no remarkable characteristics for each series. However, as shown in Figure 2-2, we can see that the Weasley and Malpoy families are high-corrected in Book 2, “The Chamber of Secrets.” Also, what can be found in Figure 2-2 is that for book 7, “Deathly Hallows,” Voldemort, Harry, and Wand are linked together. It was fascinating to visually check the main conflict of the movie through text analysis.

Harry Potter 2 and the Chamber of Secrets.txt



Harry Potter 7 and the Deathly Hallows.txt

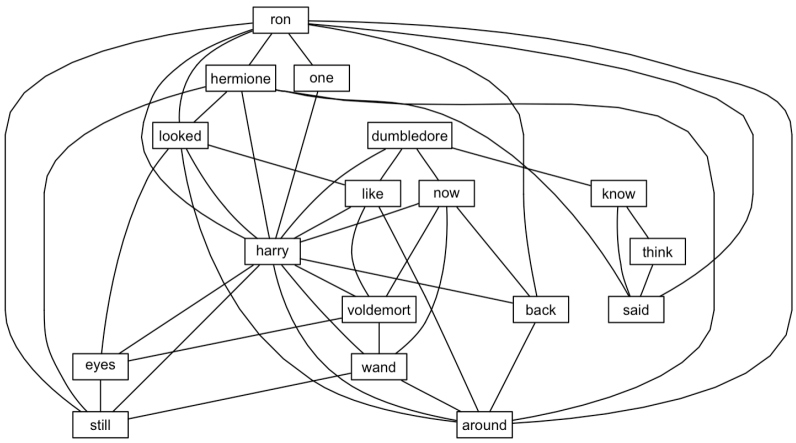


Figure 2-2. Correlation Plots for Book 2, 7

I also conducted a text analysis on the sentiment of the book series. Prior to the analysis, in order to compare the movie series and the book series one by one, the last two movies of the eight-part series in movie, “Deathly Hallows” Part 1 and Part 2, were combined into a single film. The comparison showed the most significant difference in sentiment for Movie 6, "The Half-Blood Prince." Figure 2-3 shows that "The Half-Blood Prince" had a higher presence of positive sentiment and less negative sentiment in the movie, whereas the original book contained much more negative sentiment. While the movie had more trust sentiment than negative, the original book showed negative sentiment dominating, with trust ranking only fifth. Similarly, anger sentiment appeared the 8th most in the movie, however, it appeared the 4th most in the original book.

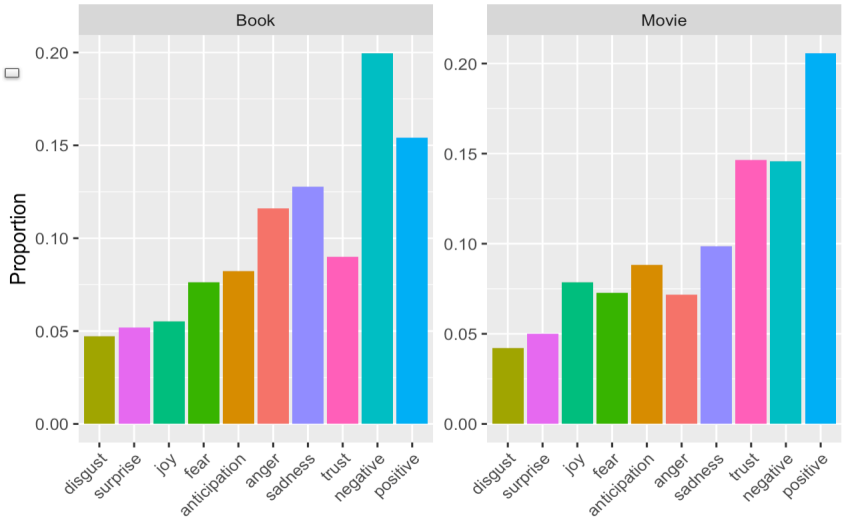


Figure 2-3. Comparison between The Half-Blood Prince

For a more detailed comparison,I organized each sentiment by series and plotted them. Figure 2-4 shows that the negative sentiment appeared more in the original book, and positive did not exceed negative sentiments in any book. Additionally, unlike the movies where the sentiment graph showed significant fluctuations, the books displayed less variation in sentiment changes. As a result, it is confirmed that the movie contains a lot of more negative sentiments, and maintained similar sentiment levels across the series. This could be due the original novel was written by Rowling alone, whereas since the movies’ directors changed several times it led to significant modifications in sentiment. Also, as shown in Figure 2-4 and 2-5, out of all the series, "The Half-Blood Prince" showed the most discrepancy between sentiments, and the series 5, "Order of the Phoenix," showed the least difference.

Figure 2-4. Comparison of Sentiment Trend of 'Positive' and ‘Negative’

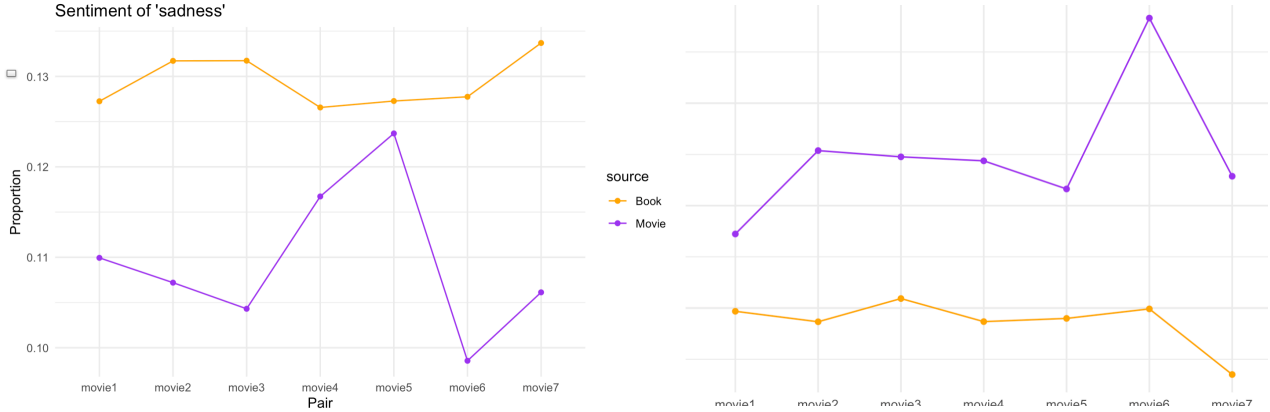


Figure 2-5. Comparison of Sentiment Trend of Sadness and ‘Trust’’

5. Comparing Dialogues by Hogwarts Houses

Finally, I analyzed the dialogue by Hogwarts house. Hogwarts has four dormitories, but the main dormitories introduced in books and movies are Gryffindor and Slytherin. To analyze the dialogue from members of these two houses, I first classified the characters into the house they are from, and then gathered all the dialogue spoken by characters from each house. As a result, I got 4483 pieces of dialogue from Gryffindor and 717 from Slytherin. This imbalanced volume of data is due to the main characters of the Harry Potter series; Harry, Ron, and Hermione are all from Gryffindor.

After classifying the dialogue by house, I proceeded with vectorization and completed data cleaning. I created word clouds for each house, as shown in Figure 3-1, it was possible to identify the words commonly used in each house. I discovered that each house refers to the same person or object by different names. For example, the word most described in Gryffindor is harry, and the word most

described in Slytherin is potter. This actually refers to the same person, Harry Potter, and you can see that each dormitory calls him by a different name. Similarly, Voldemort is referred to differently; in Gryffindor, he is called "Voldemort," while in Slytherin, he is referred to as "lord," as shown in the word clouds.

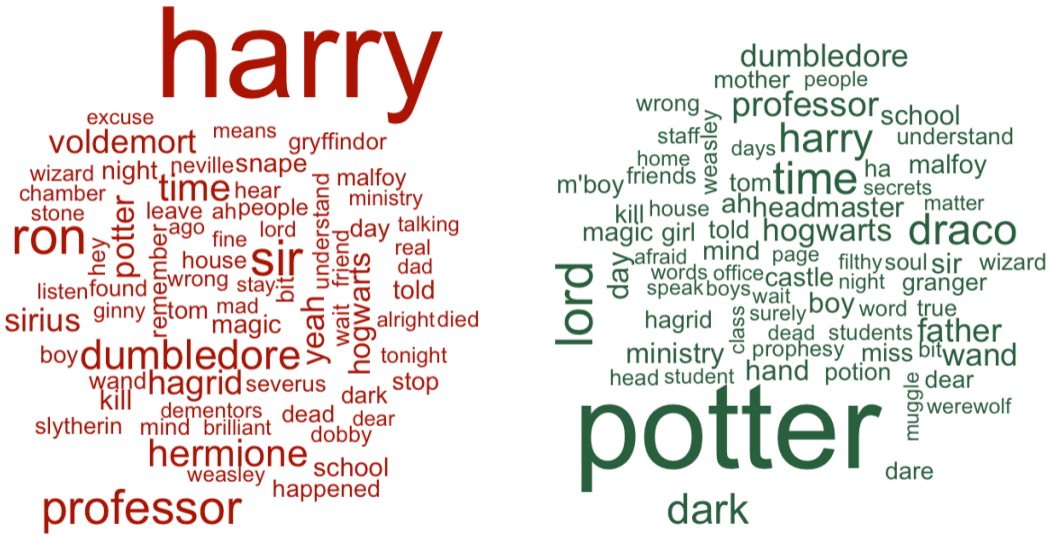


Figure 3-1. Gryffindor and Slytherin Word Cloud

In order to analyze the speaking habits of characters from Gryffindor and Slytherin, I conducted a sentiment analysis and compared themFor easier comparison, I summarized the findings in a graph Figure 3-2. Surprisingly, characters from Slytherin tend to use more positive sentiment, while those from Gryffindor tend to use more negative sentiment. I expected the opposite result because in the series, Gryffindor is home to the main protagonists such as Harry, Ron, and Hermione. However, Slytherin is home to antagonists such as Voldemort and Malfoy. In addition, Slytherin used more trust sentiments than Gryffindor, whereas Gryffindor used words related to sadness more often.

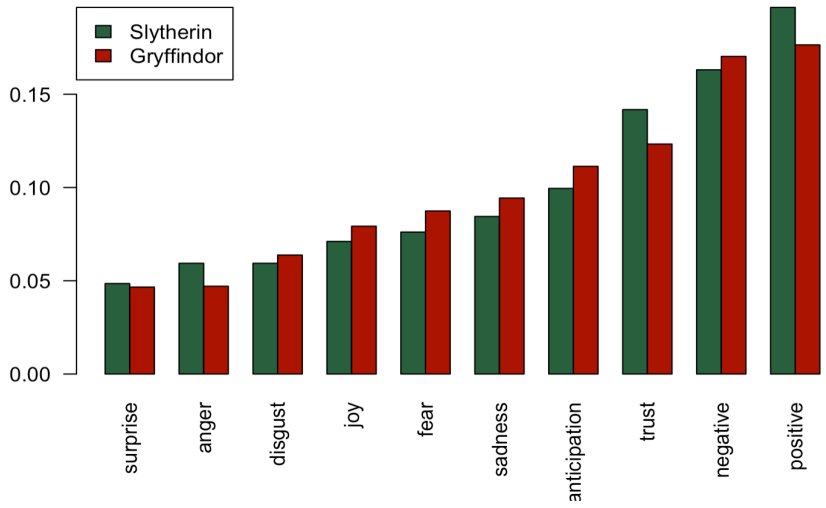


Figure 3-2. Comparison between Gryffindor and Slytherin Sentiments

Finally, I analyzed the spells most frequently used from each house. The results showed that Gryffindor favors spells like “Expecto Patronum,” “Lumos,” and “Expelliarmus,” which are more self-defensive and not intended to harm others. On the other hand, Slytherin prefers more aggressive spells such as “Crucio” and “Avada Kedavra,” which are curses that can harm or even kill others. Although the sentiment analysis showed Gryffindor revealed more negative aspects, the actual spells they use reflect their positive intention of not harming others.

Additionally, I created a dialogue classification model using the data categorized by house. I wanted to find out how accurately each line could be correctly determined by using a random forest. For this, I split the data of each house into 70 percent training and 30 percent testing. Then build a simple Random Forest model. The result of a confusion matrix is shown in Figure 3-3, and the model showed a test accuracy of 84.8%. This was significantly higher than I expected, especially since the previous sentimental analysis gave the opposite result to my anticipation. Despite the model being a simple model without any tuning, it achieved approximately 85% accuracy.

	Gryffindor	Slytherin
Gryffindor	1276	69
Slytherin	167	48

Figure 3-3. Confusion Matrix of the Random Forest Classification Model

6. Conclusion

I analyzed the Harry Potter series with three different topics. The results indicate that although the movies change darker as they progress, this shift is not reflected in the characters' dialogues according to sentiment analysis. This suggests that the overall atmosphere changing darker is likely due to tragic events in the movie or the visual elements, rather than the dialogue of the characters.

In addition, to verify the differences between the movies and the books, I conducted text analysis, focusing on sentiment analysis. The results showed a clear difference in sentiment. The sixth series exhibited the most significant changes when transforming from the original novel to the movie. Not only were there differences in each book or movie, but also in the overall flow of the work, the book showed a certain percentage of sentiments, while this unity was not met in the film.

In the analysis of dialogues from Gryffindor and Slytherin members, contrary to expectations, Gryffindor characters used more negative words and fewer positive words. Additionally, by building a Random Forest model for classifying the dialogues, the model could accurately identify the house of the characters with approximately 85 percent accuracy.

However, this analysis has several limitations. While I used sentiment to analyze the differences between the movies and the books, a more detailed analysis of word usage in addition to sentiment could yield more accurate results. Additionally, when comparing sentiments across houses, providing a comparison of specific words associated with each sentiment could offer more insights. Lastly, the classification model I created was built with imbalanced data; implementing the model with a balanced dataset could result in a more precise model.

Despite these limitations, this analysis has provided meaningful insights into the globally beloved Harry Potter series, revealing many interesting points that differ from what many of us might have expected.