Final Project: Sentimental and Thematic Analysis of the *Twilight* series

Jolie Ma and Emily Pan

December 11, 2023

Introduction

The *Twilight* series was published in the early 2000s and became one of the most popular YA series in 2008 and 2009, with all four main novels placing in the top four positions in USA Today's bestseller lists for those years (McLoughlin 2022). The series became infamous through its movies, earning at the box office $3.3 billion globally (McLoughlin 2022). As the movie series is mostly faithfully adapted from the book series (Tyler and Gorman 2023), this serves as evidence of the book series' popularity.

As a result of its popularity, the series became a cult classic and is still referenced frequently on social media like TikTok. In fact, the hashtag Twilight has 28.2 billion cumulative views on that platform (Seariac 2023). One of the reasons for its popularity is its comedic perception by its audience members. Jackson (2021) describes the series as "It's good actually, and for the same exact reasons why we all said it was bad". People ironically enjoy the series, and understanding how and why they do so could help other writers or content creators entertain their audiences.

Thus, studying these reasons and the elements of this book series could help those in the comedy industry. If they can replicate the comedic elements of this popular series, then they could use those elements to create popular content. Thus, our research question is: what are the thematic and sentimental elements of the *Twilight* series?

Data

The data we are using in this research is our own created corpus of the four main books in the Twilight series by Stephenie Meyer: *Twilight, New Moon, Eclipse, Breaking Dawn*. While there are several other books in the series, such as *Midnight Sun, and Life and Death: Twilight Reimagined*, we will stick to these four since they were released later and only as companions to the first book *Twilight*. There was also a novella in the series, *The Short Second Life of Bree Tanner: An Eclipse Novella*, which we also exclude from our analysis due to it not being one of the main novels.

The corpus consists of the text of each book's chapters, a total of 116 chapters and 584860 words. Table 1 shows the breakdown of the corpus by each book.

Table 1: Summary of Twilight Corpus.

Book	Chapters	Words (tokens)
Twilight	25	118397
New Moon	25	132154
Eclipse	27	148383
Breaking Dawn	39	185926
Total	116	584860

Methods

There are three methods we used to capture the characteristics of the *Twilight* series. We used frequency and keyness, clustering, time-series, and sentiment analysis.

We start by performing a pre-processing method called tokenization to split the text into single words, splitting at spaces. We removed punctuation, numbers, and symbols as well as converted everything to lowercase. We then calculated the absolute frequency of all tokens in the whole series as a preliminary analysis to see if Meyer frequently used any special tokens or unique lexicon that would make her series stand out. Then, we used keyness analysis to determine how the frequency of tokens changes from the first half of the book series to the last half of the book series to analyze the changes in sentiment and themes across the series using the metric of log-likelihood (LL) and their corresponding frequency and dispersion. Log-likelihood is typically used to gauge whether the difference in the use of a token is significant; if $LL < 3.84$, then $p > 0.05$ (Brezina 2018), so that's why we use it to measure keyness. The magnitude of the difference between token frequencies are reported as effect sizes. These are calculated using Hardie's Log Ratio.

We used hierarchical agglomerative cluster analysis to gain insights into how the different chapters in each book are similar linguistically (Steinbach, Karypis, and Kumar 2000). Since the two common themes of Twilight are romance and horror (Schmoop 2023), we wanted to see if the chapters would group into 2 cluster groups. Those in the same cluster group would have similar thematic and sentimental elements. We used Ward's minimum variance linkage method because it minimizes the total within-cluster variance and at each step, the pair of clusters with the smallest euclidean cluster distance are merged. We then created a dendrogram with the leaves colored by the 4 books and cut it by the number of optimal clusters identified by using the average "silhouette" method. This method computes the average silhouette of observations for different values of clusters and the optimal number of clusters maximizes the average silhouette over a range of possible values of clusters. We also used Principal Component Analysis to create a biplot to see which parts of speech contributes to each PC and their relationship to the observations and clusters.

Next, we used a time-series analysis because we wanted to visualize the frequency of certain tokens. We tested in particular four tokens from popular scenes or quotes and tested two from the romance and horror genre. It gives a snapshot of how prevalent each genre is throughout the series, and these two are the series' main genres. For the romance genre, I used the tokens "love" and "beautiful" because Bella thinks often about how in "love" she is with Edward and also uses the word to describe her feelings for other characters and also admires the appearance of vampires a lot, especially Edward, often describing him as "beautiful". For the horror genre, I used the tokens "kill" and "danger" because Edward emphasizes that his family and he are dangerous as well as the werewolves, and oftentimes Edward and Bella face antagonists that seek to kill and have to kill. Not to mention, one of Edward's most famous lines is "This is the skin of a killer, Bella!" and was even turned into a meme.

For the calculations for time-series analysis, we analyzed absolute frequency per chapter. We didn't have a function that would calculate absolute frequency across chapters, only the total absolute frequency of a corpus. Thus, we created such a function using the package `quanteda.textstats`. We then plotted the frequency of the chosen tokens throughout the book series. We also want to include different forms of the token because the forms have similar meanings, and we're testing for those meanings and not the exact token, so we used the `stringr` to do so. For example, "love" and "loved" mean the same thing, it's just the latter is in past tense.

Last, we created a sentiment analysis graph for each book we analyzed in the *Twilight* series to visualize the sentiment trends throughout the series. This way, we could see how the sentimental elements changed throughout the series. We split the corpus into sentences and calculated a sentiment score for each. We then use a discrete cosine transform function to transform the data and graph the scores along narrative time. To create these graphs, we used the `syuzhet` package.

Results

We demonstrate our results from our analysis methods discussed in the previous section. We first display our results from our frequency and keyness analysis, then clustering, then time-series, and finally sentiment analysis.

Frequency and Keyness Analysis

Table 2: Frequency and dispersion measures for all tokens in twilight series corpus. Ordered by dispersion.

	Token	AF	Per_10.5	ARF	DP
1	i	24400	4174.951	15540.210	0.093
2	the	24182	4137.650	14845.102	0.084
3	to	16772	2869.765	10770.892	0.044
5	was	11250	1924.926	6853.740	0.074
4	and	10921	1868.633	7077.137	0.059

Table 2 shows the top 5 tokens/words in the Twilight corpus. These words all appear commonly in the English language and thus it seems that these books tend to use general language given their large frequency and dispersion, which makes sense since it's targeted towards a general audience.

Table 3: Tokens with the highest keyness values in the first half of the series when compared to the last half of the series

Token	LL	LR	PV	AF_Tar	AF_Ref	Per_10.5_Tar	Per_10.5_Ref	DP_Tar	DP_Ref
i	240.62	0.29	0	11662	12738	4655.53	3814.46	0.07	0.10
he	195.23	0.41	0	4862	4881	1940.93	1461.64	0.16	0.16
mike	160.99	2.68	0	197	41	78.64	12.28	0.61	0.84
jessica	140.85	3.77	0	123	12	49.10	3.59	0.61	0.91
truck	94.34	2.61	0	119	26	47.51	7.79	0.47	0.84

Table 4: Tokens with the highest keyness values in the last half of the series when compared to the first half of the series

Token	LL	LR	PV	AF_Tar	AF_Ref	Per_10.5_Tar	Per_10.5_Ref	DP_Tar	DP_Ref
her	336.94	0.99	0	2398	907	718.09	362.08	0.28	0.37
renesmee	302.23	8.66	0	270	0	80.85	0.00	0.75	NA
seth	296.24	6.18	0	290	3	86.84	1.20	0.65	0.96
leah	181.76	6.07	0	179	2	53.60	0.80	0.68	0.96
edward	170.10	0.82	0	1649	700	493.80	279.44	0.22	0.43

Table 3 is the keyness table with first half of series as the target corpus and last half of series as the reference corpus. Table 4 is the keyness table with last half of series as target corpus and first half of series as the reference corpus. These keyness tables indicate that the biggest differences in token usage from the first half of the series and the last half of the series are the frequency of appearances of certain characters and pronouns. For example, Mike and Jessica, who are Bella's classmates, appear more frequently in the first half of the series as Bella interacts more with her schoolmates. The difference in usage of "he" and "her" is likely because the later books include more mentions of relevant female characters, like the antagonist Victoria.

Otherwise, it does not appear that Meyer made significant changes in her vocabulary usage in the Twilight series. Her vocabulary usage is also very common, as one can see from the frequency table.

Since Meyer's general language usage is so common, it's unlikely that her general language use contributed to the popularity of the series.

Clustering Analysis

We begin our clustering analysis by first performing a silhouette plot and then creating a dendrogram using the optimal number of clusters determined from our first graph.

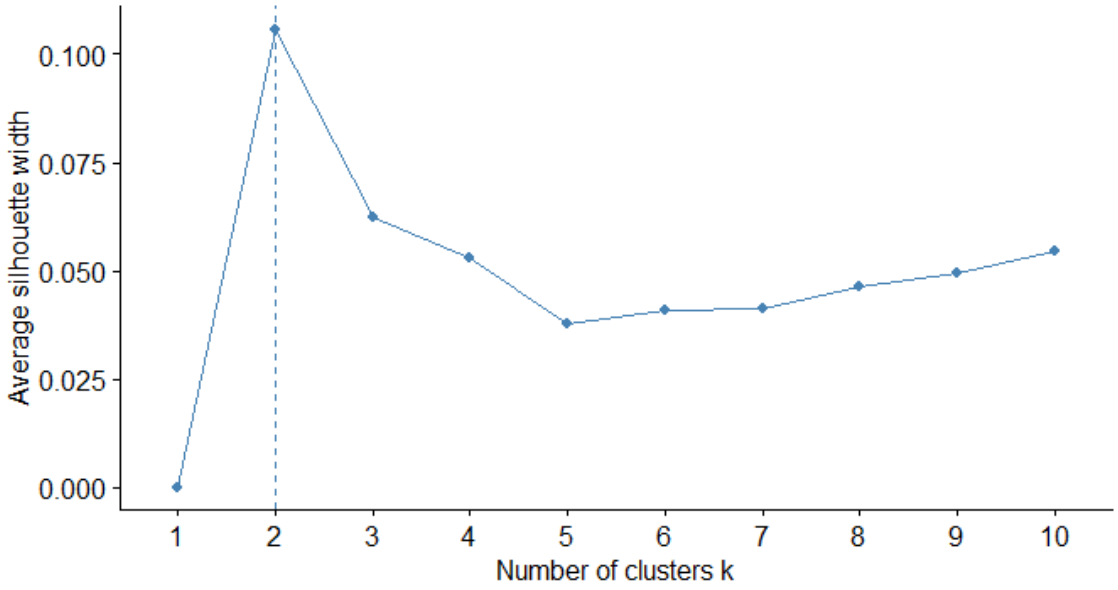

Figure 1: Silhouette plot of Twilight corpus.

Figure 1 is a silhouette plot of the Twilight corpus used to determine the optimal number of clusters for our dendrogram. The vertical dashed line helps us choose how we should "cut" our dendrogram - into 2 clusters. This is consistent with the 2 common themes of Twilight books. To further look into which chapters and books fall into which cluster, we created a dendrogram.

Figure 2 shows the cluster dendrogram of the Twilight corpus, with the leaves colored by each of the 4 books - *Breaking Dawn* is black, *Twilight* is light blue, *New Moon* is green, and *Eclipse* is red. As previously stated, we split the dendrogram into 2 clusters since the average silhouette method suggested that was the optimal number of clusters. This suggests that the 2 clusters could be split by the two main themes of the Twilight books: romance and horror. Since the Twilight series is considered more of a romance theme than horror and the cluster on the right has more leaves, it suggests that the right cluster is romance chapters while the left cluster is horror chapters.

To further explore the clusters, we look at the linguistic aspects and how they compare to the observations and clusters. Figure 3 is a bi-plot showing the top 10 parts-of-speech with their contributions to principal components 1 and 2 and the points colored by the 4 different books. We observe grouping by the different books. The clustering pattern appears to be based mostly on the 2nd principal component (Dim2). *Twilight* and *New Moon* generally have a negative Dim2 value. *Eclipse* seems to have both positive and negative values of Dim2. *Breaking Dawn* appears to have a positive Dim2 value. We do see some potential outliers on the bottom right of the bi-plot. In terms of the variables, noun_nn, pron_prp, and adp_in are positively associated with the 2nd principal component. This suggests that these parts-of-speech are more common

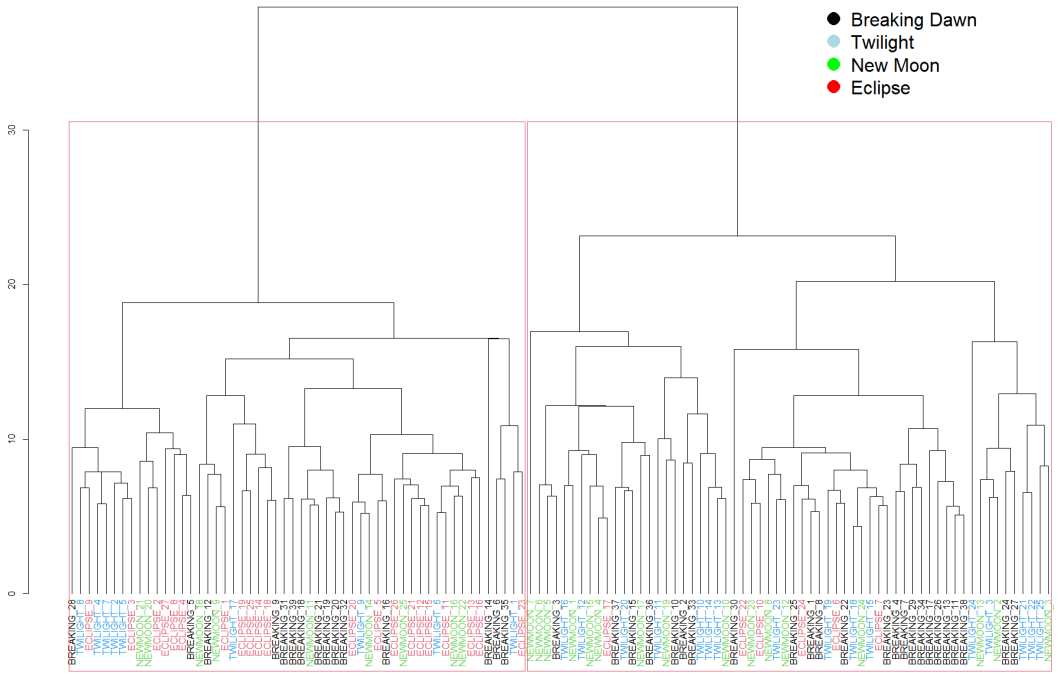

Figure 2: Cluster dendrogram (using Euclidean distances and Ward's method) of Twilight corpus cut into 2 clusters and colored by book.

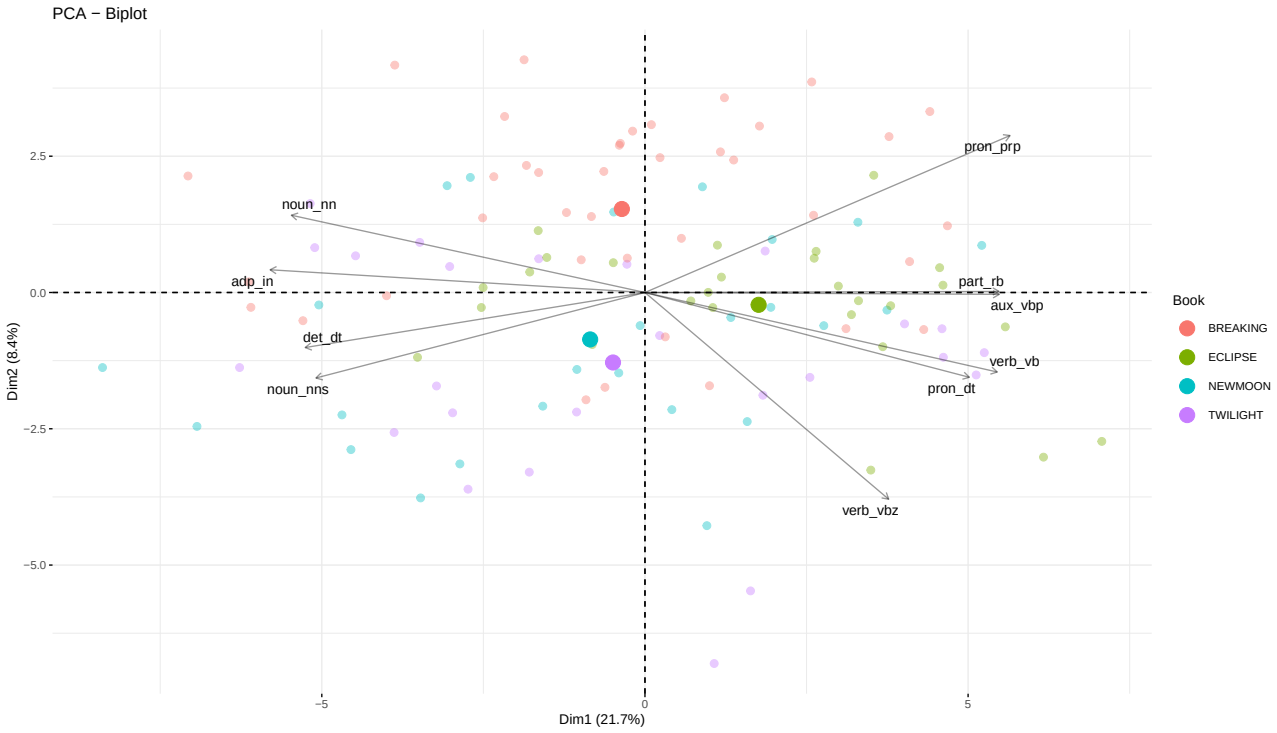

Figure 3: Biplot showing the variables with the 10 highest contributions to principal components 1 and 2 with points colored by book.

the *Breaking Dawn* chapters. On the other hand, det_dt, noun_nns, verb_vbz, pron_dt, and verb_vb are negatively associated with the 2nd principal component. This suggest that those parts-of-speech are more common in *Twilight* and *New Moon*. The variables part_rb and aux_vbp are close to pointing to the right, indicating that they may be more common in *Eclipse* chapters.

Time-Series of Tokens

We chose two tokens for a preliminary analysis of both genres. Note that in our graphs below, we created red dashed lines to divide up the graphs into the four books.

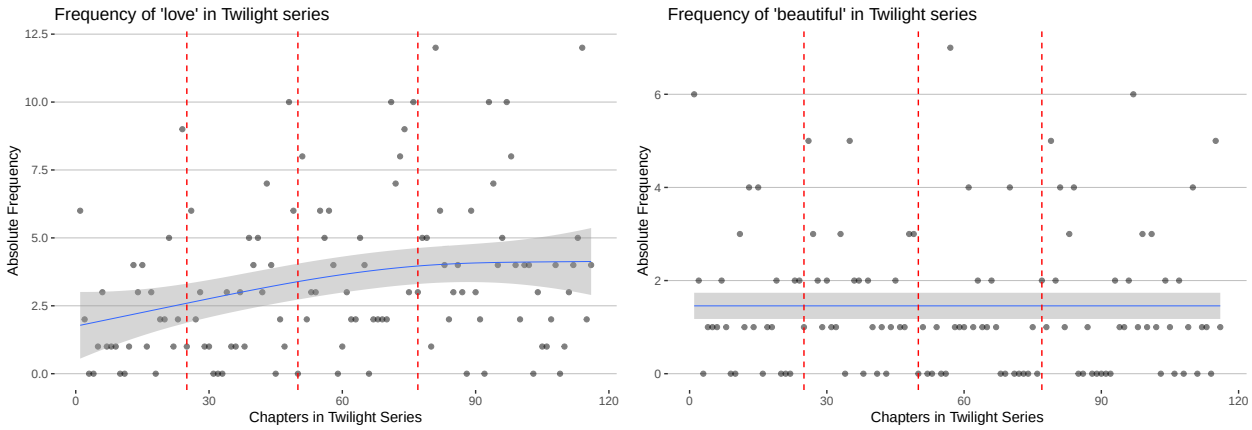

Figure 4: Time-series plots (romance genre)

For the romance genre, we analyzed the words "love" and "beautiful" since these phrases tend to happen frequently in this genre. Figure 4 graphs displayed that tokens similar to "love" slowly increased in frequency across the book series and that tokens similar to "beautiful" did not follow a pattern in frequency across the book series, but still occurred less frequently on average than tokens similar to "love". This indicates that the romance genre was generally likely equally present through the series, though perhaps present in later books compared to the first book. This makes sense as though Bella and Edward fall quickly in love in the first book, there are more demonstrations of their love throughout the series. Not to mention, there is also love between other characters especially between Bella and Jacob, which would explain the increasing frequency.

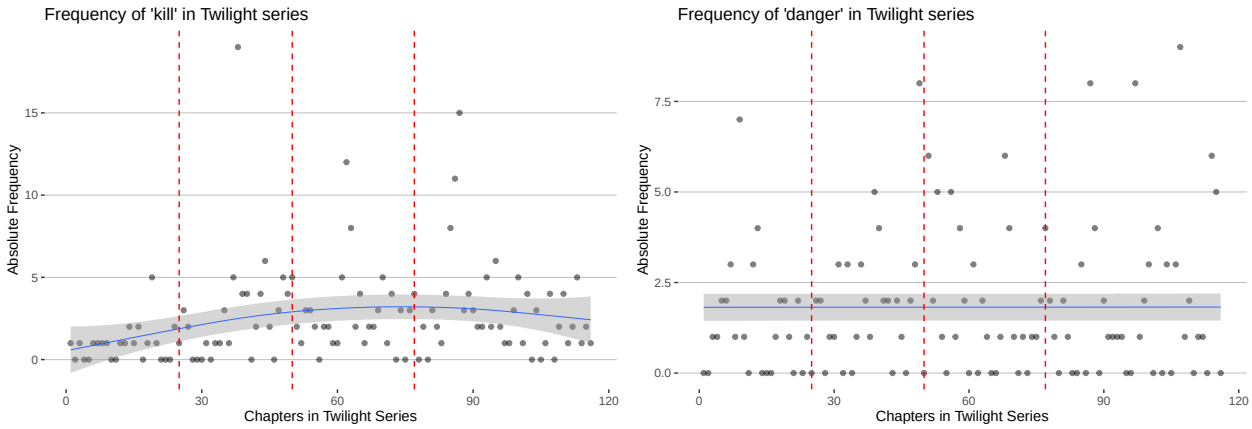

Figure 5: Time-series plots (horror genre)

For the horror genre, we analyzed "kill" and "danger" since these phrases tend to invoke this genre. Figure 5 graphs displayed that tokens similar to "kill" tend to be dispersed more evenly averaging a generally low frequency per chapter, though it increases more in later books and that tokens similar to "danger" occurred in similar frequency throughout each book, though it seems to happen more in later books. This indicates that the horror genre was generally equally present throughout the series, which makes sense as each book has antagonists that scheme to hurt Bella and/or the Cullen family.

None of these tokens have a high absolute frequency in comparison with the total number of tokens, but "love" does seem to appear more frequently than "beautiful", "kill", and "danger". This implies that romance is more present in the series than horror, which makes sense given that people primarily perceive the series as part of the romance genre.

Sentiment Analysis of Twilight Series

We graphed the sentiment trends of the main books in the *Twilight* series and displayed them in Figure 6.

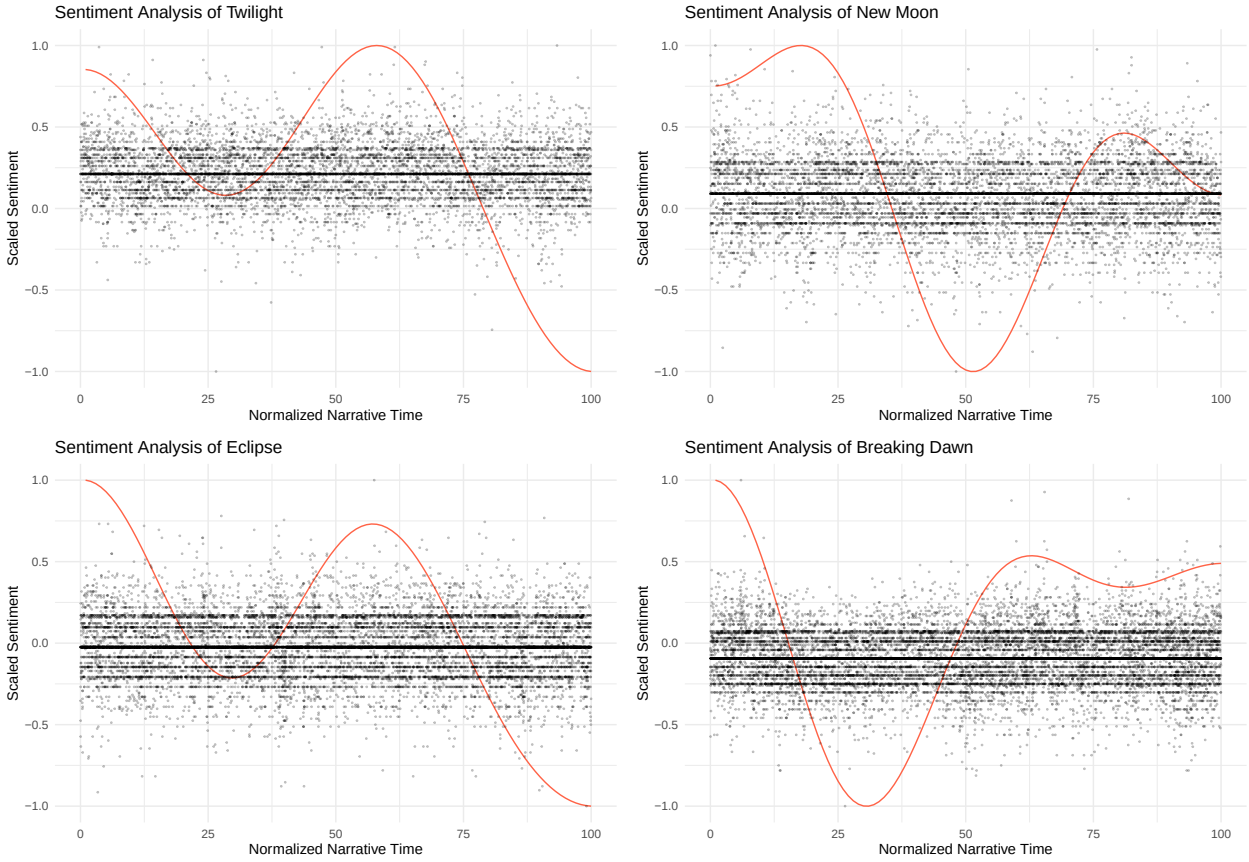

Figure 6: Sentiment Analysis Graphs of Twilight Series

The first book *Twilight* starts with a high positive sentiment in the beginning but ends with lots of negative sentiment. This makes sense since most of the earlier story was about Bella and Edward getting to know each other and falling in love, but the ending was dark since Bella nearly died from a vampire hunting her. Thus, this sentiment analysis of *Twilight* makes sense.

The second book *New Moon* starts with a high positive sentiment, decreases to a high negative sentiment in the middle, and then ends with a slightly positive sentiment. This makes sense since there's a good start

to Bella and Edward's relationship and Bella is happy, but then Edward breaks up with her and the story becomes dark even though she has Jacob for company. Bella gets depressed and commits dangerous actions because they would remind her of Edward. They end up back together in a relatively positive ending after realizing their separation was useless.

The third book *Eclipse* starts with high positive sentiment, decreases to neutral sentiment, increases again to slightly less high positive sentiment, and then decreases and ends in negative sentiment. Some of this doesn't seem to make sense. For example, there's a relatively positive ending as Bella and Edward agree to get married, but that isn't obvious in this graph.

The fourth book *Breaking Dawn* starts with maximum positive sentiment, decreases to maximum negative sentiment, and then increases to medium-high positive sentiment and generally stays and ends there. There were several dark moments throughout the book, like when Bella gets pregnant and Jacob and his pack want her to abort because they worry it'll become a monster, and later when the vampire government comes after them because they thought Bella had created an immortal child, who are generally uncontrollable and endanger vampires as a group, but it doesn't seem that the sentiment analysis graph captures all these moments.

Thus, while the first two sentiment analysis graphs seem to make sense with the plot, the last two graphs don't, but it's interesting to see that trends seem to differ from book to book. This may suggest that an interesting book series should contain variety in its sentiment trends.

Discussion

In our frequency and keyness analysis, we found that Meyer tended to use general language and that the main differences in her use of tokens from the first half of the book series compared to the last half of the book series were the result of the frequency of character appearances. Since her language use is common and meant for a general audience, it's not likely that her lexicon was a big contributor to her books' popularity, though using general language does make it easier for an author to reach a wide audience. Therefore, we would recommend content creators seeking popularity to use general language to connect with their audiences.

In cluster analysis, we found that the optimal number of clusters for our dendrogram is consistent with the 2 common themes in the *Twilight* series, horror and romance. Due to the large number of chapters we have in our corpus, it was difficult to determine how the clusters were grouped. Some further analysis of the various chapters and their linguistics would help in interpreting the clusters better. It would also help us determine what type of elements to recommend to other writers or content creators since if we could isolate the characteristics of these groups, others could then implement these characteristics to increase the popularity of their writing/content.

Through our Principal Component Analysis and biplot, we see that the clustering pattern appears to be mostly based on the 2nd principal component with *Twilight* and *New Moon* having negative values, *Eclipse* with both positive and negative values, and *Breaking Dawn* with positive values. We also see which top 10 parts-of-speech are more common in one book in comparison to another. Since there were potential outliers present, this may have shown relationships that may not be accurate. It would be interesting to further study the phrase structure of the series to see how it changes. Other writers could also learn from the changes in phrase structure and implement it in their writing to produce language similar to the book series and hopefully gain the popularity that *Twilight* has.

Our time-series analysis of the tokens showed some interesting frequency trends in the romance and horror genres, which is that the frequency was likely consistent though they both seemed to somewhat increase in later books. One exception is "love" which generally occurred more often than either of the other tokens we looked at throughout the book series. This implies that the romance genre is more prevalent than the horror genre in this book series, which makes sense as the series is known primarily for its romance. To provide a better sense of how much of each genre each chapter is made of, it would be useful to figure out what words belonged to the romance and horror genres and then measure the total frequency of those tokens. All in all, the prevalence of the two genres indicates that their presence likely contributed to the series' popularity.

Combining the two genres in other content would likely be similarly amusing. Thus, if other authors/content creators wish to produce similar content, they should use both the romance and horror genres.

The sentiment analysis trends made sense for the first two books, but there seem to be some inaccuracies for the last two books. It is interesting though that there are different trends for each book, which may imply that books in a book series should follow slightly different trends to make a series more interesting. However, we would likely need a sentiment analysis of popular fictional book series and unpopular fictional book series to see the differences between them to investigate whether that is true.

References

Brezina, Vaclav. 2018. "Vocabulary: Frequency, Dispersion and Diversity." Chap. 10 in *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge University Press. https://doi.org/10.1017/9781316410899.

Jackson, Gita. 2021. "'Twilight' Is Bad and That's Why It's Good," July 11, 2021. https://www.vice.com/en/article/bvz3mm/twilight-is-bad-and-thats-why-its-good.

McLoughlin, Danny. 2022. "Twilight (Novel Series) Statistics," October 13, 2022. https://wordsrated.com/twilight-novel-series-statistics/.

Schmoop. 2023. "Twilight Genre." https://www.shmoop.com/study-guides/twilight/genre.html.

Seariac, Hanna. 2023. "15 years later, the cultural impact of 'Twilight' is still going strong," November 20, 2023. https://www.deseret.com/entertainment/2023/11/20/23900304/twilight-15-years#:~:text=The%20franchise%20(both%20books%20and,cumulative%20views%20for%20%23twilight%20alone..

Steinbach, Michael, George Karypis, and Vipin Kumar. 2000. *A Comparison of Document Clustering Techniques.* University of Minnesota Digital Conservancy, May 23, 2000. https://hdl.handle.net/11299/215421.

Tyler, Adrienne, and Chloe Gorman. 2023. "16 Biggest Changes Twilight's Movies Made To The Books," October 5, 2023. https://screenrant.com/biggest-changes-movies-twilight-books/#:~:text=Edward's%20backstory%20as%20a%20vigilante,the%20battle%20in%20Breaking%20Dawn..