Discussion of Marvel Movie Review Sentiments Over Time

Andrew Huang and Miguel Flores

Andrew: Wrote full report, interpretation of results, debugging code, and wrote code for preprocessing.
Miguel: Wrote code for all setup and preprocessing, LDA, Sentiment Analysis, Proofread

Prepared for:
Dr. Peter Kramlinger
STA 141B
March 2025

Abstract

In this project, we used Python libraries to web scrape in order to conduct both sentiment analysis and topic modeling on Marvel Cinematic Universe (MCU) movie reviews written in Wikipedia articles. Our goal in this project is to understand the critiques and appraisals of each Marvel movie and how they have changed over time. Our focus will be on all Marvel movies made in the Marvel cinematic universe from 2008 to 2025, with the first movie being *Iron Man* and the last being *Captain America: Brave New World.* We will source the reviews from the critical response section on Wikipedia for each film via requests and lxml libraries. Then we apply nltk and re libraries in order to standardize and clean movie review data for analysis. Finally we will utilize the scikit-learn library for Latent Dirichlet Allocation for topic modeling and VADER for sentiment analysis. We discovered that much of the reviews were mixed and that many negative reviews had key points relating to theme or plot. We also found that films released in the past four years had an increased negative sentiment.

1. Introduction

There was a time in recent history when Marvel movies were synonymous with the superhero genre. This was because these films performed very well with audiences and critics alike. However, in recent years, reviews seem to not be so consistently positive. That's why for this project, we are researching Marvel movie review consensus and the key topics discussed within each review. Our goal is to explore the factors that influence critics and the general public in their appreciation or criticism of a given Marvel Movie. We are investigating the following question:

1. What are the overall critic sentiments for different Marvel movies?

2. What factors contribute to critical negative reviews?

This analysis may give Marvel producers and directors valuable insight on what makes some of their films succeed at the box office, and others fail. By understanding appraisal found in the sentiment analysis on successful Marvel movies and topic analysis, these producers may be able to create a formula to engineer box office hits. For our project we decided to web scrape our sources with Python in order to perform sentiment analysis and topic modeling. This analysis relies on the use of the critical response section of each Marvel movie's respective Wikipedia page. This section of the page focuses on critic reviews and audience reactions for each of the films.

1.1 Legal and Cost Problems

Originally, we planned on using the IMDB API to conduct our sentiment analysis and topic modeling. However, we quickly discovered that it was very costly to use their data since we needed to pay for API access. We then turned to Letterboxd and learned we had to apply in order to use their data as well. We found this to be too difficult, and time-consuming for our purposes. Lastly, we turned to Rotten Tomatoes, which had a similar problem of having an application to use their data. Eventually, we tried to instead scrape the data from any of these websites. But when we tried to web scrape the IMDB website we were unsuccessful because we needed that API key in order to access their website for data in general. Any attempt to scrape was met with a blocked request. This is because, unfortunately, web scraping goes against their terms of service (TOS) and the TOS for all three. We did not want to run the risk of getting blocked or having legal trouble so we stopped our attempts at accessing reviews through these sources. In turn, this led us to pursue Wikipedia for movie data since we were confident in its legality and ease of access.

2. Methods

1. Finding Links to Every Movie

To easily access every film's information, we began by web scraping the *List of Marvel Cinematic Universe films* on Wikipedia. This is a comprehensive source that we knew would contain all of the necessary links to every other page for our analysis. We used the request library to access the Wikipedia HTML for the list of all Marvel movies. We then used the lxml library to parse the HTML page. After parsing, we used xpath to filter through the html content in order to retrieve a list that only contains links to each movie's Wikipedia page. However, during this process, we encountered a problem. A few of the movies listed on this site were unreleased or upcoming. Each of the movies had a release date attached, so to solve this problem we used the datetime library to make sure all movies had a release date before today's date. This allowed us to extract only the films that have been released.

2. Extracting Reviews from Every Movie

The second stage in our project was extracting all information in the critical response section for each Marvel film and storing it in a dictionary. First, we initialized the base URL of Wikipedia and then created an empty dictionary to store the data. Next, we created a for loop that iterates through the movie list that we created in part 1. The "links" we stored can be concatenated to the base url to direct a user to the movie's wikipedia page. So within each iteration of the loop, we concatenated these two values for every link, therefore accessing every movie's page. We then used xpath once more in order to extract the reviews from the "Critical Response" section of each page. Lastly, we made a final for loop to append each movie with its respective review in key-value pairs of a dictionary.

One issue was finding a way to extract all of the paragraphs after the header Critical Response and before the header Accolades. We could not directly call these headers as they were nested in sister div statements. We also had to make sure that if there was an image or another sister div statement in between these two headers it would not include that section and skip over it. Originally, we were planning on using the paragraphs that appeared after the Accolades header, however, we realized not all the Wikipedia pages for Marvel Movies included information in the Accolades section that was relevant. We were unable to obtain the paragraphs after the accolade section for some of the Movies and not obtain for others; therefore we decided to only use the paragraphs in between Critical Response and Accolades.

3. Natural Language Processing

3.1. Preprocessing

Before we begin our Natural Language processing, we must start with preprocessing our raw review data. A problem that occurs in each *key-value* pair is that they have multiple strings in a list as the value for each movie. For our sentiment analysis which we will perform later on, we only want one string for each value term, as VADER, our library of choice, does not interpret list values. Additionally, for our LDA, we will have to remove a large set of custom words. We will explain in more detail why later in this report. So in order to accomplish these things, we used the nltk library for our preprocessing. We used the .join() method to combine the strings in the value lists, removed names using regular expressions, tokenized the values into words, removed stop words and custom words, and then finally joined them all back together for each movie. These new values were placed in a dictionary with their movie keys.

3.2. Topic Modeling with Latent Dirichlet Allocation

Now that the data is preprocessed, we can carry on with our topic modeling. For this step, we used Latent Dirichlet Allocation (LDA), which essentially uses probabilistic models to discover underlying topics within documents. Typically this step would return the top n topics in a set of documents, where n is the number of topics we choose. We chose an unconventional approach to using LDA when we chose only 1 topic. The idea is for the topic that is extracted to be key elements of criticism for each individual movie (e.g., plot or action) as well as emotions surrounding the movie (e.g., bad, good, boring). So this is why we only chose 1 topic; in order to extract the top 5 factors for each film's review. This is also why we incorporated a custom list in our preprocessing step. This is due to a problem we encountered, in that at this stage of the process much of our output was filled with neutral language, making it incredibly hard to conduct any type of analysis, as the words were too neutral, and did not provide information on critical factors of each Marvel film. To solve this problem we had to manually process the output from the LDA and remove any words that were irrelevant to the movie characteristics, essentially forcing the topic to become what we want it to be. For example, we removed words like: "tells", "title", "office", and "pulled" while keeping words like "plot", "action", and "emotional." We believe the reason we are encountering this problem is due to our source, Wikipedia. Had we used a source like IMDB we would not be having this issue as the reviews on IMDB are much more opinionated in comparison to Wikipedia.

For our process, we used the scikit-learn library for our LDA. We created a list called lda_results and iterated through each review in our pre-processed reviews. We then made a matrix of tokens and applied an LDA transformation, and with this transformed data we fit the data and specified the one topic. This returned our top 5 words per review.

Originally, we planned to only output words with the sentiment (without terms like "plot" and "action"), however after we did this we found that it removed all nuance for each movie's review, especially for movies with mixed reviews. So instead, we kept these terms as they add more clarity on review details and nuances.

3.3. Sentiment Analysis

The preprocessing for sentiment analysis was the same as for the LDA step. We proceeded with this step by using VADER for its .polarity_scores() method. This returns the percentage of negative, neutral, and positive words. Normally, for this analysis you would want to use the "compound" score which returns an overall sentiment score on the data. However, the data that was returned to us was largely unusable, as the compound score for each Marvel movie was roughly the same, usually in the high .99 range. This is most likely due to Wikipedia being an unbiased source, thereby using very neutral language. The only category between neg, neu, and pos that seemed to have changed and varied between films was the negative category. Using solely the negative category we can get a basic understanding of the sentiment of each film. As the scale is from 0 to 1, a higher number would signify a higher overall negative critic sentiment on the film, and a lower number would signify a lower critic sentiment. The idea can be applied to positive sentiment, in that a high negative critic sentiment would signify a lower positive sentiment.

3.4. Bag of Words Attempt

Initially we considered conducting a Bag of Words word frequency analysis in order to identify the words that appeared most across each review. However, we later learned that it provided no further analysis or answer to our question. Bag of Words provides a frequency count for words, but the frequency count of words is not useful for our analysis since we are looking for sentiment and common themes in a Wikipedia page. Since this is our source, the amount of times a useful sentiment word appears is negligible because neutral words are much more prominent as this site is not opinionated.

3. Results

Our analysis revealed that the reviews were overall highly mixed, indicating that the movies tend to be very polarizing. Over time we see that sentiment has worsened in terms of the amount of negativity within reviews. We also found that, in general, audiences associate Marvel films with action, fun, and humor. Interestingly, most of the movies in our list that are associated with higher negativity scores lack these 3 words as part of their top 5 words. We will explore these results in depth below.

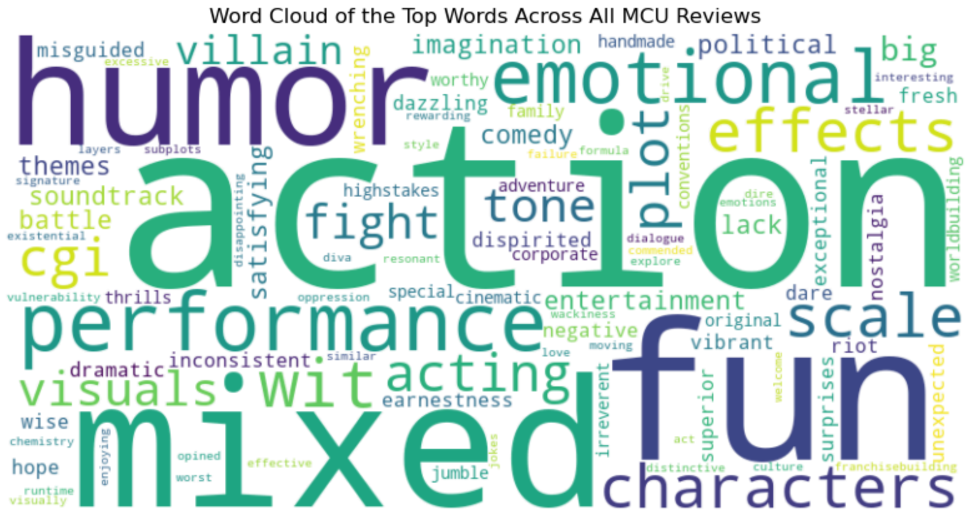

Figure 1. Word Cloud of Top Words Across Marvel Movies

Figure 1 highlights the most prominent words: action, fun, mixed, humor, and performance. The prominence of the word mixed shows many of the movies were polarizing, meaning that many people had widely differing opinions.

From Figure 1, we can see common discussion points on the film included, plot, villains, acting, effects, and soundtrack. This shows a key focus on discussing Marvel movie production and storytelling. Additionally, terms like CGI, visuals, and scale, help to suggest that the visuals and cinematics of the film were also a major focus.

Figure 1 also reveals both the enjoyable and negative aspects of these films. Positive terms such as fun, humor, satisfying, fresh, and entertainment, highlight what audiences appreciated and enjoyed about the films, While negative terms such as, misguided, lack, inconsistent, and dispirited, show the audience's criticisms of the films.

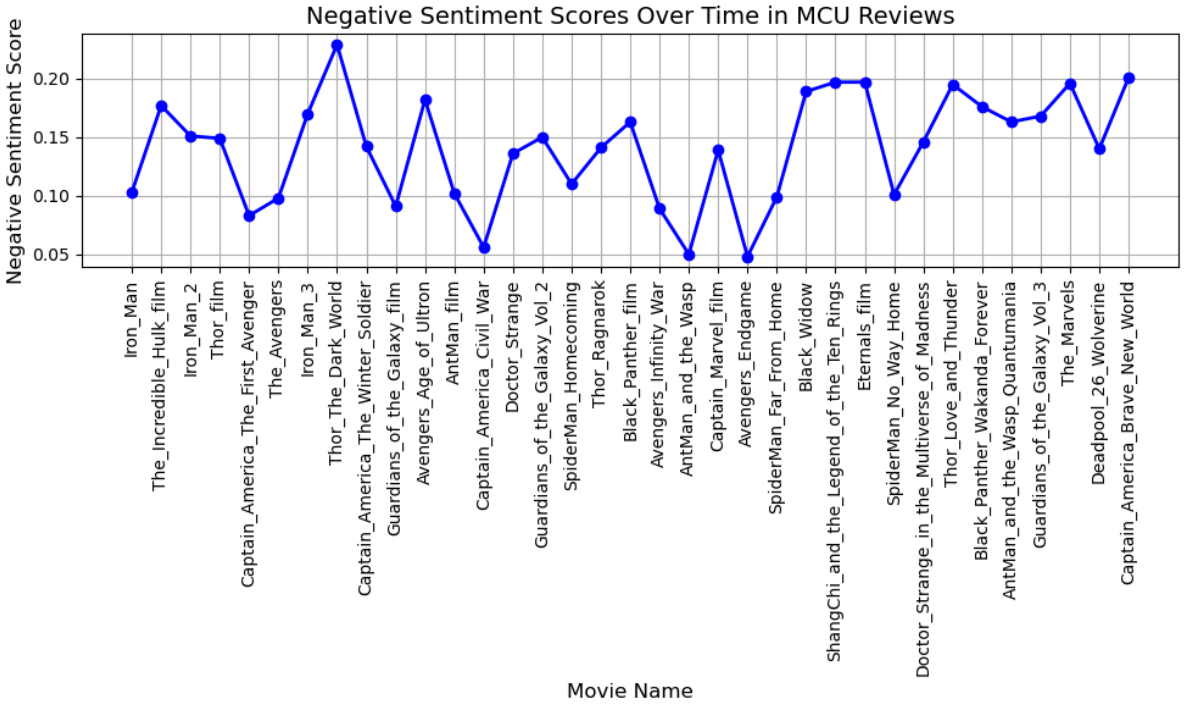

Figure 2. Negative Sentiment Scores for Each Marvel Movie over time

Figure two shows that *Thor: Dark World* had the most negative sentiment score, meaning the most criticism. Additionally, we found that any movie that was released after 2021 to 2025 has a significantly higher negative sentiment score, suggesting that more recent Marvel films have been reviewed more harshly, or have been of worse quality.

In contrast, movies from 2013 to 2019 (its prime mainstream years) received the lowest negative sentiment scores. Marvel movies are categorized into phases: Phase 1 (2008 - 2012), Phase 2 (2013 - 2015), Phase 3 (2016 - 2019), Phase 4 (2021 - 2022), and Phase five (2023 - 2025). From Figure 2 we can see that there were more movies released in a shorter time frame in Phase 4 and Phase 5 than in previous phases. This raises the question of whether the accelerated production schedule has contributed to the increase in negative sentiment.

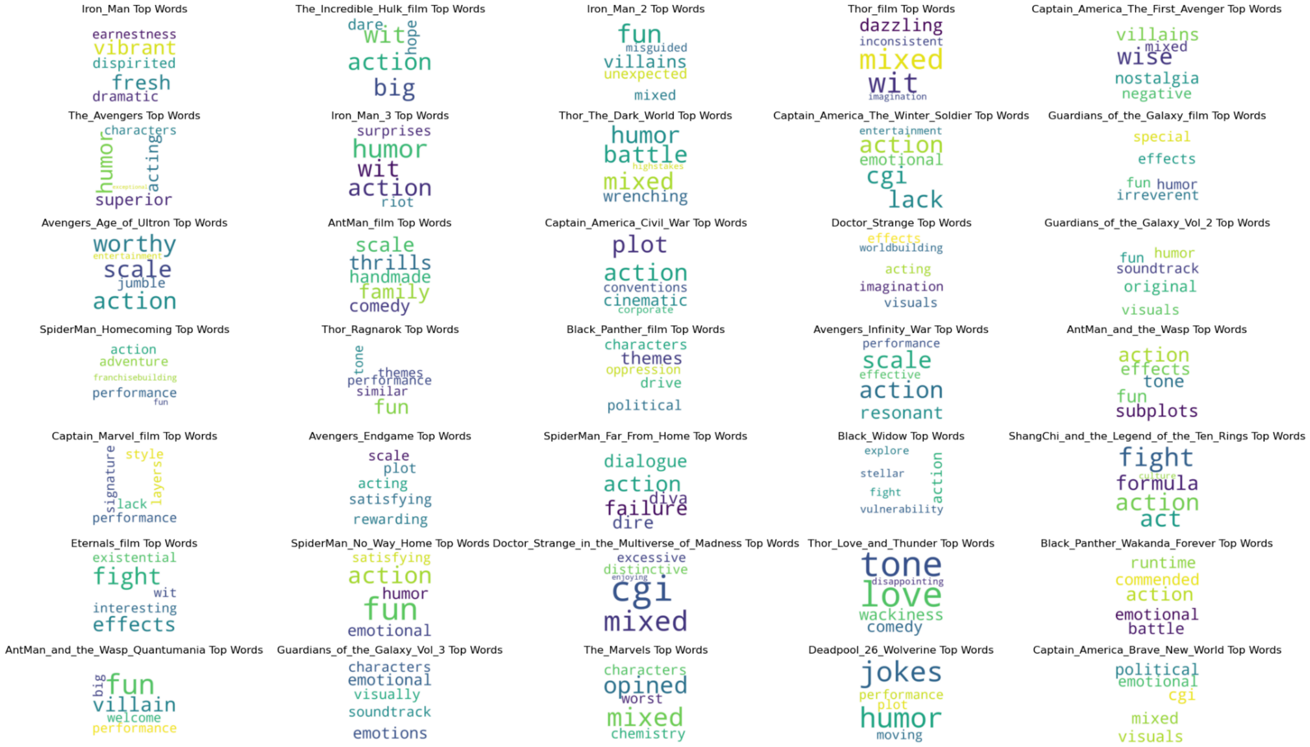

Figure 3. Top words in Review for each Marvel Movie

From Figure 2, we can observe that some of the movies with the highest negative sentiment scores: *Thor: the Dark World, The Marvels,* and *Captain America Brave New World* all share the top word “Mixed”. In Figure 3, we can see that these movies have a higher negative sentiment. This allows us to correlate negative sentiment with having a top word mixed. Additionally, the top 3 words for Marvel which we highlighted action, fun, and humor, seem to be missing from some of the newer poorly reviewed movies such as *Captain America: Brave New World, The Marvels, and Thor: Love and Thunder*. This seems to suggest that these qualities that typically make up what people expect from a Marvel movie are missing in these films and so makes them less favorable to critics and audiences.

On the other hand, some movies with the lowest negative sentiment are: *Captain America Civil War, Avengers Endgame,* and *Antman and the Wasp,* and have the top words from Figure 3 “Action” and “Plot”. This implies that movies with the least negative sentiment have reviews focusing on action and the plot.

One caveat is the fact that *Antman and the Wasp* has an incredibly negative review on review sites, yet our sentiment analysis shows this film has very little negative sentiment. Therefore we must take our results with a grain of salt and must do further research on each film before making any conclusions.

4. Conclusion

The goal of this project was to gain an understanding of the sentiment in reviews for each Marvel film using web scraping to perform sentiment analysis and LDA. We examined all MCU movies released up to today's date for our analysis. Through our analysis, we found that Marvel films have become increasingly critiqued, as they have received significantly more negative reviews in the last four years. This trend may be due to an increased rapid production in Marvel's Phase 4 and 5. This may also be due to a lack of qualities such as “action” and “fun” which seem to be less prominent with these films, straying away from the identity that gave Marvel it's household name. We also found that Marvel movies tend to have incredibly mixed reviews, reflecting audiences polarizing opinions. This indicates that maybe the negative aspects have always existed but now they are far more discussed, perhaps because of the expectations set up by Marvel themselves. Our study shows how sentiment analysis and topic meddling may be used to understand audience reception and trends in the film industry. These insights could shed light on what may be causing more positive or negative sentiment on films.

5. Limitations

Our source for data was an incredibly large limitation for our project in that Wikipedia produces unbiased information, which made it hard for us to get a confident understanding of the audience's sentiment. An additional limitation of our study was not being able to explain the reason for the sentiments that appeared in each film. Perhaps, in the future, we will build a model that helps us understand the factors that influence either positive or negative sentiments. Our analysis also only focuses on Marvel Movies, so a broader focus could also be more beneficial to the film industry, in which we compare by year to other movies released around similar times.