**Final Project Deliverable 4**

**Angela Wilburn**

**Information Technology, Walsh College**

**IT 445 VI Spring 2024**

**Chris Heiden**

**June 12, 2024**

Further Text Analysis of IMBD Sentiment Reviews

In the latest project deliverable, our data analytics team leveraged the logistic regression model of the IMDB sentiment reviews to make insightful recommendations. They investigated further using two advanced data mining techniques: Word clouds and Named Entity Recognition (NER). These techniques were expertly applied to analyze the unstructured data of movie reviews, showcasing the team's proficiency and the robustness of our findings.

The team's analysis examined 20 top-rated and 20 lowest-rated movies, which involved 45,316 reviews, comprising 10,727 positive and 34,589 negative reviews. This extensive analysis enabled the team to uncover trends and offer valuable insights to the marketing team.

Choice of Text Analysis Techniques Reasoning

The IMDB movie review dataset consists of unstructured text data, making it well-suited for applying text analytics techniques. Word clouds and Named Entity Recognition (NER) were chosen as complementary approaches to gain insights from this dataset due to their ability to analyze and extract meaningful information from natural language text.

Word clouds provide a visually intuitive way to identify the most frequently occurring words in each body of text. By generating separate word clouds for positive and negative reviews, we can quickly discern the common themes, topics, and sentiment-laden words that distinguish well-received movies from poorly received ones. This technique offers a high-level overview of the key concepts driving audience perception.

However, word clouds alone may not capture the full context and nuances of the reviews, so Named Entity Recognition (NER) was employed to complement the analysis. NER is a powerful technique automatically detecting and classifying named entities within unstructured text, such as movies, actors, characters, locations, and events. By extracting and analyzing the entities mentioned in positive and negative reviews, we can pinpoint the specific elements (e.g., popular franchises, acclaimed actors, cultural themes) that resonate positively with audiences or the elements that contribute to negative sentiment.

The combination of word clouds and NER provides a multi-faceted view into the factors influencing audience perception of movies. Word clouds offer a broad overview of sentiment-laden language, while NER dives deeper into the specific entities driving positive or negative reactions. Together, these techniques enable a comprehensive understanding of the unstructured review data, surfacing insights that may not be immediately apparent through manual inspection alone.

Results

Data mining techniques carefully analyzed the sentiment of the top 20 highest and the 20 lowest sentiment reviews. This focused approach allowed the team to provide recommendations based on impactful reviews. Common words such as 'movie,' 'film,' 'we,' 'one,' and 'ha' were removed to ensure the word clouds capture less frequently used words without losing the overall value. The most used words in the top 20 positive reviews include 'time,' 'story,' 'character,' and 'life.' Other prominent words are 'love,' 'watch,' and 'think.' (Appendix 1) As for the bottom 20 negative reviews, the most frequently used words include "character," "like," "bad," "even," "plot," and "would."(Appendix 2)

Next, our data team delved deeper into the top 20 movie reviews with positive and bottom 20 based on sentiment and extracted the top 20 most frequently mentioned entities related to each movie review. This rigorous process allowed us to pinpoint the key factors influencing the sentiment of the reviews, providing a solid foundation for our recommendations.

The most mentioned geopolitical entities (GPEs) Hollywood is the most frequently cited, with close to 700 mentions, which suggests that Hollywood is often associated with positive sentiment in the reviews. India and France follow Hollywood in frequency. Events such as 'The Oscars' and historical events like 'World War II' also are mentioned several times in the revies. Iconic movies like 'Toy Story,' 'The Great Escape,' and 'Finding Neverland' are frequently mentioned in positive reviews. Film characters like Marty from Back to the Future and Gandhi and actors like Micheal J Fox and filmmaker Robert Zemeckis are often discussed positively. Fictional places like Hill Valley from Back to the Future and Europe are top locations mentioned in positive Reviews. (Appendix 3- 8)

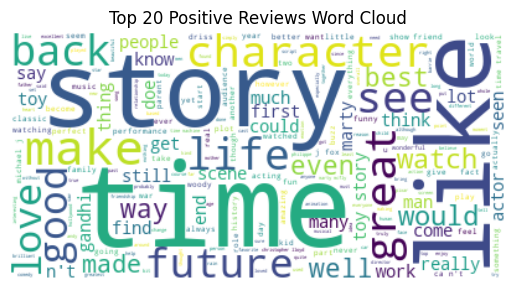
In the negative reviews, Hollywood was also the top-mentioned GPE. Other entities surrounding scary movies also showed up; for example, events had all Titles of scary movies, prominently Get Out. For works of art, almost all were scary movie titles. For people, they were all fictional sci-fi characters; for example, Alien was the top-mentioned person in negative reviews. The top location mentioned was Earth, followed by Europe (9-13)

Recommendations

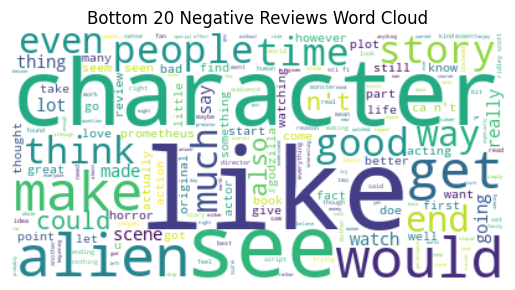
After analyzing the sentiment of IMDB movie reviews, the data team found some key insights to help Wanted Brother Studios make informed decisions about movie investments:

1. Prioritize Character-Driven Stories: Positive reviews often mentioned "character," "story," and "life," indicating a preference for well-developed characters and engaging narratives. Wanted Brother Studios should focus on promoting movies with solid character development and compelling storylines.
2. Highlight Emotional Connections: Positive reviews frequently included words like "love" and "think," suggesting that movies evoking emotional connections are well-received. The marketing team should emphasize movies' emotional aspects and resonance with viewers.
3. Leverage Nostalgia and Familiarity: Positive reviews frequently mention iconic movies, characters, and actors, indicating a positive sentiment towards familiar and nostalgic elements. Wanted Brother Studios could promote movies with recognizable elements or tie-ins to beloved franchises or actors.
4. Avoid Negative Associations: Negative reviews often included words like "bad," "plot," and "would," signaling dissatisfaction with the storyline or execution. Additionally, negative reviews mentioned scary movie titles and fictional sci-fi characters. The marketing team should be cautious about promoting movies associated with these elements or themes.
5. Highlight Cultural Significance: Positive reviews frequently mentioned geopolitical entities like "Hollywood," "India," and "France," indicating a positive sentiment towards movies with cultural significance or representation. Wanted Brother Studios should promote movies celebrating diverse cultural perspectives or highlighting critical historical events.
6. Emphasize Critical Acclaim: Positive reviews mentioned events like "The Oscars," suggesting critical acclaim and industry recognition influence audience sentiment. Wanted Brother Studios could highlight any prestigious awards or accolades received by the movies they are promoting.
7. Explore Opportunities in Historical/Period Films: Positive reviews often reference events like World War II, suggesting that audiences respond favorably to films that explore significant historical periods or events, likely appreciating the educational value, authentic representation, and compelling storytelling that can arise from such subject matter.

By implementing these recommendations, the Wanted Brothers Film Studio Marketing Team can leverage the insights from the text analysis to make informed decisions, promote movies that resonate with audiences, and ultimately drive higher profitability and success for the studio.Appendix 1



Appendix 2



Appendix 3

A graph of a bar graph

Description automatically generated

Appendix 4: A graph of events in blue and white

Description automatically generated

Appendix 5A graph with blue and white bars

Description automatically generated

Appendix 6

A graph of a number of people

Description automatically generated

Appendix 7 A graph of a number of people

Description automatically generated

Appendix 8

A graph of a number of locs

Description automatically generated

Appendix 9

A graph of a graph

Description automatically generated

Appendix 10

A graph with numbers and text

Description automatically generated

Appendix 11

A graph of a bar graph

Description automatically generated

Appendix 12

A graph showing the top 20 most retired persons in negative review

Description automatically generatedAppendix 13

A graph showing a number of people

Description automatically generated with medium confidence