Georgia Institute of Technology

Online Master of Science in Analytics

Analyzing Movie Sentiments

Utsab Mitra

Nabin Kumar Karki

Academic semester Fall 2024

A Project submitted in partial fulfillment of the requirements for the Course ISYE 6740

Table of contents

I. Problem Statement

Movies released over the years have been a way to help human mind escape their mundane life. While serving as a source of entertainment, a movie's rating tells us about the success or failure, and a description of the movie gives us an insight into what to expect. While we can understand what to expect from the movie, how does it correlated to the audience reaction?

Sentiment Analysis [**1**], also known as opinion mining is a major topic in machine learning which aims to capture the subjective information from textual basis. Sentiment analysis is an application of natural language processing (NLP) technologies that train computer software to understand text in ways similar to humans. NLP aims to subjectively comprehend texts as being positive, negative or neutral. Some models may even be used to classify as sad, joy, happy, and other human emotions.

This project aims to build a comprehensive sentiment analysis on over 16k movie descriptions and correlate the findings with Metacritic ratings. The aim is to understand the effect of movie description on the audience reaction.

II. Methodology

The project will utilize Natural Language Processing (NLP) technique to analyze the movie descriptions. Of the many NLP Python libraries available, this project will use the Hugging Face Transformer [**2**]. Hugging Face is most notable for its transformers library built for natural language processing applications. Its Python package contains open-source implementations of transformer models for text, image, and audio tasks. It is compatible with the PyTorch, TensorFlow and JAX deep learning libraries and includes implementations of notable models like BERT and GPT-2 [**3**]. Since we are going to be taking an unsupervised learning approach, we have found the pre-trained model for these transformers to be helpful for our analysis.

The analysis will be carried out in different dimensions:

1. **Sentiment Trend:** Study how the sentiment in movie descriptions has evolved over the decades, and whether there have been significant shifts.

2. **Director's Influence on Sentiment:** Investigate whether certain directors are associated with more positively or negatively described movies, thus assessing the impact of directorial influence on public sentiment.

3. **Critical Response Sentiment Analysis:** Similar to the Director's influence, this study will involve correlating the Metacritic rating with sentiment scores, and determine if a positive description aligns with the higher critical score. This can be an excellent provider of future predictions of movies.

III. Data Collection and Pre-processing

Data Source

The source of the data is from Kaggle [**3**]. The dataset contains detailed information on over 16,000 movies released between 1910 and 2024, along with their corresponding Metacritic ratings. The key features include: Title, Release Date, Description, Rating, Number of Persons Voted, Directed by, Written by, Duration and Genres. With Sentiment Analysis, we could be looking at how the movies fair against directors, writers, duration, number of people voted, among other things.

Data Pre-processing

Firstly, we noticed the data types for certain columns were not as expected, for example the Released data were not of DateTime category, and Number of People voted were not numerical. There were also certain values missing. The missing values were mainly in the ratings column, and thus for the sake of analysis they were removed. Another observation we made was that there were duplicate data points. The duplicate data points were also dropped.

IV. Analysis

The data set was analyzed using the Hugging Face Transformer. This is a publicly available, pre-trained model that includes pipelines for transforming data and producing valuable results for analysis. The model used has been specifically been used for movie reviews analysis [**4**]. The model was trained and tested on a supervised data set. The dataset used was "Large Movie Review Dataset" provided by the Stanford University and specifically said to be used to train Hugging Face Transformers [**5**]. Their dataset contained 25000 training points and 25000 testing point. For our analysis however they were combined and then split randomly to have 80:20 ratio for training and testing respectively. [**Table 1**] displays the evaluation metrics. Thus we can see with an accuracy score of 93.1%, this model can be now used with our data set.

Metric	**Value**
eval_loss	0.3822307586669922
eval_accuracy	0.931
eval_f1	0.9302466639708855
eval_precision	0.930058621386699
eval_recall	0.9304347826086956
eval_runtime	354.557
eval_samples_per_second	28.204
eval_steps_per_second	3.526
epoch	3.0

Table 1: Hugging Face Transformer Model Evaluation Results

After the data set was cleaned and pre-processed, the Hugging Face Transformer was able to take the data set and sign sentiments based on the movie description. To have a better understanding of the sentiment, the labels assigned were either "Positive" or "Negative". The label "Neutral" was not taken into consideration, and any "Neutral" label was considered "Negative". The model also assigned a confidence label to the sentiment to indicate the model's certainty in its predictions. Finally, the sentiment labels and scores were merged back into the main dataset for a comprehensive analysis.

V. Results

The first thing that was looked into is to observe how the sentiments generated faired with the genre. Each movie was either represented by a single genre or categorized into multiple genres. There were 23 unique genres that the movies were categorized into. We decided to work on categorizing the movies into multiple genres and analyze the sentiments from there on. **Figure 1** displays the result.

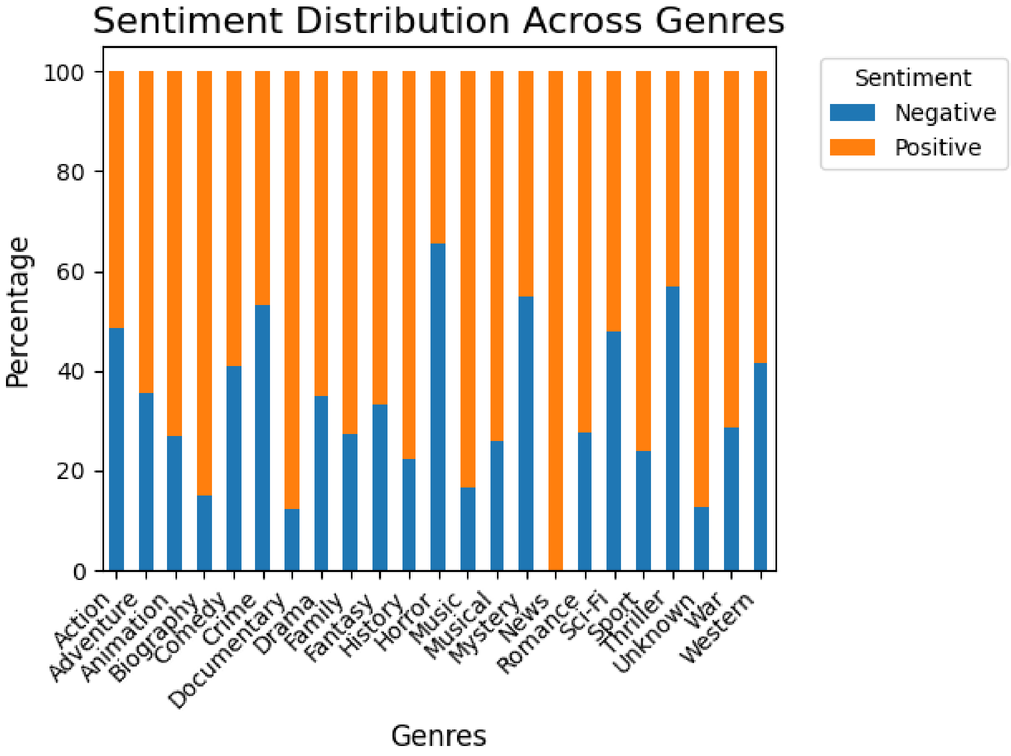

Figure 1: Sentiment distribution across genres

Here we see that movies that were generally considered positive were in a larger quantity over the decades of movie releases. It makes obvious sense that cheerful and joyous movies (that is movies with genre like Comedy, Romance) have a greater positive sentiment than movies with a darker tone (like Crime, Horror). Only a few movies crossed the threshold of 50% negative sentiment, suggesting that movies are generally intended to have a positive vibe.

For analysis looking into how the ratings correlate with sentiments, the ratings were categorized into bins, for example bin (0, 1], bin (1, 2], etc. Over the board, it was observed that the sentiments were fairly constant. People generally rated the movies to the same extent over the entire rating spectrum. **Figure 2** displays the result of such an outcome.

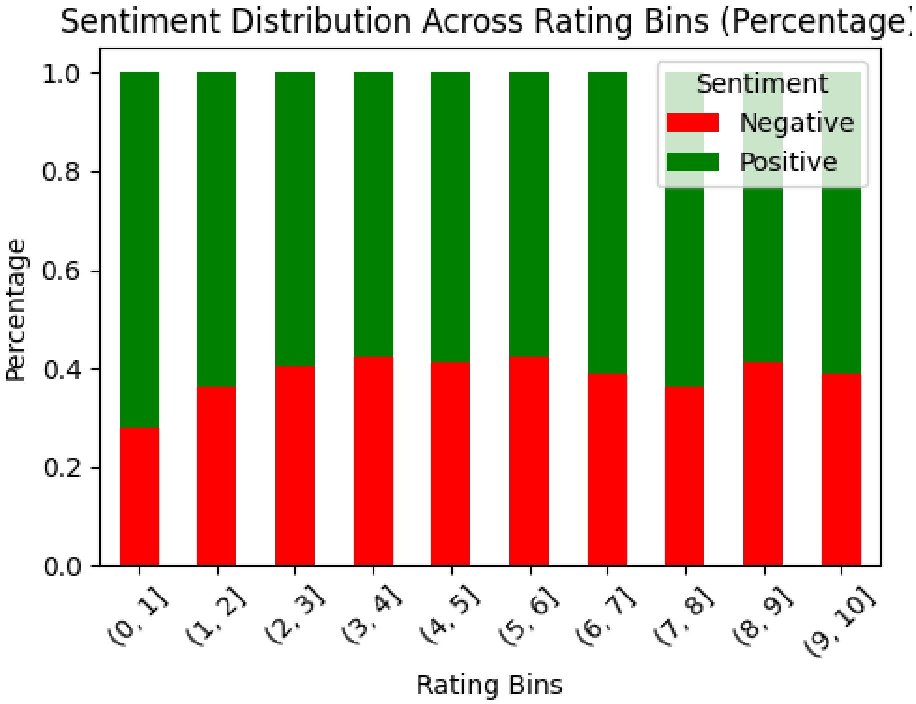

Figure 2: Sentiment distribution across rating bins

Another interesting analysis that came out of sentiment observation is how certain director align themselves with positive toned movies and other directors prefer the darker tone. We made a venn diagram to display this (**Figure 3**).

Some examples of the “Positive” directors are Josh Lawson, Fred Cavayé, Ken Marino, Michael Dweck, Gregory Kershaw; on the other hand some “Negative” directors are Gerard Johnson, Katie Aselton, Gregg Bishop, Spike Lee, Danya Taymor, Meg Ryan. This gives us an idea of what kind of movies an audience might expect coming from these directors.

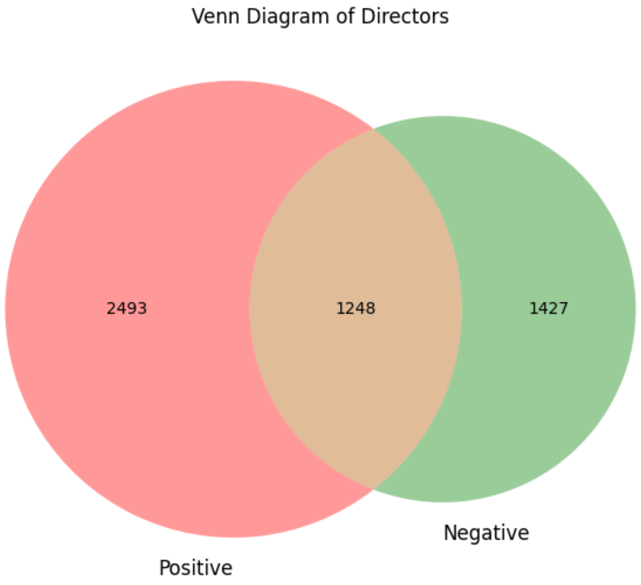

Figure 3: Venn Diagram of directors

VI. Conclusion

The sentiment analysis provided valuable insights into the emotional/sentiment tones prevalent in different movie description. By leveraging NLP and advanced transformer models, it was possible to quantify from textual concept and visualize these sentiments effectively.

By understanding these trends, industry professionals can better tailor their content and marketing efforts to meet audience expectations, ultimately improving viewer satisfaction and engagement.

V. Bibliography

[1] IBM. (2023). What is sentiment analysis? — IBM. www.ibm.com. https://www.ibm.com/topics/sentiment-analysis

[2] Hugging Face. (2024). Hugging Face – On a mission to solve NLP, one commit at a time. https://huggingface.co/

[3] jake3375. (2024, September 5). 16000+ Movies dataset. Kaggle.com; Kaggle. https://www.kaggle.com/code/jake3375/16000-movies-dataset

[4] sarahai/movie-sentiment-analysis · Hugging Face. (2024, January 4). https://huggingface.co/sarahai/movie-sentiment-analysis

[5] Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. (2011). Learning Word Vectors for Sentiment Analysis. The 49th Annual Meeting of the Association for Computational Linguistics (ACL 2011).