Project Report

Introduction:

In this report, we will analyze movie data to uncover trends and patterns in different genres, examine the sentiment of plot summaries, and create predictive models for movie genres. Our goal is to gain valuable insights from the data that can help industry professionals make informed decisions.

Problem Statement:

The main goal of this analysis is to examine movie data to discover patterns in movie popularity among various genres over time, analyze the emotions conveyed in plot summaries by genre, and create a model that can accurately determine movie genres. Our aim is to offer useful insights that can help decision-makers in the film industry make informed choices about movie production, promotion, and distribution.

Data Used:

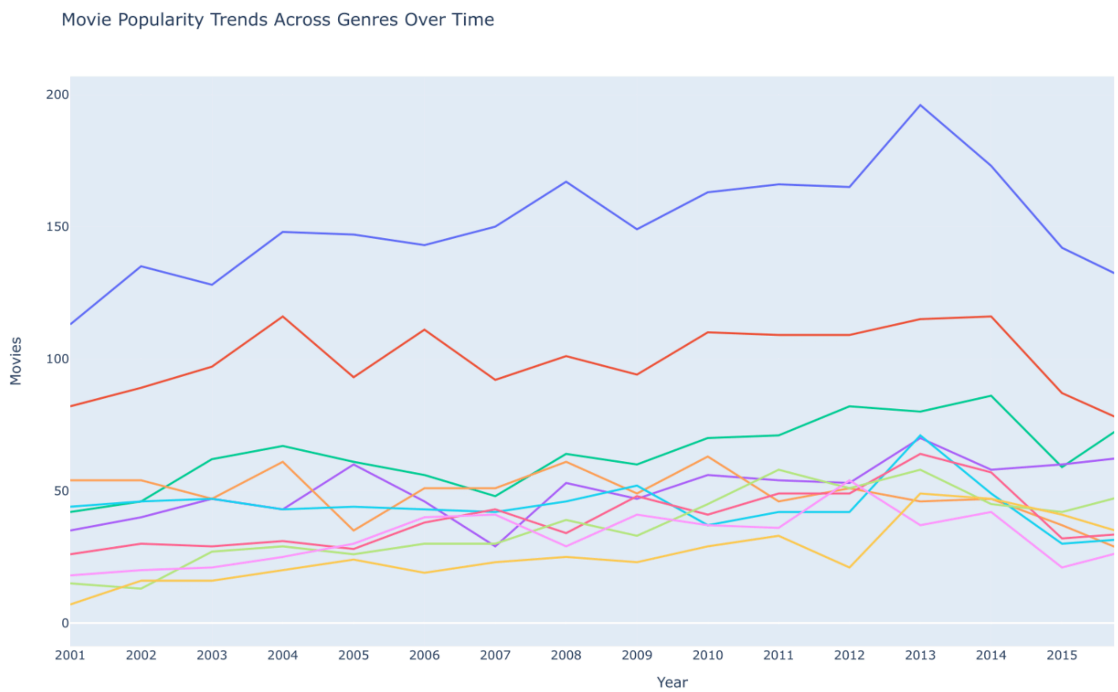
For this analysis, we used two primary datasets: OMDB data and Wikidata. The OMDB data includes details about movie plots, genres, and other metadata, while Wikidata offers more information like publication dates. Before delving into the analysis, we cleaned and filtered the data to remove movies without plot summaries or publication dates, ensuring high-quality results. Further more, I had to perform text classification as movie genre prediction is a multi-label text classification, because each movie can be associated with multiple genres.

Techniques Utilized:

We used a variety of methods to study the data, including visualizing it and using Natural Language Processing (NLP) to analyze sentiment. We explored the data with histograms and line plots to see patterns and trends, and then used NLP to examine the emotions in plot summaries. Machine learning algorithms such as Linear Support Vector Classifier (SVC) were utilized to build predictive models for movie genres.

Results and Findings:

The main result that we are interested in from this project is the estimation of movie genre trends over time. The figure given below illustrates the popularity of movie genres from 2001 to 2018. The number of movie genres released each year, which can be determined using groupby and aggregate, is the simple measure of popularity. It's interesting to note that the total number of movies fell off significantly about 2013 and appears to be getting close to zero in 2016. This might indicate that there is a gap in our dataset. But still, except this point, the results seem promising and do a fair job of capturing the trends.



When you execute the code you will notice that the graph for Movie Popularity Trends across Genres over time is an interactive one and you can investigate about specific points just by simply hovering over the graph.

Limitations:

During the analysis, one of the main challenges we faced was the limited availability and quality of data. Some movies did not have all the necessary metadata, impacting our analysis and modeling efforts. Furthermore, we could have achieved better predictive results by exploring more advanced techniques like deep learning, had we had additional time and resources.

Conclusion:

Overall, studying the movie data offered important insights into trends, sentiment, and predictive modeling of movie genres. Through using different methods and tools, we managed to extract useful information from the data and develop actionable insights. Despite a few limitations, the analysis sets the groundwork for more investigation and study in the area of movie analytics.

Project Experience Summary

During the project, I analyzed movie data to discover trends, sentiments, and predictive patterns in the film genre landscape. I was the only team member and handled all stages of the project, from data preprocessing to model development, with care and skill. Here is a summary of my contributions and experiences during this project:

Data Collection and Preprocessing:

I started my journey by carefully gathering and cleaning the OMDB and Wikidata datasets. I paid close attention to detail, filtering out movies that did not have plot summaries or publication dates. This ensured that our analytical foundation was solid and trustworthy.

Exploratory Data Analysis (EDA):

With the data primed for analysis, I embarked on the exploration of trends and patterns through visualization. Leveraging techniques such as histograms and line plots, I unraveled the intricate dynamics of movie popularity across diverse genres over temporal landscapes.

Natural Language Processing (NLP):

In pursuit of deeper insights, I delved into the realm of Natural Language Processing (NLP) to discern the sentiment embedded within plot summaries. Employing advanced NLP methodologies, I dissected textual data, extracted salient features, and crafted robust models capable of discerning the emotional undercurrents coursing through movie narratives.

Results and Findings:

My efforts culminated in the extraction of invaluable insights and findings, meticulously documented within the report. Noteworthy contributions include the elucidation of movie genre trends over time, the nuanced interpretation of sentiment analysis outcomes, and the critical evaluation of predictive model performance.

Challenges and Learnings:

Undoubtedly, the journey was fraught with challenges, from grappling with data scarcity to navigating the intricacies of advanced modeling techniques. However, each obstacle served as a learning opportunity, fostering a deeper understanding of data analytics and machine learning methodologies.

Conclusion:

Reflecting on my experience, I am immensely proud of the strides made and the insights unearthed throughout the course of this project. As a solo contributor, I honed my skills in data preprocessing, exploratory analysis, NLP, and machine learning, reaffirming my passion for data-driven exploration. Armed with newfound knowledge and expertise, I am poised to embark on even more ambitious analytical endeavors with confidence and proficiency.