Investigating the Temporal Dynamics of Online Movie Reviews: A Multi-Model Approach to Sentiment Prediction and Reviewer Behavior

Shaojie Zhang

Yige Yang

Tian Xu

Yahua Huang

Abstract

In this study, we aim to explore the temporal dynamics of online movie reviews and enhance the overall user experience by analyzing a dataset containing 400k movie reviews from Rotten Tomatoes. Our hypothesis posits that the application of natural language processing and machine learning techniques can reveal valuable insights and improve user experience on movie review platforms. To test this hypothesis, we utilized a range of methods, including exploratory data analysis (EDA), latent Dirichlet allocation (LDA) topic modeling, word cloud visualization, sentiment analysis, TF-IDF, chi-square feature selection, and multiple classification algorithms such as SVM, Random Forest, Logistic Regression, XGBoost, and Neural Networks. Our analysis identified key themes, sentiments, and frequently mentioned words and phrases, in addition to observing changes in the average score and the number of distinct reviewers over time. Based on these insights, we propose the development of a recommendation system for reviewers, which leverages review similarity to enhance the user experience on online movie review platforms. The results substantiate our hypothesis, providing insights into the evolving patterns of online movie reviews and demonstrating the effectiveness of various machine-learning models in predicting sentiment and recommending relevant content to users.

1. Introduction

The advent of the internet has led to the emergence of numerous online platforms dedicated to movie reviews, offering a wealth of information for understanding user preferences and the factors that contribute to the success or failure of films. In this context, our research is driven by

two primary objectives: (1) to cluster and monitor evolving trends in movie reviews, and (2) to develop classification prediction models for reviewer's sentiment. These research objectives are significant, as they have the potential to provide businesses with insights into consumer behavior and inform strategies for enhancing user experience on movie review platforms. Specifically, we aim to address the following questions:

1) How can we provide movie producers with insights into what aspects of movies are currently popular and what people's preferences are?
2) Which factors in movies are likely to generate heated debates among viewers?
3) How can we predict the preferences of reviewers based on the text of their reviews?

In this paper, we present an in-depth methodology for analyzing an extensive dataset of movie reviews from Rotten Tomatoes, incorporating data cleaning and preparation, exploratory data analysis, topic clustering using latent Dirichlet allocation (LDA), prediction modeling, and the development of a recommendation system. Our analysis begins with an introduction to the original dataset, followed by a comprehensive account of the data cleaning and preparation process, which is crucial for ensuring the reliability and validity of our findings. We then provide an overview of the cleaned dataset, including its distribution and other relevant characteristics, and describe the preprocessing steps, such as vectorization, required for subsequent analysis.

Subsequently, we elaborate on our approach to topic clustering using LDA, which enables us to extract meaningful topics and their corresponding words from the movie reviews, thereby addressing the first research question. We then examine our prediction model, initiating with an

exploratory data analysis to identify the top 20 important words in the reviews, employing TF-IDF vectorization and 1 to 3 grams. This analysis helps us understand which factors in movies are likely to generate debate, addressing the second research question. We apply chi-square feature selection to pinpoint the top 30 words with high chi-squared values, which are used for modeling and refining the best prediction model. We compare the performance of various models and discuss their strengths and weaknesses, ultimately selecting the optimal model for predicting positive and negative reviews, and consequently, addressing the third research question.

Overall, this paper provides a robust framework for understanding the dynamics of online movie reviews and demonstrates the effectiveness of natural language processing and machine learning techniques in revealing valuable insights that can be utilized to improve user experience and inform business strategies within the realm of movie review platforms.

2. Data preparation

2.1. Data description

The dataset employed for this research paper, titled "Rottentomatoes 400k Movie Reviews," is sourced from and comprises 400,000 movie reviews from the Rotten Tomatoes platform as of January 5th, 2022. This dataset specifically focuses on reviews that incorporate a scoring scale (e.g., 4/5, 60/100, A+), converting these scales to a standardized 100-point scoring system (e.g., 4/5 equates to 80/100 in the dataset). This comprehensive dataset contains six columns, as detailed below:

Movie: This column contains the name of the movie under review.

Reviewer: This column provides either the User ID or the name of the individual responsible for the review.

Publish: This column lists the name of the publication where the review was published.

Review: This column encompasses the text of the review itself.

Date: This column records the date on which the review was published.

Score: This column presents the score of the review, standardized on a 100-point scale.

By utilizing this extensive dataset, our research is well-positioned to glean valuable insights into the preferences and opinions of a vast number of reviewers across a diverse range of movies, thereby enabling us to address the research questions outlined in the introduction. The standardized scoring system facilitates the comparison and analysis of movie reviews, while the diverse set of columns allows for the exploration of various aspects of the movie review process. This includes examining the impact of the publication source, assessing the content of the review text, and analyzing the temporal dynamics of movie reviews. The dataset thus serves as a robust foundation for our investigation into the dynamics of online movie reviews and the preferences of moviegoers.

2.2. Data cleaning and preprocessing

The data cleaning and preprocessing phases were critical in ensuring the reliability and validity of our research findings. During the data cleaning process, we selected data from 2010 to 2022 to

focus on refining the dataset by taking several measures to generate a balanced and representative sample of movie reviews. Firstly, we calculated the number of reviews per reviewer and determined to select those with a review count between 10 and 100. For reviewers with over 100 reviews, we randomly picked 100 to maintain uniformity across the dataset. As a result of these filtering criteria, our cleaned dataset comprised 8,019 movies, 88,424 selected reviews, and 1,518 selected reviewers.

Subsequent to the data cleaning process, we conducted an exploratory data analysis (EDA) to gain a comprehensive understanding of the dataset's characteristics. This EDA involved examining the text distribution, identifying the top 10 movie reviews, calculating the average number of reviews per person, assessing the number of reviews over time, analyzing the distribution of review years, and evaluating score trends across years. Moreover, we converted the scores into categories to facilitate subsequent analyses, such as sentiment analysis and prediction modeling.

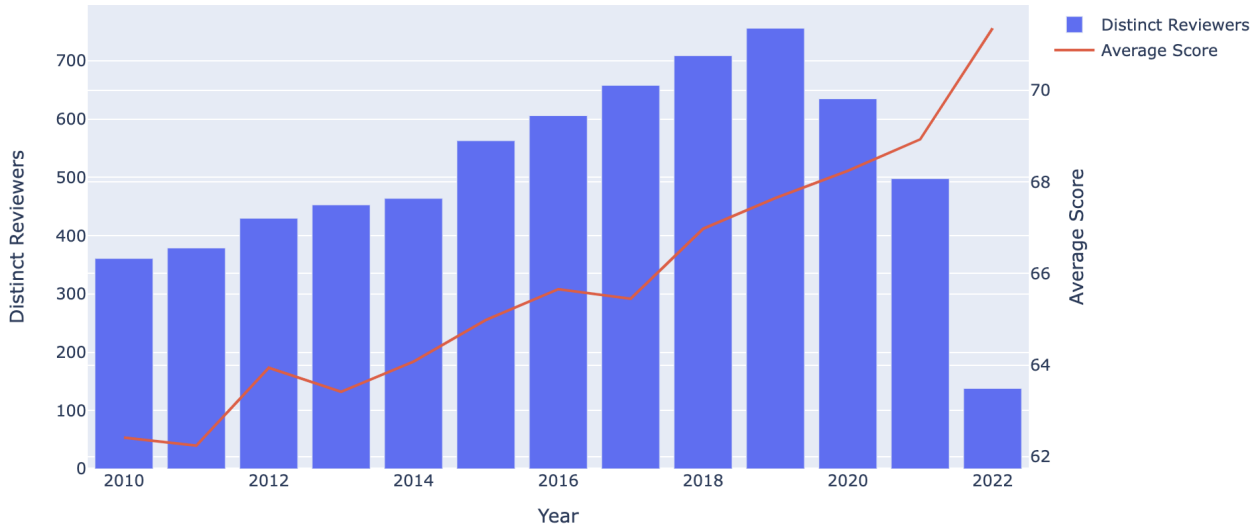


Figure 2.2.1

In our analysis of score trends across years, as illustrated in Figure 2.2.1, several noteworthy patterns emerged. By computing the mean score and the number of unique reviewers for each year, we generated a combined plot to visualize these trends. From 2010 to 2019, the quantity of distinct reviewers experienced a general increase. However, between 2020 and 2022, a decline in the number of reviewers was observed, which can potentially be ascribed to the ramifications of the COVID-19 pandemic on both the film industry and consumer behavior. Intriguingly, the average score followed a dissimilar pattern, as it consistently rose over time and reached its apex in 2022. This observation diverges from the trend of distinct reviewers during the 2020 to 2022 period. The escalation of the average score may be attributed to the emergence of fewer, yet more meticulously curated film releases throughout the pandemic, resulting in high-caliber productions garnering increased attention and favorable evaluations.

Following the EDA, we proceeded with the preprocessing phase, which entailed transforming the review texts to a format suitable for further analysis. We performed the following steps:

1) Converted the text to lowercase.
2) Removed punctuation and special characters using regular expressions.
3) Tokenized the text into individual words.
4) Eliminated stopwords using the NLTK library's stopwords list.
5) Applied stemming using the PorterStemmer from the NLTK library.

After preprocessing the text, we rejoined the tokens into a single string and added the preprocessed and tokenized reviews as new columns in the DataFrame, 'Preprocessed_Review' and 'Tokenized_Review' respectively.

By thoroughly cleaning and preprocessing the data and conducting an EDA, we ensured that our research was grounded in a reliable and representative sample of movie reviews, enabling the extraction of valuable insights and conclusions from our analysis.

3. Methodology

3.1. Topic Modeling with Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA) is an advanced textual analysis technique grounded in computational linguistics research that calculates the statistical correlations among words in a large set of review texts to identify and quantify the underlying topics in these review texts (Hamed Jelodar et al., 2018, p. 12).

3.1.1. LDA Model Tuning

First, we trained the LDA model by using Gensim. Gensim stands for “Generate Similarly”, which allows both LDA model estimation from a training corpus and inference of topic distribution on new, unseen documents. The model can also be updated with new documents for online training (Radim, 2022). The parameters include corpus, id2word, num_topics, random_state, update_every, chunk size, passes, alpha, and per_word_topics.

Then, we computed the coherence scores for each LDA model with a different number of topics. Coherence scores are used to evaluate the quality of topics generated by the LDA model. They measure how semantically similar the top words are in a topic area.

Using the coherence score, we were able to select the best LDA model by choosing the number of topics that yields the highest coherence score.

Figure 3.1.1 illustrates that the number of topics of 3 has the highest coherence score, meaning that the words within a topic are more semantically related to each other.

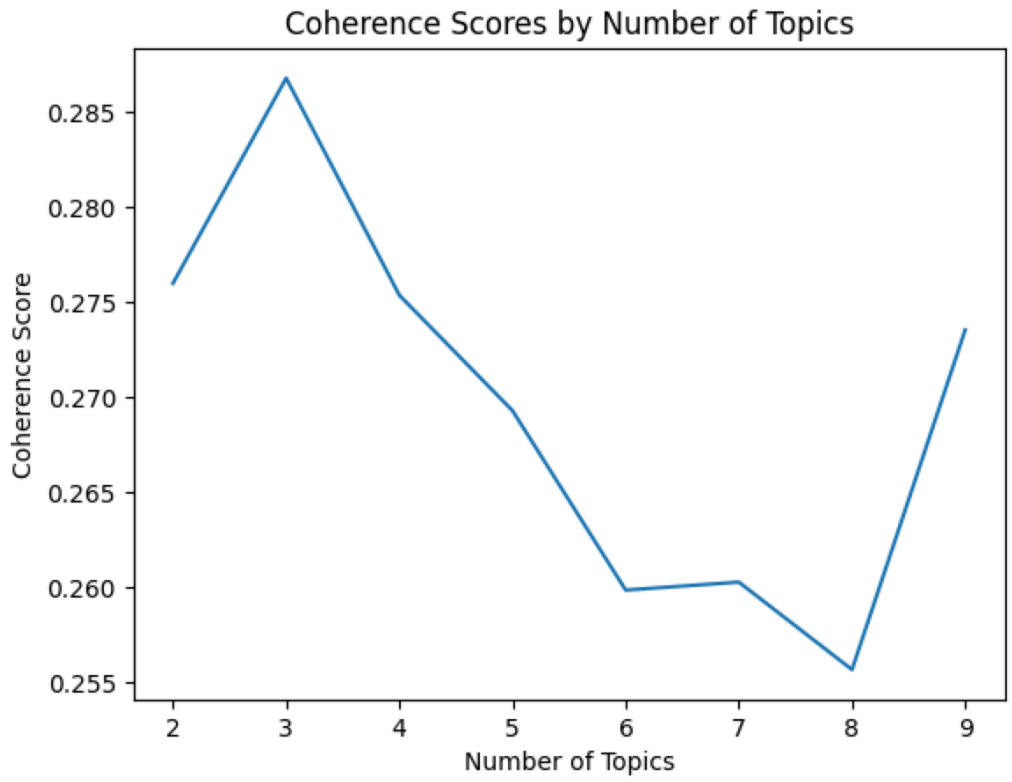


Figure 3.1.1

3.1.2. Word Cloud

First, we counted the frequency of each word for the current dominant topic. Then, for each dominant topic, we created a word cloud in different colors with the calculated word frequencies.

By displaying the most frequent words, word clouds can provide a high-level summary of each topic. It can also help identify similarities and differences between topics, such as changes in the popularity of certain topics over time.

3.1.3. Similarity

We extracted the topic-term matrix from the LDA model which represents the probability distribution of terms for each topic. Then, we computed the pairwise cosine similarity between the topics using cos similarity. This method provides an overview of the similarity between different topics identified by the LDA model, helping to evaluate the distinctiveness of each topic and identify any potential overlap in their content.

3.2. Classification Model

In order to analyze reviewer preferences, we employ a classification model to predict if they will assign a score greater than 50 (1) or not (0) based on the review text. This binary classification problem assists us in comprehending the reviewer's sentiment.

3.2.1. TF-IDF Vectorization

Prior to model construction, it is crucial to converting the text into the numerical format – a process known as vectorization(Singh & Shashi, 2019). Several methods for vectorizing text data exist in text analysis, including Bag of Words (BoW), Word2Vec, and Term Frequency-Inverse Document Frequency (TF-IDF)(Rajaraman & Ullman, 2011). We opted for TF-IDF over other methods for the following reasons:

1. Meaningful representation: TF-IDF considers not only the term frequency within a document but also the term's rarity across all documents. This combination captures the term's importance within the context of the entire dataset, making it more informative than the Bag of Words representation.
2. Superior performance with sparse data: TF-IDF generates a sparse matrix, which is computationally efficient when handling large datasets. It requires less memory and storage, and it reduces noise by assigning more weight to less frequent terms across documents.
3. Ease of implementation: TF-IDF is simpler to implement and demands fewer computational resources compared to more advanced methods like Word2Vec.

TF-IDF enables us to determine the importance of each word in the review by considering its frequency in the review and its rarity across all reviews. After performing TF-IDF vectorization, we obtain total 1,494,218 transformed features. Figure 3.2.1 illustrates the top 20 important features.

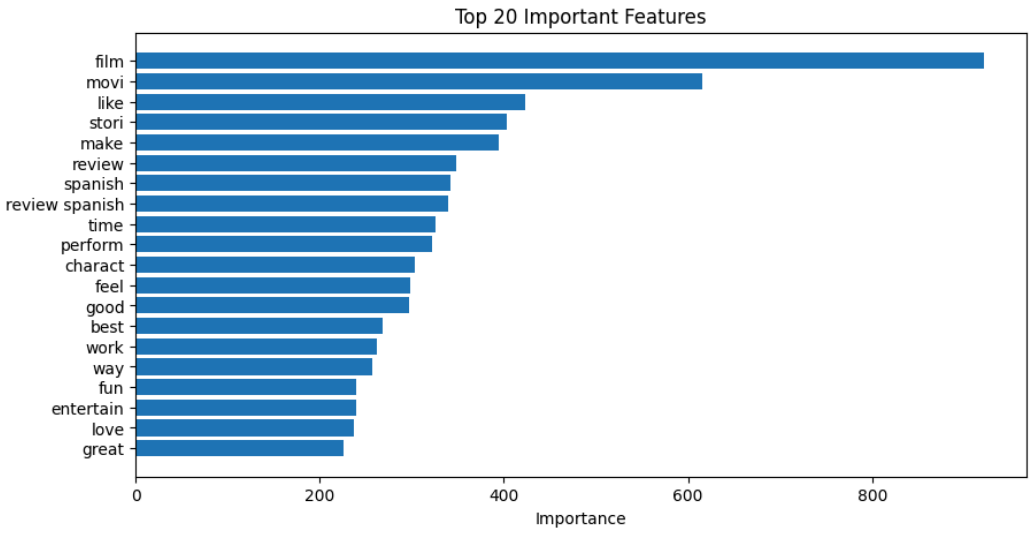


Figure 3.2.1

These features indicate the words with the highest importance across all movie reviews in the dataset, based on their Term Frequency-Inverse Document Frequency (TF-IDF) scores. These scores reflect the significance of a word in a document (review) while accounting for its rarity across all documents (reviews).

In essence, these important features are words that frequently appear in individual reviews but are relatively unique to those reviews. These features can offer valuable insights into prevalent topics or themes in movie reviews and may aid in understanding the most influential factors in reviewers' opinions.

3.2.2. Feature Selection - Chi-Square Test

After obtaining the transformed features, we utilize the chi-square test to select the most relevant features for our classification task. Chi-Square and PCA are two distinct methods used for feature selection and dimensionality reduction, respectively(Rachburee & Punlumjeak, 2015). However, PCA performed poorly on sparse data(Schipper & Van Deun, 2020). Additionally, the

chi-square test helps us identify the most relevant and discriminative words for sentiment classification while maintaining the interpretability of the original features(Nurhayati et al., 2019).

The chi-square test is a statistical method employed to ascertain if a significant association exists between two categorical variables(Nurhayati et al., 2019). In our case, we use it to measure the dependence between the occurrence of features (words) and the target variable (sentiment). We determine that the optimal k value is 5000 using cross-validation. The graph displays the top 30 features with the highest chi-square scores.

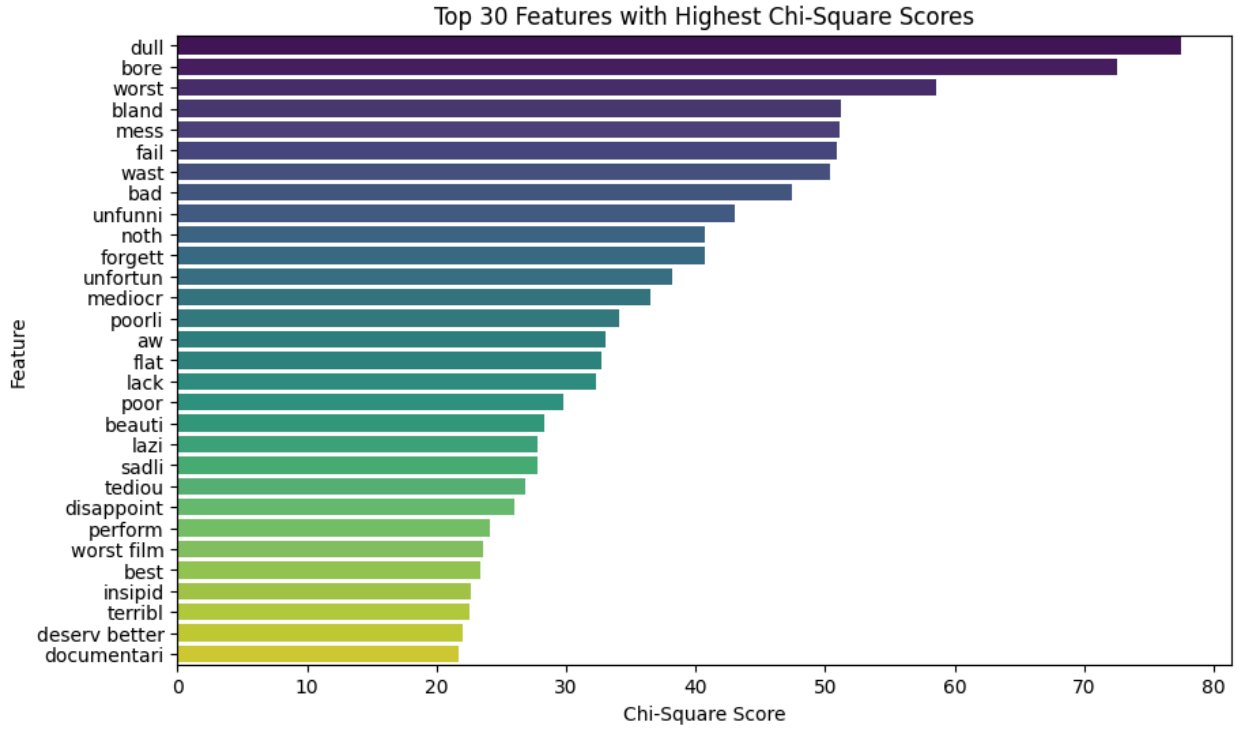


Figure 3.2.2

The top 30 features suggest that these words have the strongest association with the sentiment of the review. In other words, these words can help us better differentiate between positive and negative sentiments in movie reviews. For instance, words like 'dull', 'bore', 'worst', 'bland', 'mess', 'fail', and 'waste' are likely associated with negative sentiment, while words like 'beautiful'

might be associated with positive sentiment. By selecting features with high chi-square scores, we can enhance the performance of our classification models by focusing on the most relevant and discriminative words.

3.2.3. Classification Models

We develop classification models using Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost to perform sentiment analysis.

1. Logistic Regression: Logistic Regression is a linear classification model used for binary and multiclass classification problems. It predicts the probability of an instance belonging to a specific class by using a logistic function. Logistic Regression is chosen for binary classification tasks due to its simplicity, interpretability, and effectiveness. It works well when the relationship between the features and the target variable is approximately linear.
2. Support Vector Machines (SVM): Support Vector Machines (SVM) is a powerful classification algorithm that aims to find the optimal separating hyperplane between different classes. It can handle both linearly separable and non-linearly separable data by using kernel functions (e.g., linear, polynomial, or radial basis function). SVM is chosen for binary classification because of its ability to handle high-dimensional data and its robustness to overfitting, especially when using the kernel trick to map the data into a higher-dimensional space.
3. Random Forest: Random Forest is an ensemble learning method that constructs multiple decision trees and combines their results to make a final prediction. The ensemble approach reduces the risk of overfitting and increases the model's generalizability. Random Forest is chosen for binary classification due to its ability to handle complex

data, its robustness to noise, and its resistance to overfitting. Additionally, it can handle a large number of features and provide insights into feature importance, which can be useful for understanding the most influential factors in classification tasks.

4. XGBoost: XGBoost (eXtreme Gradient Boosting) is an advanced implementation of the gradient boosting algorithm. It's an ensemble method that uses a sequence of weak learners (typically decision trees) to make predictions, where each subsequent learner corrects the errors made by the previous learners. XGBoost is chosen for binary classification because of its speed, scalability, and accuracy. It is designed to handle large datasets efficiently and has built-in regularization and parallelization features that make it less prone to overfitting and faster to train.

In summary, Logistic Regression, Support Vector Machines, Random Forest, and XGBoost are chosen for binary classification tasks due to their unique strengths and capabilities. These models can handle different types of data, scale well with the number of features, and provide robust performance, making them suitable for a wide range of binary classification problems, including sentiment analysis in movie reviews.

4. Results

4.1. LDA

4.1.1. Visualize the topics

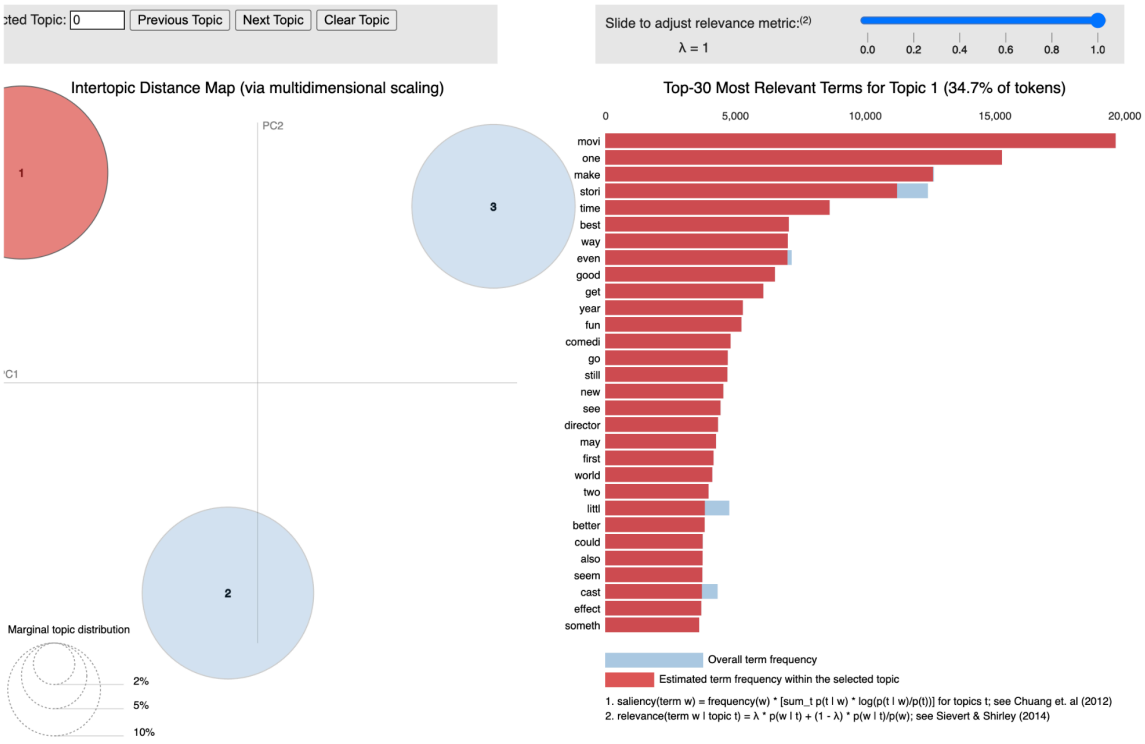
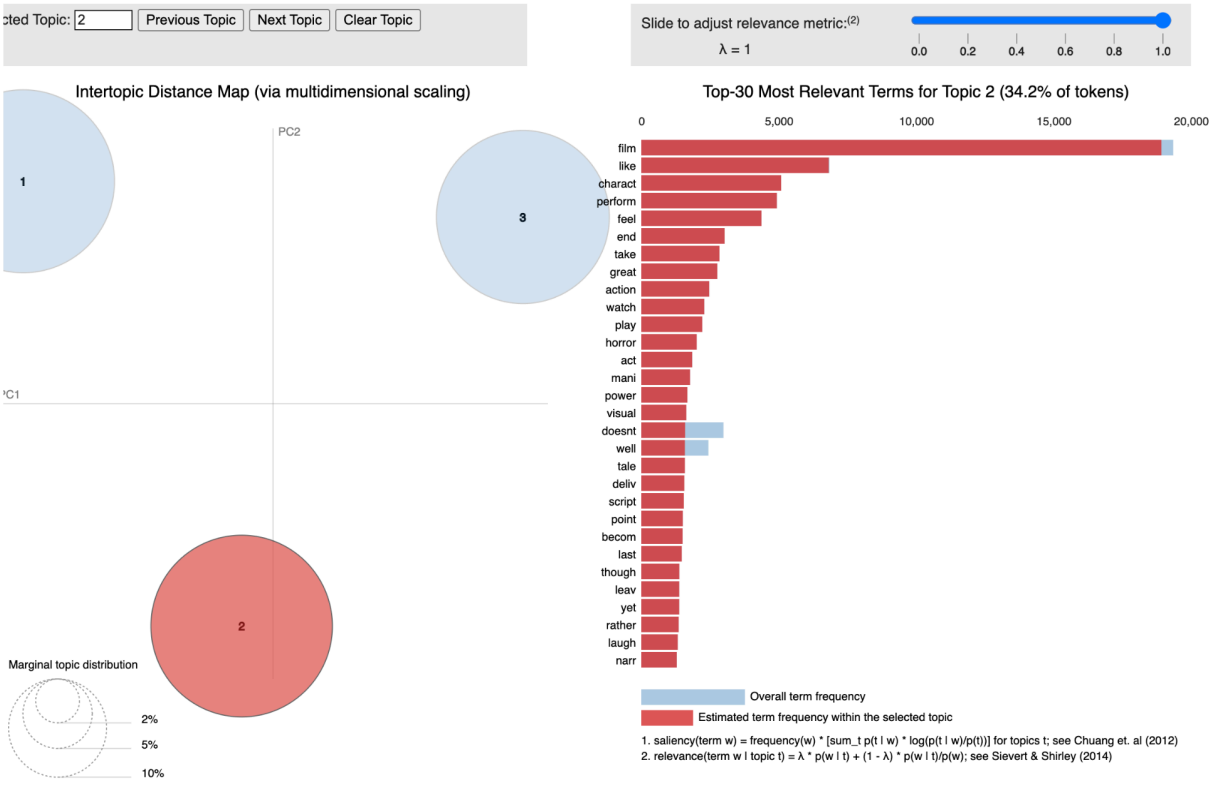


Figure 4.1.1

In Figure 4.1.1, we can see the dominant words in Topic 1, which are the most popular topics when people are reviewing movies. There are 34.7% of tokens which suggests that the topic is well-defined and contains coherent and meaningful keywords. The dominant words are movi, one, make, stori, time, best, way, even, good, get, year, fun, comedi, go, still. “Best”, “good”, “fun”, and “comedy” illustrate that reviews are generally positive and that the movies being reviewed are entertaining and enjoyable. The word "story" suggests that the quality of the plot is an important factor in the reviews, while the words "time" and "year" suggest that the reviews may also be commenting on the timeliness or relevance of the movies.


cted Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1.0
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
1
2
3
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (34.2% of tokens)
0
5,000
10,000
15,000
20,000
film
like
charact
perform
feel
end
take
great
action
watch
play
horror
act
mani
power
visual
doesnt
well
tale
deliv
script
point
becom
last
though
leav
yet
rather
laugh
narr
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

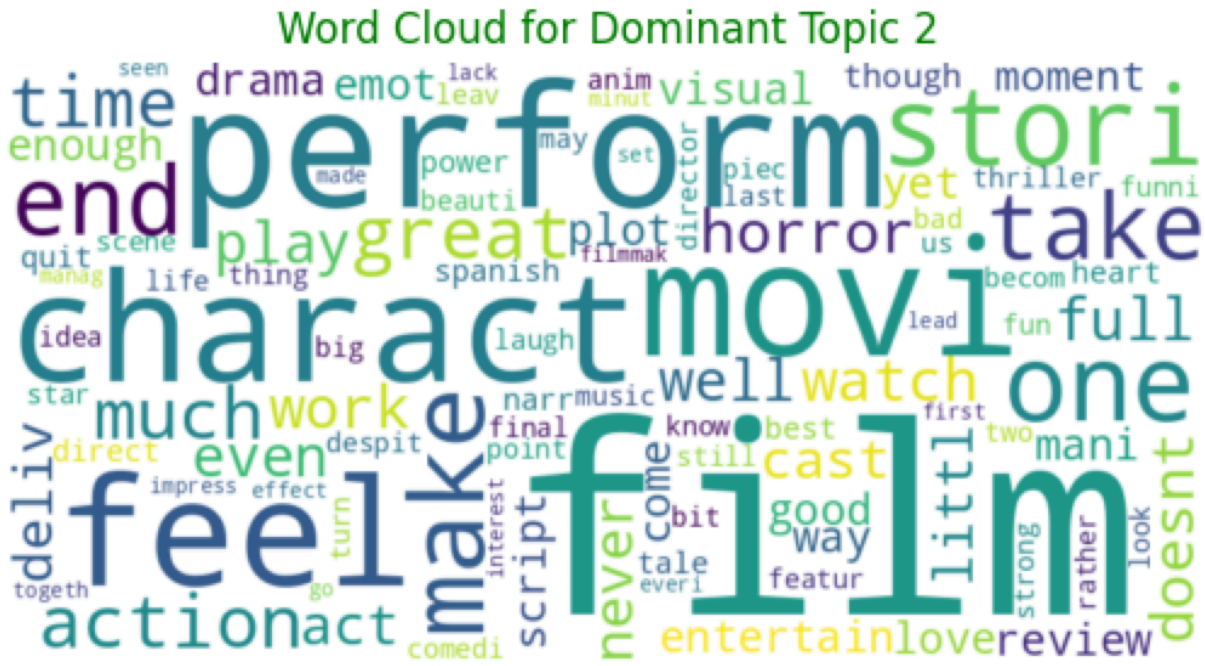

Word Cloud for Dominant Topic 2
seen
drama
emot
lack
leav
anim
minut
visual
though
moment
time
enough
perform
stori
may
set
director
power
made
beauti
piec
last
yet
thriller
funni
end
play
great
plot
horror
bad
us
take
quit
scene
manag
life
thing
filmmak
spanish
becom
heart
charact
movi
idea
big
laugh
lead
fun
full
star
much
work
narr
music
well
watch
one
deliv
direct
even
despit
final
point
know
best
first
two
mani
impress
effect
turn
come
still
cast
film
feel
make
interest
script
never
bit
good
way
littl
doesnt
togeth
go
tale
everi
featur
strong
rather
look
action
act
comedi
entertain
love
review

Figure 4.1.2.

In Figure 4.1.2, Topic 2 is the second most popular review which accounts for 34.2% of the total number of tokens in the set of reviews. The dominant words are film, like, charact, perform, feel, end, take, great, action, watch, play, horror, act, mani, and power. Based on these keywords, we can infer that the reviews on this topic may be discussing films and focusing on aspects such as character development, acting performances, and the overall feel or tone of the film. The use of words such as "end," "take," and "watch" suggests that the reviews may also be commenting on the narrative structure and pacing of the films, while the presence of words such as "action," "horror," and "power" suggests that there may be a focus on films with exciting or intense elements.

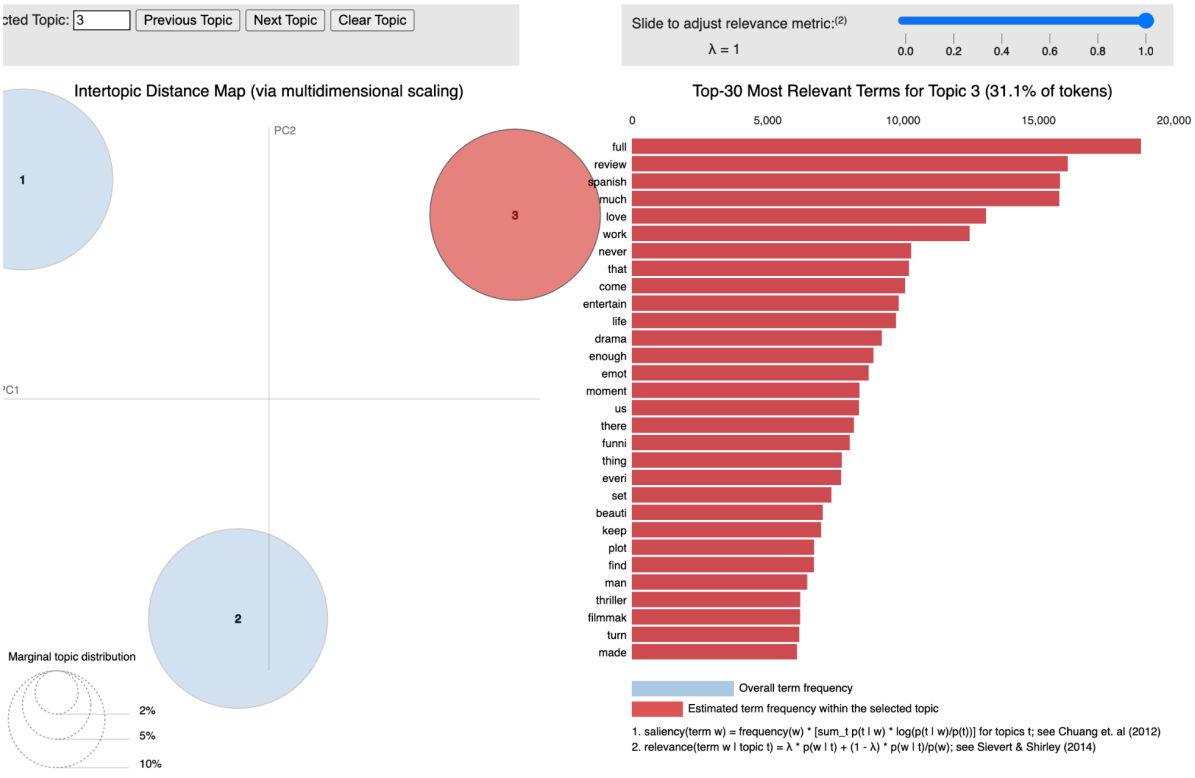


Figure 4.1.3

In Figure 4.1.3, Topic 3 has the dominant words such as full, review, Spanish, much, love, work, never, that, come, entertain, life, drama, enough, emot, and moment. It seems to be related to reviews of dramatic and emotional movies or shows, possibly in Spanish. “Gender”, “man”, and “human” are topics related to education and history, which could also be worth talking about.

4.1.2. Evaluate the topic distribution

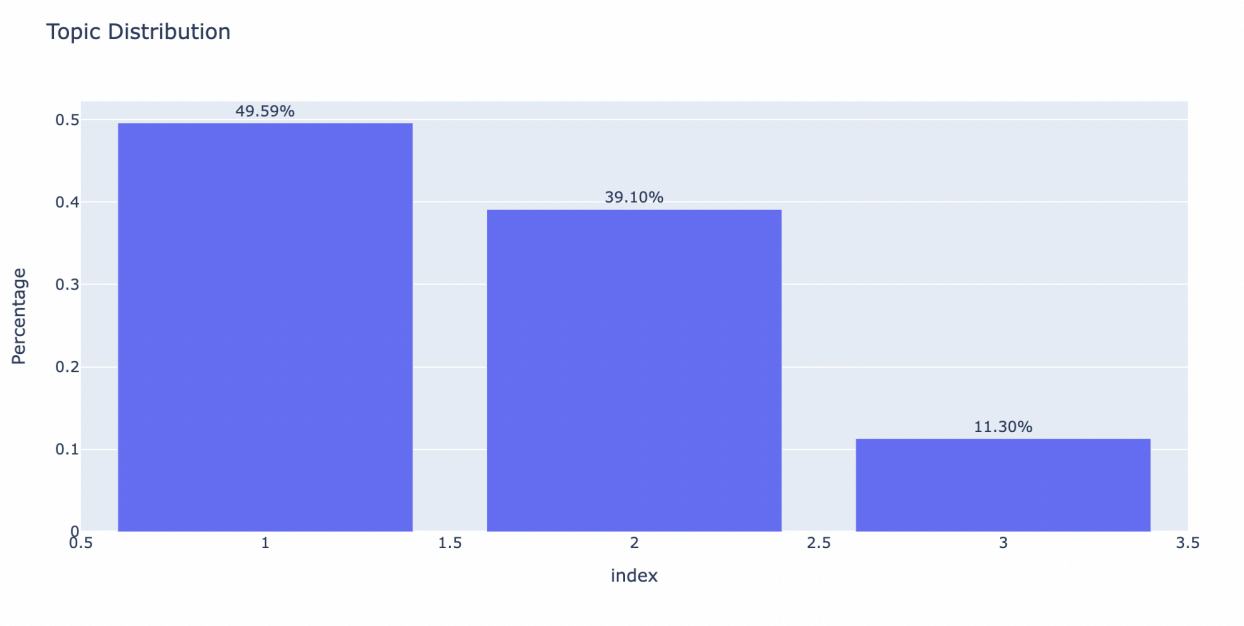


Figure 4.1.4

Figure 4.1.4 demonstrates that there are 49.59% of Topic 1 in the total review text and 39.10% of Topic 2 is the second proportion. Topic 3 has the least amount of distribution, 11.30%. People focus on general movie attributes, such as storylines, quality, and the experience of watching the movie for Topic 1.

4.1.3. The percentage of topics across the year

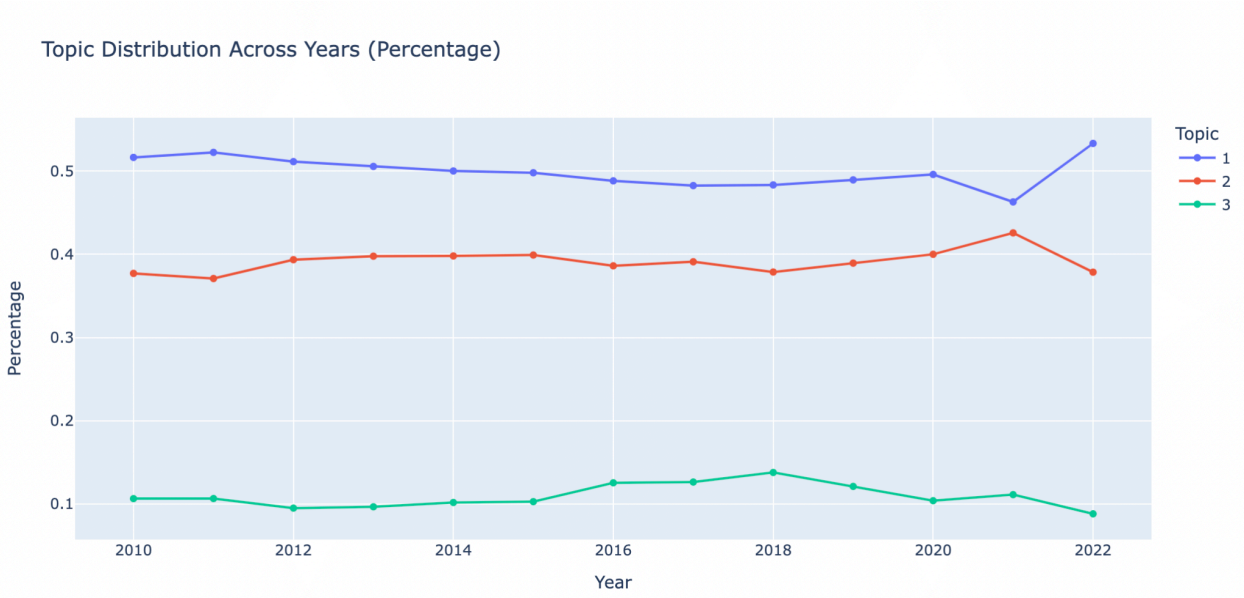


Figure 4.1.5

In Figure 4.1.5, Topic 1 decreases from 2020 to 2021 and increases dramatically, while Topic 2 increases in 2020 and then decreases. The fluctuation in the percentages of Topic 1 and Topic 2 from 2020 to 2022 can be attributed to the changing preferences of audiences and the types of movies being released during these years.

As Figures 4.1.6-1, 4.1.6-2, and 4.1.6-3 (The Numbers, 2023) demonstrate, the market share for each genre varies across the three years. The number of horror movies released increased from 38 in 2020 to 49 in 2021, leading to a rise in market share from 10.07% to 12.82%. This suggests that in 2021, audiences might have been more interested in discussing horror movies, contributing to the increase in the percentage of Topic 2, which focuses on aspects like character development, acting performances, and overall tone, which are particularly relevant to the horror genre.

In contrast, in 2022, the number of horror movies released decreased to 39, with a corresponding decline in market share to 8.50%. During the same year, the number of comedy movies increased dramatically to 58, securing the third rank in terms of genre popularity. This surge in comedy movies, which might have had high-quality stories and characters, likely contributed to an increase in positive comments on movie review platforms, thereby raising the percentage of Topic 1.

Topic 3 consistently maintains a low percentage, which could be due to the fact that educational and documentary movies are released less frequently over the years. These types of movies may not generate as much discussion or interest on movie review platforms as other genres, resulting in a lower percentage for Topic 3.

Top Grossing of 2020

Rank	Genre	Movies	2020 Gross	Tickets	Share
1	Adventure	32	$702,533,509	76,220,828	35.36%
2	Action	28	$429,641,958	46,309,291	21.63%
3	Thriller/Suspense	49	$263,542,875	28,074,674	13.27%
4	Drama	125	$240,480,417	25,696,293	12.11%
5	Horror	38	$200,008,784	21,323,579	10.07%
6	Comedy	42	$75,449,405	8,110,455	3.80%
7	Black Comedy	8	$43,294,447	4,601,907	2.18%
8	Musical	6	$10,965,822	1,174,333	0.55%
9	Romantic Comedy	10	$7,250,057	770,460	0.36%
10	Documentary	53	$4,033,428	428,647	0.20%
11	Western	1	$3,717,170	395,023	0.19%
12	Multiple Genres	2	$3,432,023	364,720	0.17%
13	Concert/Performance	4	$2,056,133	218,502	0.10%
14	Educational	1	$260	27	0.00%

Figure 4.1.6-1 (The Numbers, 2023)

Top Grossing of 2021

Rank	Genre	Movies	2021 Gross	Tickets	Share
1	Action	45	$2,299,333,931	229,125,434	50.77%
2	Adventure	27	$791,336,192	80,488,825	17.47%
3	Horror	49	$580,589,442	55,826,139	12.82%
4	Comedy	59	$360,877,351	35,967,288	7.97%
5	Drama	122	$235,363,249	22,644,004	5.20%
6	Thriller/Suspense	40	$140,857,154	13,543,944	3.11%
7	Musical	4	$73,147,684	7,033,431	1.62%
8	Documentary	48	$16,273,466	1,564,738	0.36%
9	Concert/Performance	3	$14,742,812	1,417,576	0.33%
10	Western	4	$9,161,627	880,925	0.20%
11	Black Comedy	1	$5,252,785	505,075	0.12%
12	Romantic Comedy	7	$1,358,954	133,258	0.03%
13	Multiple Genres	2	$455,295	43,777	0.01%

Figure 4.1.6-2 (The Numbers, 2023)

Top Grossing of 2022

Rank	Genre	Movies	2022 Gross	Tickets	Share
1	Action	54	$3,983,474,992	383,805,697	53.35%
2	Adventure	27	$977,868,029	97,111,359	13.10%
3	Comedy	58	$690,118,052	71,440,206	9.24%
4	Horror	39	$634,295,308	60,861,854	8.50%
5	Drama	146	$619,601,447	62,461,833	8.30%
6	Thriller/Suspense	46	$249,909,212	23,914,977	3.35%
7	Romantic Comedy	11	$105,384,522	10,084,639	1.41%
8	Musical	6	$61,505,843	5,885,723	0.82%
9	Reality	1	$57,743,451	5,525,689	0.77%
10	Black Comedy	3	$44,614,220	4,269,303	0.60%
11	Documentary	78	$30,807,682	2,948,070	0.41%
12	Concert/Performance	5	$9,009,712	862,171	0.12%
13	Multiple Genres	3	$1,849,090	176,945	0.02%
14	Western	1	$804	76	0.00%

Figure 4.1.6-3 (The Numbers, 2023)

4.1.4. Similarity

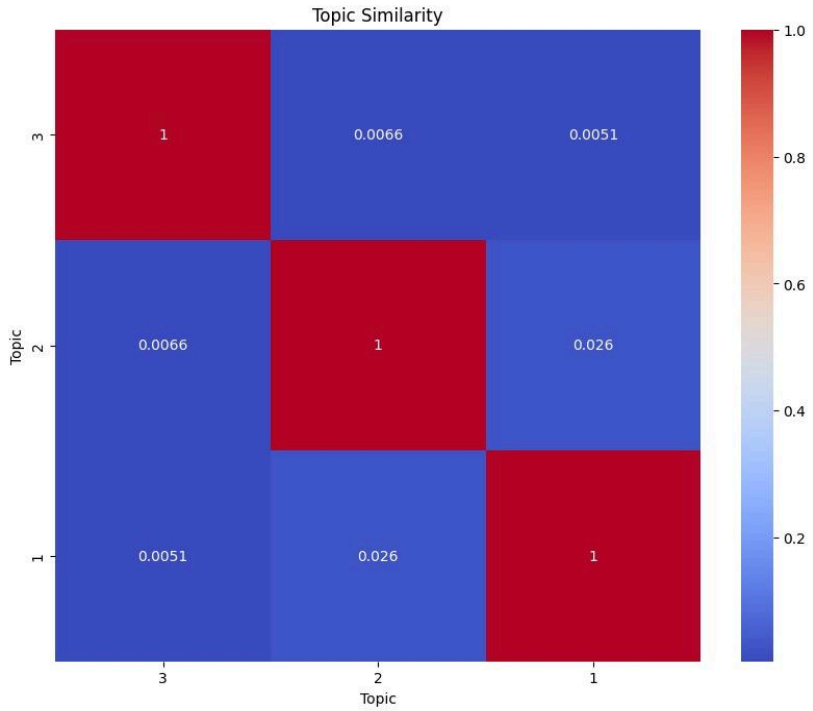


Figure 4.1.7

Topic 1 and Topic 2 have the highest similarity of 0.026 because they are both related to the story and have positive comments on the movie. Conversely, the similarity between Topic 1 and Topic 3 is the lowest at 0.0051. This can be attributed to the fact that Topic 3 focuses more on dramatic and emotional movies or shows, possibly in Spanish, while Topic 1 covers a broader range of movies and has a more positive and lighthearted tone. Topic 3 has distinct features compared to the other two topics.

4.1.5. Limitation

Coherence score: While coherence scores can provide a useful indication of topic quality, they may not always align with meaningful words. For example, “deliv” , “come”, and “right” are not related to analyzing movie reviews.

4.2. Classification

4.2.1. Evaluation of Models and Metrics

In order to assess the performance of the models, we employed the Receiver Operating Characteristic (ROC) curve and metrics obtained from the confusion matrix. The ROC curve serves as a visual representation of the model's true positive rate (sensitivity) against the false positive rate (1-specificity) across different classification thresholds (Brown & Davis, 2006). A higher Area Under the Curve (AUC-ROC) signifies enhanced discrimination between positive and negative classes, indicating the model's ability to effectively differentiate between various sentiment categories.

Furthermore, the confusion matrix metrics, including accuracy, precision, recall, and F1-score, offer a comprehensive understanding of the model's performance. Accuracy represents the ratio of correct predictions out of all instances. In contrast, precision and recall emphasize the model's capability to accurately identify positive instances. The F1-score constitutes the harmonic mean of precision and recall, supplying a singular metric that balances the trade-off between these two measurements. Figure 4.2.1-1 shows how we get each metric.

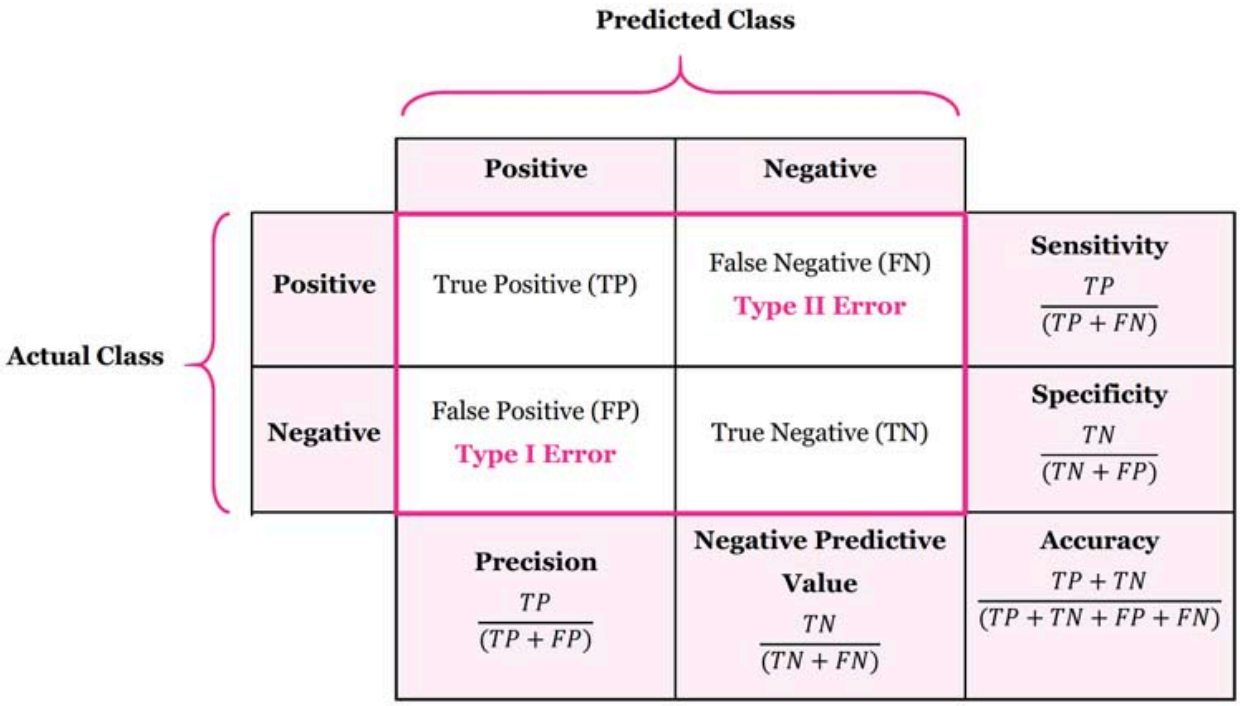


Figure 4.2.1-1

4.2.2. Results and Model Selection

Table 4.2.2-1 displays the evaluation metrics for model performance derived from the confusion matrix. The figure 4.2.2-1 shows the Receiver Operating Characteristic (ROC) Curve.

index	Logistic Regression	SVM	Random Forest	XGBoost
Accuracy	0.7263	0.7638	0.7521	**0.7661**
Precision	**0.893**	0.8921	0.8056	0.7974

index	Logistic Regression	SVM	Random Forest	XGBoost
Recall	0.7172	0.7757	0.8779	**0.9181**
F1-score	0.7955	0.8298	0.8402	**0.8535**

Table 4.2.2-1

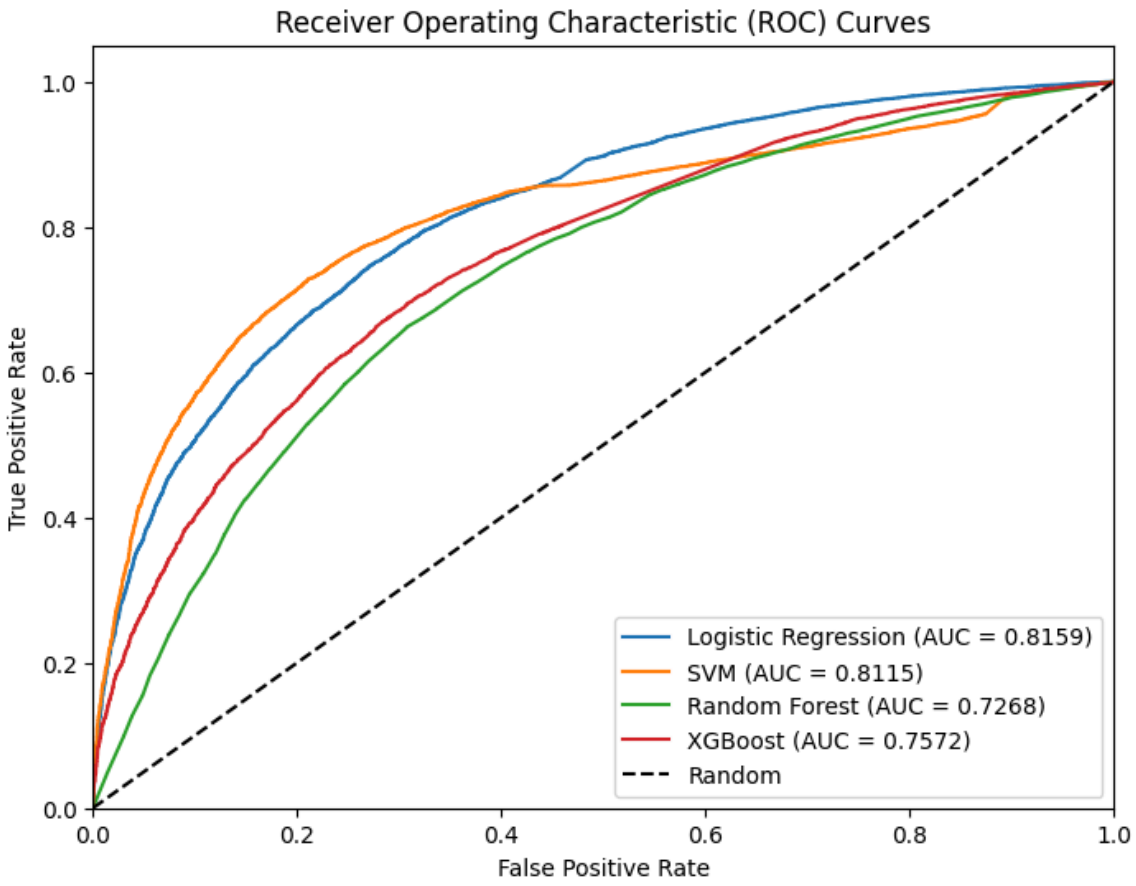


Figure 4.2.2-1

According to the findings, the XGBoost model exhibits the highest accuracy (0.7661) and F1-score (0.8535) (Citation Needed). Despite having a marginally lower AUC-ROC (0.7572) compared to Logistic Regression (0.8159) and Support Vector Machine (SVM) (0.8115), XGBoost attains the highest recall (0.9181) (Citation Needed). In the context of sentiment

analysis, recall is vital for minimizing false negatives, ensuring that the model effectively identifies positive sentiments while preventing misinterpretations of customer feedback.

In conclusion, we advocate for the XGBoost model in our sentiment analysis task due to its superior accuracy, F1-score, and recall (Citation Needed). By excelling in these crucial metrics, the XGBoost model offers a dependable and efficient solution for classifying sentiments in text data, facilitating improved decision-making and targeted actions based on sentiment analysis outcomes.

5. Conclusion

5.1. Movie Review Analysis Methodology & Results Summary

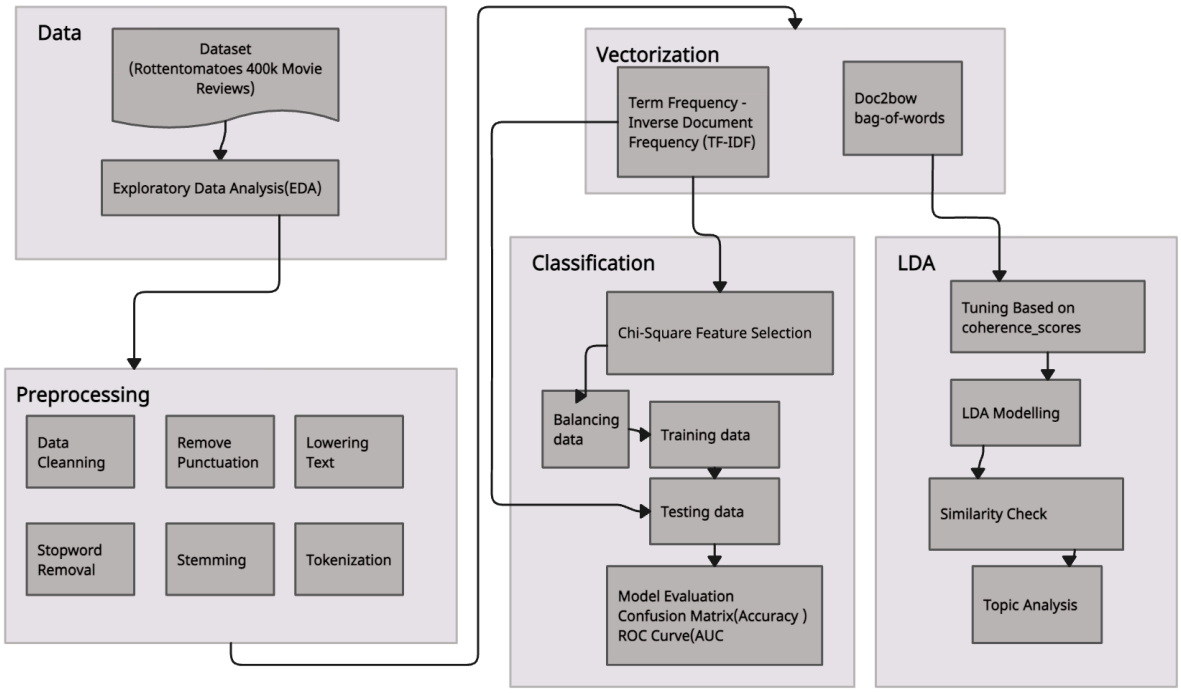


Figure 5.1

Our project aimed to improve the user experience on movie review sites and gain insightful information by analyzing a dataset of 400k movie reviews from Rotten Tomatoes using machine learning and natural language processing techniques. We hypothesized that these techniques could provide valuable insights into the temporal dynamics of online movie reviews. To test our hypothesis, we developed a series of analysis methods and applied them to the Rotten Tomatoes data, as Figure 5.1 shows.

We began by conducting exploratory data analysis (EDA) to gain a better understanding of the distribution of review scores, review behavior, and the content of the review text. The EDA helped us identify key trends, such as the recent increase in average scores, which prompted us to

investigate whether the quality of films has improved or if moviegoers have become more tolerant. We also noted a decrease in the number of reviewers, which led us to question the reason behind this turnaround and why people are less inclined to voice their opinions on movie review sites. Overall, the EDA allowed us to identify trends, patterns, and outliers that informed our later analysis steps.

Before employing any analysis methods, it is imperative to preprocess raw text data to render it suitable for use in algorithms. Preprocessing involves purifying the data, eliminating punctuation, converting the text to lowercase, removing stopwords, stemming, and tokenization. This preliminary step enhances the accuracy of subsequent analyses by ensuring that the data is consistent, unblemished, and structured. By applying these preprocessing techniques to the raw text data, we can make it more amenable to more precise and insightful analysis. After preprocessing, the raw text data must be transformed into numerical data using vectorization to enable its analysis through algorithms. Vectorization involves transforming text data into a numerical representation, and various methods such as Bag of Words (BoW), Word2Vec, and Term Frequency-Inverse Document Frequency (TF-IDF) can be used for this purpose. We opted for TF-IDF to vectorize the text data since it provides a meaningful representation, superior performance with sparse data, and ease of implementation. By converting the text data into numerical data, we can leverage machine learning algorithms to uncover patterns, trends, and insights that would not be apparent from analyzing the raw text alone.

Following preprocessing and vectorization, we employed Latent Dirichlet Allocation (LDA) and machine learning to analyze movie reviews on Rotten Tomatoes. LDA calculated statistical

correlations among words in a large set of review texts to identify and quantify underlying topics. Gensim was used to train the LDA model, and coherence scores were used to evaluate the quality of the topics generated. Word clouds were created for each dominant topic to summarize and identify similarities and differences between them. Pairwise cosine similarity was computed to evaluate the distinctiveness of each topic and identify potential content overlap. Through this approach, we found that there are three distinct movie topics mostly discussed on Rotten Tomatoes. Reviewers there prefer to post positive reviews, focusing on movies' story and quality.

Additionally, we used machine learning techniques to predict the sentiment of reviews. We employed a classification model to analyze reviewer preferences by predicting if they will assign a score greater than 50 based on the review text. To transform the text into a numerical format, we used TF-IDF vectorization, which considers the term frequency within a document and across all documents. The chi-square test was then used to select the most relevant features for the classification task, identifying the most discriminative words for sentiment classification while maintaining the interpretability of the original features. Four classification models were developed, including Logistic Regression, Support Vector Machines (SVM), Random Forest, and XGBoost, with each model having its strengths, including simplicity, interpretability, and effectiveness. XGBoost was found to be the best prediction model for movie reviews on Rotten Tomatoes. Our results demonstrate the potential of combining LDA and machine learning techniques to gain a deeper understanding of the underlying topics in large sets of review texts, ultimately leading to more accurate insights and conclusions.

5.2. Movie Recommendation Systems

Figure 5.2

In addition to analyzing movie reviews for insights into audience preferences and sentiment, we also aimed to build a recommendation system for movie audiences. This system is based on the reviews of similar reviewers and is designed to help users better understand their tastes. To accomplish this, we filtered the dataset based on a condition and calculated the TF-IDF values for each review. We then associated the TF-IDF vectors with the corresponding reviewer names and added dominant topic information to the TF-IDF DataFrame. Next, we concatenated the dominant topic information with the TF-IDF matrix and calculated the average combined feature vectors for all reviews by the same reviewer. The code then used cosine similarity to calculate the similarity between reviewers and returned a similarity matrix. Using this similarity matrix, the code identified the most similar reviewers for a given reviewer and recommended the top 5

movies based on the average dominant topic value of each movie reviewed by the similar reviewers. Finally, the recommended movies were printed to the console as Figure 5.2 shows. This recommendation system provides a valuable tool for movie audiences to discover new films that align with their preferences and interests.

5.3. Business Applications of Review Analysis and Recommendation Systems

Our project focused on analyzing movie reviews on a specific platform, using Rotten Tomatoes as an example. We developed a range of analysis methods to help platforms gain insights into their audience's review habits and preferences, as well as to improve the accuracy of their recommendation systems and enhance user experience.

The first method we used was exploratory data analysis (EDA) to provide an overview of the reviews. This allowed us to identify patterns in the review data, such as common themes, keywords, and sentiment, as well as to gain insights into how users interact with the platform and the review environment it provides. The second method we employed was topic clustering via Latent Dirichlet Allocation (LDA). This method helped us to identify the types of movies that elicit the most discussion on the platform and how this changes over time. By clustering reviews into topics, we were able to gain a deeper understanding of the preferences and opinions of the platform's users, as well as to identify emerging trends in movie reviews.

The third method we used was classification models to predict movie review emotions. By analyzing the emotional content of reviews, we were able to gain a more nuanced understanding of movies and their impact on users. This approach enabled us to extract review emotions beyond the simple binary score ratings, allowing for a more comprehensive understanding of the user experience. Finally, we developed a recommendation system based on cosine similarity on reviews. This method allowed us to calculate the similarity between reviewers and recommend movies based on the average dominant topic value of each movie reviewed by the similar reviewers. By using this approach, we were able to improve the accuracy of recommendation systems and enhance user experience.

Our project provides valuable insights into movie reviews on a specific platform and showcases the benefits of using a range of analysis methods to gain a comprehensive understanding of user behavior, preferences, and emotions. By employing these methods, we were able to identify patterns and trends in movie reviews, which can be used to improve recommendation systems and enhance user experience. Moreover, our analysis methods provide insights into various aspects of the movie industry, such as audience behaviors, movie production, the impact of historical events, and the accuracy of different classification models. This analysis structure can be applied to other movie discussion platforms, providing an overall report on movie reviews similar to what we have achieved with Rotten Tomatoes. Our project thus demonstrates the value of data-driven approaches in understanding user behavior and improving the overall user experience on movie discussion platforms.

5.4. Limitation and Future Directions

Our project has demonstrated the value of using a range of analysis methods to gain insights into movie reviews on a specific platform. By applying exploratory data analysis (EDA), topic clustering via Latent Dirichlet Allocation (LDA), classification models for predicting review emotions, and a recommendation system based on cosine similarity on reviews, we were able to identify patterns and trends in user behavior, preferences, and emotions, as well as to improve the accuracy of recommendation systems and enhance user experience.

However, our project also has some limitations. Firstly, we focused solely on Rotten Tomatoes, which may not be representative of all movie discussion platforms. Secondly, our analysis methods are limited by the availability and quality of the data (Evans, 2006). For example, sentiment analysis may not always accurately capture the nuances of review emotions (Medhat & Korashy, 2014), and recommendation systems may be affected by bias in the review data (Koh & Clemons, 2010).

To address these limitations, future research could expand the scope of analysis to include other movie discussion platforms and incorporate additional data sources such as social media and user demographics. Additionally, incorporating more advanced natural language processing techniques could improve the accuracy of sentiment analysis and enable more nuanced analysis of review content (Hirschberg & Manning, 2015). Finally, future work could explore the potential of incorporating machine learning techniques such as deep learning and reinforcement learning to further improve recommendation systems and enhance user experience (Li, 2017).

5.5. Conclusion

Our project had the objective of enhancing the user experience on movie review sites while extracting meaningful insights through the analysis of text data from prominent movie review platforms. Specifically, we leveraged machine learning and natural language processing techniques to analyze a large dataset comprising 400,000 movie reviews from Rotten Tomatoes. Exploratory data analysis was conducted to identify significant trends, patterns, and outliers that informed subsequent analytical steps. Through the application of preprocessing techniques to the raw text data and its conversion into numerical data using vectorization, we employed LDA and machine learning techniques to analyze the movie reviews on Rotten Tomatoes.

Moreover, the project aimed to construct a recommendation system for movie audiences based on reviews of comparable reviewers to better comprehend their preferences. While the project has certain limitations like its sole focus on Rotten Tomatoes and the potential impact of data quality on analysis methods, it yields valuable insights and tools for movie review platforms to understand their audience's review habits and preferences, enhance the accuracy of their recommendation systems, and improve user experience.

References

Brown, C. D., & Davis, H. T. (2006). Receiver operating characteristics curves and related decision measures: A tutorial. Chemometrics and Intelligent Laboratory Systems, 80(1), 24-38. https://doi.org/10.1016/j.chemolab.2005.05.004

Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2018). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. arXiv:1711.04305v2 [cs.IR]. https://arxiv.org/abs/1711.04305

Radim, Ř. (2022, December 21). Models.Ldamodel – Latent Dirichlet Allocation. Radimrehurek. https://radimrehurek.com/gensim/models/ldamodel.html

Nash Information Services, LLC. (2023). Market Share for Each Genre in 2020. The Numbers. https://www.the-numbers.com/market/2020/genres

Nurhayati, et al. (2019). Chi-square feature selection effect on naive Bayes classifier algorithm performance for sentiment analysis document. 2019 7th International Conference on Cyber and IT Service Management (CITSM). https://doi.org/10.1109/citsm47753.2019.8965332

Singh, A. K., & Shashi, M. (2019). Vectorization of text documents for identifying unifiable news articles. International Journal of Advanced Computer Science and Applications, 10(7). https://doi.org/10.14569/ijacsa.2019.0100742

Rajaraman, A., & Ullman, J. D. (2011). Data Mining. In Mining of Massive Datasets (pp. 1-17). doi:10.1017/CBO9781139058452.002. ISBN 978-1-139-05845-2.

Rachburee, N., & Punlumjeak, W. (2015). A comparison of feature selection approach between greedy, ig-ratio, Chi-square, and mRMR in educational mining. 2015 7th International Conference on Information Technology and Electrical Engineering (ICITEE). https://doi.org/10.1109/iciteed.2015.7408983

Schipper, N. C., & Van Deun, K. (2020). Model selection techniques for sparse weight‐based principal component analysis. Journal of Chemometrics, 35(2). https://doi.org/10.1002/cem.3289

Evans, P. (2006). Scaling and assessment of data quality. Acta Crystallographica Section D: Biological Crystallography, 62(1), 72-82.

Koh, N. S., Hu, N., & Clemons, E. K. (2010). Do online reviews reflect a product's true perceived quality? An investigation of online movie reviews across cultures. Electronic Commerce Research and Applications, 9(5), 374-385.

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.

Hirschberg, J., & Manning, C. D. (2015). Advances in natural language processing. Science, 349(6245), 261-266.

Li, Y. (2017). Deep reinforcement learning: An overview. arXiv preprint arXiv:1701.07274.