Environment and Social Psychology (2023) Volume 8 Issue 3
doi: 10.54517/esp.v8i3.1958

RESEARCH ARTICLE

Analyze IMDb movies by sentiment and topic analysis

Ningjing Ouyang

School of Communication, Hong Kong Baptist University, Hongkong 999077, China; 22444173@life.hkbu.edu.hk

ABSTRACT

Movie is an important cultural form, carrying multiple levels and meanings such as art, entertainment and social value. Movie review and rating data sets are huge, and deep learning and natural language processing methods are widely used today. Advances in big data and deep learning offer unprecedented opportunities to understand moviegoer behavior and preferences while providing a cost-effective way to gain insights relevant to the entertainment industry. This project conducts sentiment analysis, topic modeling, and visual statistical analysis based on the IMDb movie data set to identify key factors and deeper insights that influence successful decision-making in film production. This project first uses the word embedding method to vectorize the movie review text, and then uses Bidirectional Long Short-Term Memory (Bi-LSTM) to perform sentiment classification. In addition, statistical methods such as visualization were used to discover conclusions such as the highest average number of movies released in November, and identify trends, patterns and relationships between the variables of IMDb movies. Finally, the Latent Dirichlet Allocation (LDA) topic modeling model was constructed to find out that the important topic with increased demand is light entertainment movies, highlighting the commercial feasibility of comedy movies as a profitable business model. In summary, this project uses an emotion-topic fusion analysis method based on the Bi-LSTM emotion classification method and the LDA topic modeling method. The results show that the Bi-LSTM model can better identify positive and negative emotions in movie reviews, and the LDA topic model performs well in mining popular topics.

Keywords: movie; nature language processing; sentiment analysis; topic analysis; Bi-LSTM; LDA

1. Introduction

1.1. Background and motivation

Watching movies has gradually become an inseparable part of people's daily lives in recent years. As social media networks including Twitter, Facebook, and Instagram develop rapidly, a tremendous amount of user-generated data is available for analysis. Analyzing these data can provide useful insights for movie production companies to develop successful films. Therefore, social media data mining about movies has become increasingly important as it can provide valuable insights into consumer tastes and preferences, identify emerging trends and patterns, and reveal opportunities for market expansion. By analyzing user-generated data, we can gain insights about moviegoers' preferences and sentiments towards different movies, actors, and directors. It is valuable to movie studios and distributors. With millions of online users daily sharing their opinions on movies, social media data mining allows businesses and researchers alike to unprecedentedly comprehend the behaviors and preferences of moviegoers across the globe. Furthermore, social media data

ARTICLE INFO
Received: 31 July 2023 | Accepted: 18 September 2023 | Available online: 25 October 2023

CITATION
Ouyang N. Analyze IMDb movies by sentiment and topic analysis. *Environment and Social Psychology* 2023; 8(3): 1958. doi: 10.54517/esp.v8i3.1958

mining offers a cost-effective method for movie businesses, researchers, and stakeholders to extract meaningful insights related to the entertainment industry in general, thereby, advancing our understanding of this field in interesting and insightful ways.

This project aims to analyze the preference of moviegoers from social media data including basic information, plot summary, reviews and ratings of movies on movie websites to find the key factors that help decision-making for movie production companies. The reviews are analyzed in terms of sentiment while key information about the plot summary of high-rated movies is extracted and visualized by using topic analysis. Study the relationship between movie ratings and sentiment and movie topics and take other factors such as movie genre and duration into account, resulting in suggestions for improving the popularity of the movies.

The basic idea is that user reviews and plot synopsis texts were converted into word vectors and analyzed applying Latent Dirichlet Allocation (LDA) models[1] and Bidirectional Long Short-Term Memory (Bi-LSTM)[2]. Conduct a sentiment analysis of user reviews and discover the topics told in the plot descriptions of high-scoring films. Statistical analysis was used to present the relationship among movie genre, duration, sentiment and movie ratings. Topic analysis was used for figuring out what the main content (plot summary) of the highly rated film was about.

This study seeks to thoroughly investigate the elements that lead to highly rated movies and subsequently improve the production and promotion operations of films. The results provide recommendations for producing high-rating movies based on multidimensional analysis.

This research findings have important implications for the data mining based on text data analysis. Specifically, it is critical in utilizing these analyses to identify the essential factors contributing to high-rated movies. The inclusion of violence, aliens, school, and dreams as plot elements has been prominently favoured by audiences and producers who want to make high-scoring films can take note of these preferences. Niche films with high-quality content are more likely to earn higher scores. Furthermore, there is a positive correlation between the sentiment of the review and the rating of the film. It is important to manage Internet word-of-mouth (IWOM) well. Producers can improve the review sentiment of their films by actively engaging with audiences and responding to their feedback and comments, leading to better ratings. These results provide useful insights for studios to develop high-rated films.

1.2. Research question and objective

This thesis aims to conduct in-depth data analysis of film reviews through text mining methods including sentiment analysis and LDA topic modelling, to mine the tendency judgment of film reviews and to mine and analyse the hidden information. The following sub-questions will help us better address the main problem:

1) Use sentiment analysis methods such as text mining to classify the films of IMDb Movies positively and negatively and identify the emotional tendencies of movie users.

2) Build a topic modelling model to extract topics and find high-rated movie topics.

3) Use statistical methods to determine the relationship between movie genre, duration, emotion, and movie rating.

1.3. Contributions

The method of this thesis based on natural language processing includes sentiment analysis and topic analysis for movie review text mining, and the language used is python language. The major contributions include:

- Use NLTK to perform data processing on the text data of the IMDb movie data set. The clean data set can be used as the basis for any relevant analysis.

- Use the word embedding method of the Keras library to vectorize text data.
- Use the LSTM sequential network of the deep learning method Recurrent Neural Network (RNN) to classify the emotions of movies.
- Use TF-IDF to visualize high-frequency words.
- Build an LDA model for topic discovery.
- Use visual methods to explore patterns of relationships between variables in a data set.

This project proposes an emotion-topic fusion analysis method based on Bi-LSTM and LDA to mine film review public opinion. This fusion analysis method avoids the shortcomings of previous studies in which topic mining and sentiment analysis were studied in isolation.

2. Literature review

Previous studies have been reported to analyse the social media data about movies to help decision making. Topal and Ozsoyoglu[3] suggested employing an emotion map made by clustering movies with ratings and reviews with the application of the k-means clustering technique to assist moviegoers in choosing which film to see. Sharma et al.[4] use natural language processing and a variety of machine learning classifiers. Trivedi et al.[5] set out to offer sentiment analysis of an Indian movie review corpus. To solve the sentiment analysis challenge, Hybrid CNN-LSTM Model combines deep convolutional neural networks (CNN) with LSTM[6]. Researchers have shown a great deal of interest in sentiment analysis and its applications because of the vast amounts of internet data that are available at the same time that social media is expanding[7]. To find out how earlier studies have approached this problem, Hourrane et al.[8] review the state of the art. Using neural networks trained on Stanford's "Movie Review Database" and two large sets of positive and negative phrases, the challenge of extracting opinions from movie reviews has been accomplished[9]. There have been numerous studies aimed at classifying films based on sentiment. In terms of each sentiment, textual data can be classified in multiple ways using a natural language method, according to Arora et al.[10]. The movie review is classified by Chirgaiya et al.[11] into the appropriate category using a classifier model that was developed using feature extraction and feature ranking. One of the biggest obstacles to Turkish natural language processing (NLP) research is a lack of funding. Acikalin et al.[12] suggested two ways for Turkish sentiment analysis to get around these drawbacks which are enhancing BERT's multilingual model and utilizing the BERT primary model following automatic English to Turkish text translation. For upcoming text classification challenges, Wu et al.[13] plan to fine-tune the pre-trained model BERT. Using a variety of classification algorithms, Kaushik et al.[14] examine sentiment analysis based on movie reviews: a review. IMDb's Sentiment Analysis for movie reviews explains the reviewer's overall sentiment or opinion of a film. Examples include any feeling or opinion expressed by people regarding audits, surveys, web journals, smaller-scale websites, and a series of informal organizations.

3. Methodology

The overall flow chart of IMDb film review text mining and data visualization mining is shown in **Figure 1**. The pipeline comprised of five main stages: data collection, data preprocessing, sentiment classification model construction, statistical analysis and visualization, and topic analysis. Following are introductions of each stage.

Figure 1. The flowchart of data mining in this paper.

3.1. Key methods

3.1.1. Sentiment classification method based on Bi-LSTM

A typical method of employing text in neural networks is to represent words as vectors, which is a key concept in NLP. In this study, before training an emotion classification model based on deep learning, use the word-embedding method to vectorize text data. Use the Keras library, which supports the TensorFlow backend and provides a convenient method for text sequence preprocessing. The text corpus is vectorized by the tokenizer into a list of integers. The keys in the dictionary are the words, and each integer corresponds to a value in the dictionary that encodes the complete corpus. The parameter *num_words* sets the size of the words dictionary. It is set to 6000. Text sequences in most cases have different numbers of words and can be filled with zeroes for sequences that are not long enough using *pad_sequence().* After calculating the average length of a comment to be 120 words, in order to remove any words that are longer than 130, it is appropriate to add the maxlen option, which specifies the length of the sequence to be 130.

The sentiment categorization model is trained using a Recurrent Neural Network (RNN), which aims to classify the movie reviews as positive or negative. The input of the model is movie review text, and the output is the sentiment classification result. **Figure 2** shows the architecture of the sentiment classification model which consist of 6 layers. The input layer encodes the movie reviews as a sequence of words with one-hot encoding with a maximum length of 130. Then, the word embedding layer transforms the high-dimensional and one-hot-encoded word vectors of a text corpus into dense, continuous, and low-dimensional representations that capture semantic relationships between words with the embedding size of 128. These embedded vectors are then passed on to Bi-LSTM layer to learn representation for movie reviews for sentiment analysis. Each direction LSTM transforms the embedding with an output vector with size of 32. And Max Pooling layer was employed to compress features and reduce the complexity of the neural network by aggregating the output of Bi-LSTM layer. Finally, a linear binary classification layer with the sigmoid as the activate function was fed the output of the Max Pooling layer. The model's training parameters total 810,537.

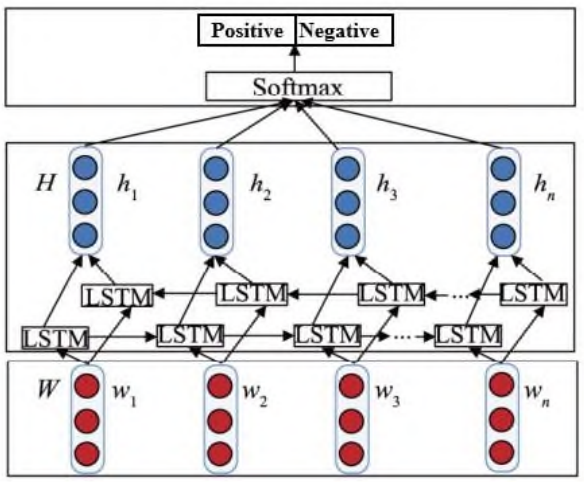

Figure 2. The architecture diagram of the sentiment classification model.

Figure 2 is the emotion classification method based on Bi-LSTM. From bottom to top are the word embedding layer, feature learning layer and emotion classification layer.

Bi-LSTM is composed of forward LSTM and backward LSTM. Learning sentence information from both forward and reverse directions can improve the accuracy of emotion classification. Its calculation formula is:

$$X^L = [h_{t-1}^L, \omega_t] \quad (1)$$

$$i_t^L = sigmoid\left(W_i^L \cdot X^L + b_i^L\right) \quad (2)$$

$$f_t^L = sigmoid\left(W_f^L \cdot X^L + b_f^L\right) \quad (3)$$

$$o_t^L = sigmoid(W_o^L \cdot X^L + b_o^L) \quad (4)$$

$$c_t^L = f_t^L \odot c_{t+1}^L + i_t^L \odot tan\, h(W_c^L \cdot X^L + b_c^L) \quad (5)$$

$$h_t^L = o_t^L \odot tan\, h(h_t^L) \quad (6)$$

$$X^R = [h_{t-1}^R, \omega_t] \quad (7)$$

$$i_t^R = sigmoid\left(W_i^R \cdot X^R + b_i^R\right) \quad (8)$$

$$f_t^R = sigmoid\left(W_f^R \cdot X^R + b_f^R\right) \quad (9)$$

$$o_t^R = sigmoid(W_o^R \cdot X^R + b_o^R) \quad (10)$$

$$c_t^R = f_t^R \odot c_{t+1}^R + i_t^R \odot tan\, h(W_c^R \cdot X^R + b_c^R) \quad (11)$$

$$h_t^R = o_t^R \odot tan\, h(h_t^R) \quad (12)$$

$$h_t = [h_t^L, h_t^R] \quad (13)$$

Among them, Equations (1)–(6) are forward LSTM calculation formulas, and Equations (7)–(12) are backward LSTM calculation formulas. Equation (13) integrates the results of bidirectional LSTM to obtain the final latent vector of each word.

3.1.2. Popular topic mining based on LDA

Topic Modeling is a powerful statistical technique used for discovering the underlying themes and latent patterns within a corpus of textual data. It can automatically capture a set of relevant words as a topic from collections of text documents without labeled data by using a specified number of topics and a corpus of documents as input. Consequently, topic modeling is well-suited for tasks such as document clustering, information retrieval, and organization of massive amounts of unstructured text data. In this study, Topic Modeling was employed to find the key elements for high rating movies. Following are details of topic analysis.

In this study, films with a score above 7 are considered to be high-scoring films. High-scoring Films were screened to obtain thematic analysis.

In the realm of NLP, the term Frequency-Inverse Document Frequency (TF-IDF)[15] is frequently used to weight terms in text documents. With regard to a wider corpus of papers, it seeks to illustrate the significance of each word in a given document. By assigning higher weights to words that are more frequent in a particular document and less frequent across the corpus as a whole, TF-IDF can effectively capture the meaningfulness and relevance of words within documents. In this study, it was used to represent the document of plot summary. Words that appeared in less than 15 comments and words that appeared in more than 50% of the comments were removed, and the words were vectorized using the doc2bow[16] method, constructing bow bags of words. After cleaning the text to be used for analysis, word frequency statistics are done to obtain the number and frequency of occurrences, and TF-IDF values.

Latent Dirichlet Allocation is a popular model for topic modeling and text analysis[17]. It is an unsupervised learning technique that uses word pattern analysis to find the underlying subjects in a huge corpus of writings. According to LDA, each document consists of a variety of themes, each of which is a probability distribution over a set of words. LDA determines the most likely themes of each document and the related word distribution for each topic by estimating the probability distributions of topics and words in the corpus. The main advantage of the LDA model is its ability to extract meaningful themes and patterns from unstructured text data without the need for supervised labels. As a result, LDA has been widely used in many natural language processing applications, including information retrieval, sentiment analysis, and content recommendation systems. This study will use LDA for topic analysis.

The number of topics, the number of words, and the number of iterations are three crucial factors in the LDA model.

- Number of topics: how many topics will be taken out of the corpus. To get better clustering results, choose the appropriate amount of subjects.
- Number of topic words: the words in each topic. This study needs to extract topics for different clusters, so more topic words are chosen.
- Number of iterations: the most iterations necessary for LDA to converge.

To find the ideal number of subjects, the LDA model's performance is assessed. Two common methods of evaluation are used: (1) The test dataset is first labelled and classified as the true result, and then compared to the clustering results using an algorithm such as NMI. (2) Instead of classifying and labelling the test dataset, the trained model is used directly to predict the results. The most commonly used evaluation metrics are perplexity and coherence. Perplexity refers to the uncertainty of a trained model in text analysis in identifying which topics are contained in certain documents. David M. Blei[17] used the Perplexity value as a criterion in LDA experiment. A lower value therefore means less uncertainty and a better final clustering result. The complexity of the model decreases as the number of topics increases, but models with too many topics tend to be overfitted, so cannot only rely on perplexity to judge a model.

To determine whether the words within a topic are coherent, coherence is employed. A group of words is coherent when they are mutually supportive. A conditional probability-based technique was utilized to compute coherence based on the concept of topic coherence after Newman et al.[18] proposed using PMI to calculate topic coherence. Musat's et al.[19] proposal to use the WordNet idea of hierarchy to capture linkages between subjects in the same year. After Baroni[20] then suggested a strategy based on distributional similarity to discover coherence. For now, a few of the more common methods are those collated in Roder et al.[21], and there are packaged functions in Genism that can be called directly, such as C_{uci}, $C_{npmi}$and C_v. This study uses the C_v (Coefficient of variance) method, on the basis of a sliding window, for one-set partitioning of subject terms, and indirectly obtains coherence using normalized point mutual information (NPMI) and cosine similarity.

3.2. Data

3.2.1. Data collection

To investigate the factors influencing movie ratings and sentiment, three datasets were collected from Kaggle[22], which is an open platform for data modelling and data analysis competitions. The first dataset[22] consists of 50,000 movie reviews with sentiment labels (positive or negative), which is used for training the sentiment classification model. **Table 1** shows the example data of the dataset. It was a balanced dataset containing 25,000 reviews each for positive and negative sentiment. The other two datasets include movie reviews (573,913 records) and their basic movie information such as duration, genres (1,572 records) and ratings respectively, shown in **Table 2**.

Table 1. IMDb movie reviews with ratings dataset.

Review	Rating	Sentiment
Kurt Russell's chameleon-like performance, coupled with John Carpenter's flawless filmmaking, makes this one, without a doubt, one of the finest boob-tube bios ever aired. It holds up, too: the emotional foundation is strong enough that it'll never age; Carpenter has preserved for posterity the power and ultimate poignancy of the life of the one and only King of Rock and Roll. (I'd been a borderline Elvis fan most of my life, but it wasn't until I saw this mind-blowingly moving movie that I looked BEYOND the image at the man himself. It was quite a revelation.) ELVIS remains one of the top ten made-for-tv movies of all time.	10	1
It was extremely low budget (it some scenes it looks like they recorded with a home video recorder). However, it does have a good plot line, and its easy to follow. 8 years after shooting her sexually abusive step father Amanda is released from the psychiatric ward, with the help of her doctor who she is secretly having an affair with. The doctor ends up renting her a house and buying her a car. But within the first 20 minutes of the movie Amanda kills him and buries him in her backyard. Then she see's her neighbor Richard sets eyes on him and stops at nothing until she has him. She acts innocent but after another neighbor Buzz finds out that Amanda killed that doctor and attempted to kill Richard's wife Laurie (this is after Amanda and him get it on in the hot tub). Then she stops acting so Innocent and kills Buzz and later on attempts to kill Richard whom she supposedly loves and cares for. And you'll have to rent the movie to find out if Amanda dies or not. Overall good movie, reminds me a lot of my life you know the whole falling for the neighbor and stopping at nothing until you have him part.	8	1
Who the heck had the "bright"(?) idea of casting Lucille Ball in this film??? It should have been Angela Lansbury's baby all the way. At the very least Lucy should have had her singing dubbed. There is some compensation in the fact that Jerry Herman's score is pretty well kept intact except for "That's How Young I Feel", and we do get performances by the original Broadway cast members Jane Connell and Bea Arthur. I suppose Robert Preston had to be given a song, hence the inferior "Loving You". Overall, I think in this one the wrong redhead was cast.	3	0
David Lean's worst film. Even 'In Which We Serve' wasn't as bad as this. Usually a film with a really good reputation like this one, has at least some redeeming qualities, which makes one understand why it might be considered a classic. But after watching this I just could not get why this piece of crap was liked so much even back in 1945! I disliked the acting, stiff upper lip British mannerisms, story, script (which may be quite witty at times but totally unfunny) and soundtrack. The elvira character is meant to be alluring and attractive, but was in actual fact ugly and had a weird and annoying voice. Just another film that has convinced me not to trust a films reputation. Another very overrated 'british classic'.	2	0

Table 2. IMDb movie reviews with ratings dataset.

Movie ID	Plot Summary	Duration	Genre	Rating	Release date
tt0040897	Fred C. Dobbs and Bob Curtin, both down on their luck in Tampico, Mexico in 1925, meet up with a grizzled prospector named Howard and decide to join with him in search of gold in the wilds of central Mexico. Through enormous difficulties, they eventually succeed in finding gold, but bandits, the elements, and most especially greed threaten to turn their success into disaster. Written by Jim Beaver	2h 6min	Adventure, Drama, Western	8.3	1948-01-24
tt0286716	Bruce Banner, a brilliant scientist with a cloudy past about his family, is involved in an accident in his laboratory causing him to become exposed to gamma radiation and Nanomeds (A tiny life-form that is supposed to heal wounds but has killed everything with which they have made contact). Confused and curious about his survival, Banner discovers that since the accident, whenever he becomes angry he transforms into a giant green monster destroying everything in sight in an act of fury. Bruce's mysterious past and the answer to why the radiation had this effect becomes revealed to him as his Birth Father David Banner intervenes with hopes to continue experimenting on him. Written by Séamus Hanly	2h 18min	Action, Sci-Fi	5.7	2003-06-20

Table 2. (*Continued*).

Movie ID	Plot Summary	Duration	Genre	Rating	Release date
tt0243155	Bridget Jones is an average woman struggling against her age, her weight, her job, her lack of a man, and her many imperfections. As a New Year's Resolution, Bridget decides to take control of her life, starting by keeping a diary in which she will always tell the complete truth. The fireworks begin when her charming though disreputable boss takes an interest in the quirky Miss Jones. Thrown into the mix are Bridget's band of slightly eccentric friends and a rather disagreeable acquaintance who Bridget cannot seem to stop running into or help finding quietly attractive. Written by Anuja Varghese	1h 37min	Comedy, Drama, Romance	6.7	2001-04-13

3.2.2. Data preprocessing

Text data cleaning is a critical procedure for further analysis or modeling. The punctuation marks, website links, and stop words in the original text of the movie review are irrelevant information for sentiment analysis and topic analysis. This irrelevant information will waste computing resources and affect the accuracy and effectiveness of NLP models[23]. Removing irrelevant information improves the effectiveness of NLP models. The Natural Language Toolkit (NLTK), which offers numerous text processing programs and test datasets, was used to clean the text data in order to achieve this goal. The following steps are carried out:

- Remove the html tag and punctuations.
- Convert all letters to lower case.
- Identify the phrases and restore them.
- Lemmatization, removing affixes from words, for example by changing third person words to first person and past and future tense verbs to present tense.
- Stemming, simplifying words to their root form.
- Filter out meaningless "stop words".

Data preprocessing plays a crucial role throughout the process of data analysis. Data cleaning, data integration, and data transformation are the three basic stages of data preprocessing, which transform raw data into a suitable format for further analysis. It plays a vital role in ensuring that the input data is reliable, high quality, and accessible to data analysis techniques. Raw data often contain various issues such as missing values, outliers, inconsistencies, and errors, which may hinder the accuracy and effectiveness of any subsequent analysis. The following steps are carried out in this paper:

- Data cleaning: handling of outliers and missing values.
- Data integration: Eliminate duplicate columns that are unnecessary for this study and create necessary columns in accordance with specifications, such as dividing the composite genre column into a single label.
- Data transformation: standardization of the data; discretization, including rating 1–10 continuous processing into high, medium and low rating groups; Boolean values of sentiment mapping into 0, 1; calculation conversion, such as duration (hh/mm) to time (minutes).

4. Model and results

Both Bi-LSTM sentiment model and LDA topic modeling are based on the python platform.

4.1. Sentiment classification model construction

4.1.1. Corpus

The format of the training data for the experiments was <text, label > format, i.e., one review text and one label per row, separated by tabs, with labels of 0 or 1, i.e., positive emotions were labeled as 1 and negative

emotions as 0.

4.1.2. Parameters

Although there are many limitations of more advanced and well-trained algorithms, there are not many artificial intelligence algorithms for text mining, and Bi-LSTM implements two-way sentiment evaluation, which can more accurately classify the sentiment of movie reviews. Use Bi-LSTM to extract the feature attributes between sentences, max pool the output of the hidden layer of Bi-LSTM, then dropout, and then inputs to the activation function, and finally outputs the judgment result. The following were the settings for the experiment's parameters: Set the vocabulary size to 6000 and the word vector dimension to 128. There was a 0.2 Dropout setting. **Table 3** contains a summary of the sentiment classification model.

Table 3. Summary of sentiment classification model.

Layer	Output shape	Parameter
Embedding	(None, None, 128)	768,000
Bi-LSTM	(None, None, 64)	41,216
Max pooling	(None, None, 64)	0
Linear	(None, 20)	1300
Dropout	(None, 20)	0
Output	(None, 1)	21

The parameter setting code of Bi-LSTM is as follows:

```
max features= 6000 # the size of vocabulary
embed size = 128 # the dimension of word vector
model = Sequential()
model.add(Embedding(max features, embed size))
model.add(Bidirectional(LSTM(32， return sequences = True)))
model.add(GlobalMaxPoo11D())
model.add(Dense(20, activation="relu"))
model.add(Dropout(0.85))model.add(Dense(1, activation="sigmoid"))
model.compile(loss='binary_crossentropy, optimizer='adam', metrics=['accuracy'])
```

4.1.3. Training

Accuracy and loss values are used to determine whether the model is over-fitted and to determine the values of batch size and epoch. 15% data was used as the validation set.

Epoch equal to 3 is an inflection point, as shown in **Figure 3**. When epoch is greater than 3, the model performs increasingly well on the training set, but the loss value in the validation set rises, indicating that the model has been overfitted.

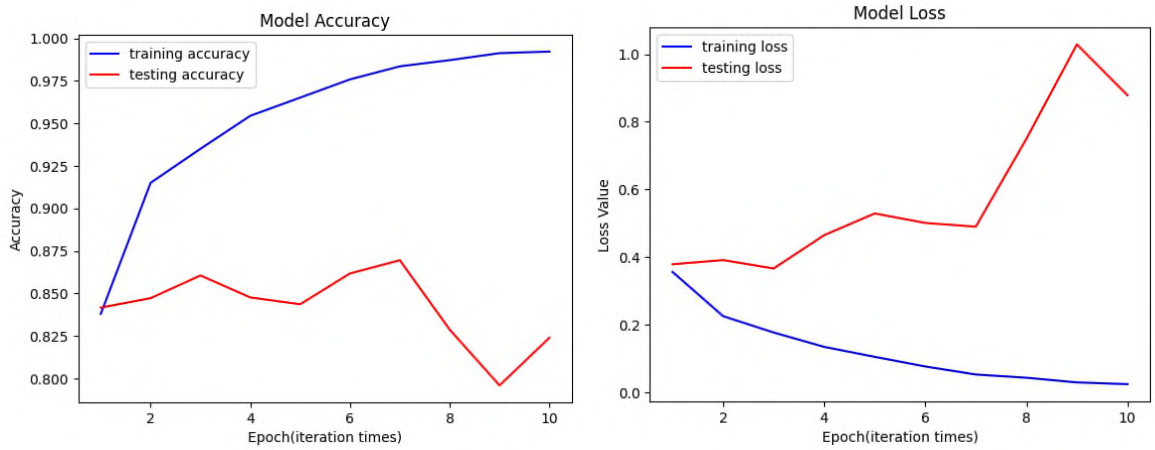

Figure 3. Learning curve of accuracy and loss vs epoch.

4.1.4. Model evaluation

Accuracy in the testing dataset and confusion matrix were used to assess the sentiment categorization model's effectiveness. The test accuracy of the model was 83.11%. The performance of the model is good, as evidenced by the high number of samples that were accurately predicted, as shown in **Figure 4**.

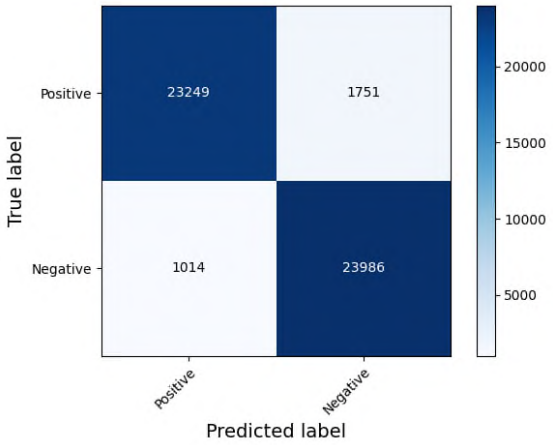

Figure 4. Confusion matrix of model in the testing set.

4.2. Statistical analysis and visualization

Statistical analysis and visualization are used to draw insights and conclusions from datasets by identifying trends, patterns, and relationships between variables. **Figure 5** indicates that a large number of films are released each year in November, December, June, July and May, with November having the highest average number of movies released at 568. However, the average rating of movies for each month presented in **Figure 6** shows that there is no significant relationship between the month and the rating a movie receives.

Figure 7 shows the frequency of different movie genres in the dataset. The top five genres among 21 film genres with the highest frequency count are Drama and Comedy, Action, Adventure and Crime. (Note: Due to the existence of cross-genre movies, a movie may contain more than one genre tag, the count is the number of occurrences in each genre tag.)

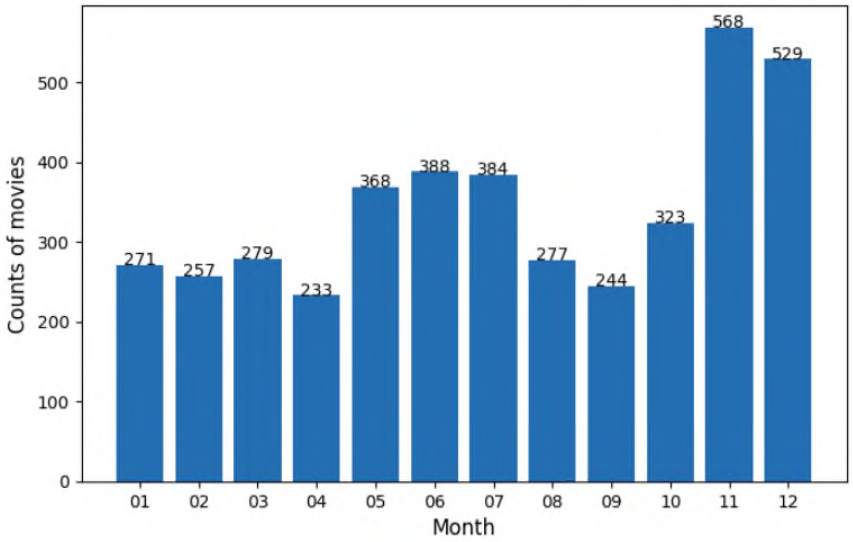

Figure 5. Bar plot of counts of films released on each month.

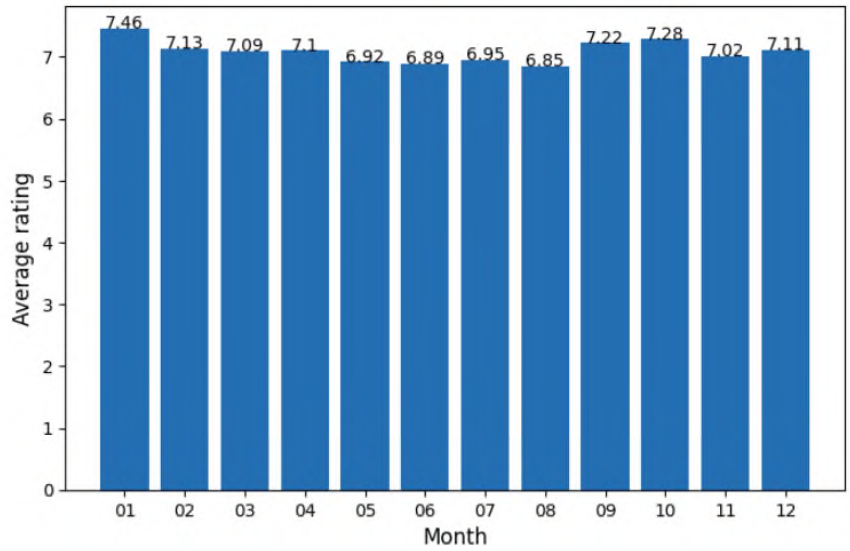

Figure 6. Bar plot of average ratings of films released on each month.

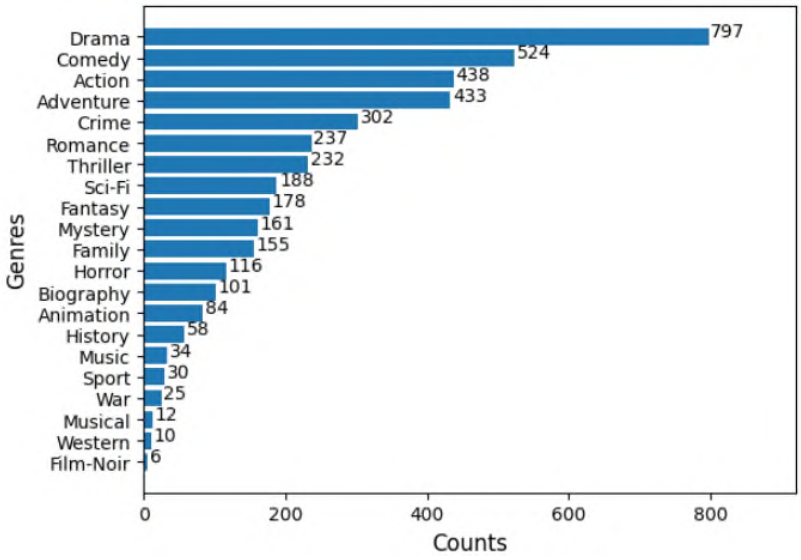

Figure 7. The frequency distribution histogram of movie genres.

The trend of movies is shown in **Figure 8**, with drama being the first to develop and the fastest growing genre. It was not until the 1990s that other genres of movies began to enter the market and develop rapidly. From 2010, comedies and other genres besides drama are rapidly pulling away, with a tendency to catch up with dramas. In addition to this, action, adventure and science-fic movies are also growing rapidly, following closely behind comedies. However, the music and western fantasy genres were relatively rare and did not grow significantly.

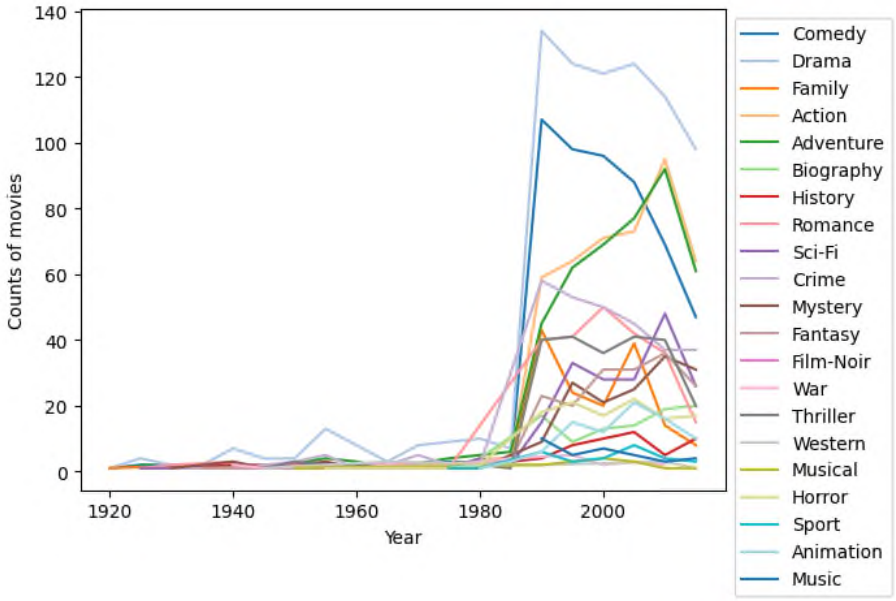

Figure 8. Plot of counts of films by genres.

However, the average of their rating is not as high as their frequency, shown in **Figure 9**. On the contrary, genres such as Film-Noir and War have low frequency counts, but they received very high ratings.

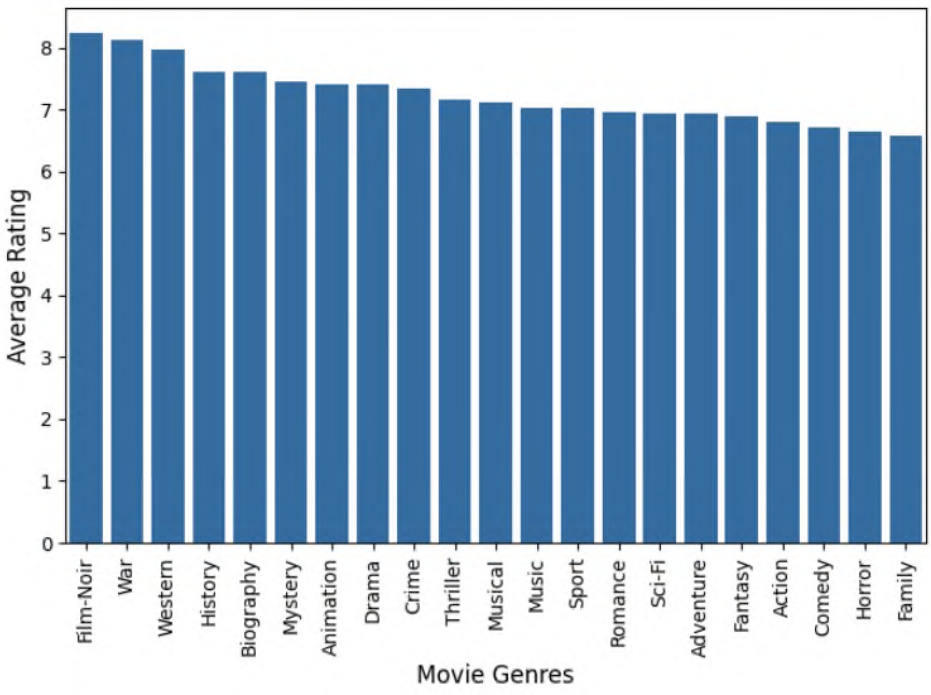

Figure 9. The average rating of movie genres.

Figure 10 shows the average positive sentiment rate of different movie genres. The top five in terms of positive sentiment rate are Film-Noir, Biography, Animation, Music, and History, while action, horror, and Sci-Fi genres do not have a very high positive sentiment rate.

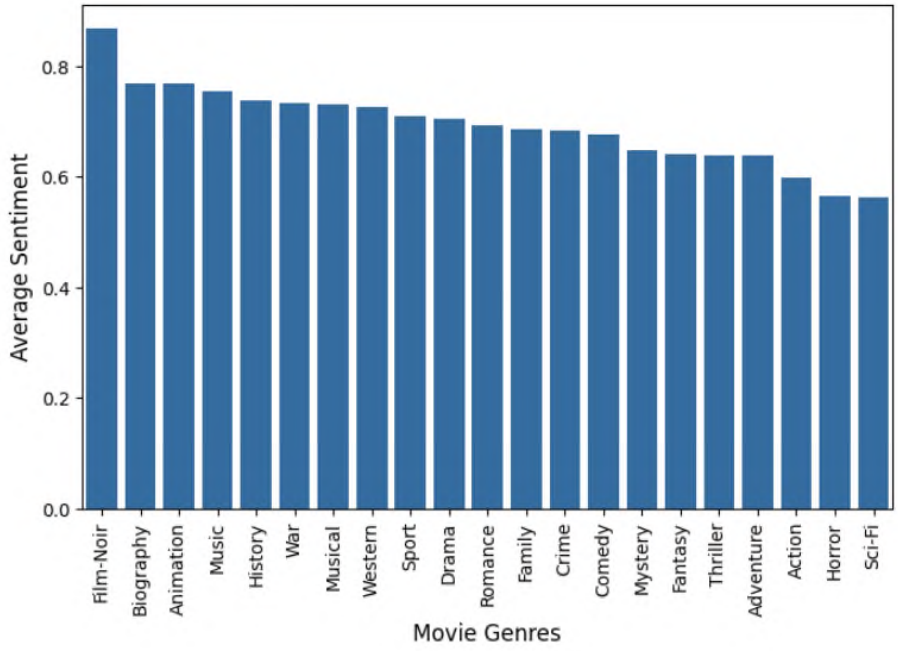

Figure 10. Bar plot of average positive sentiment rate of different movie genres.

To study the relationship between rating and sentiment positive rate, the scatter plot shown in **Figure 11** displays the data between average rating and average positive sentiment rate across different genres of movies. It reveals a clear positive correlation between the average rating vs average positive sentiment rate. The ratings and sentiment for horror movies were the lowest, while those for Film-Noir movies had the highest ratings and positive sentiment rate.

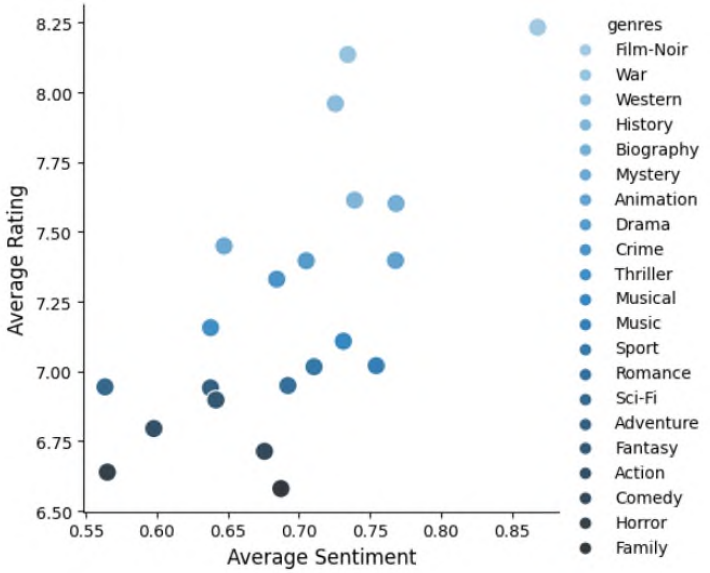

Figure 11. Scatter plot of average rating vs average positive sentiment rate among different genres.

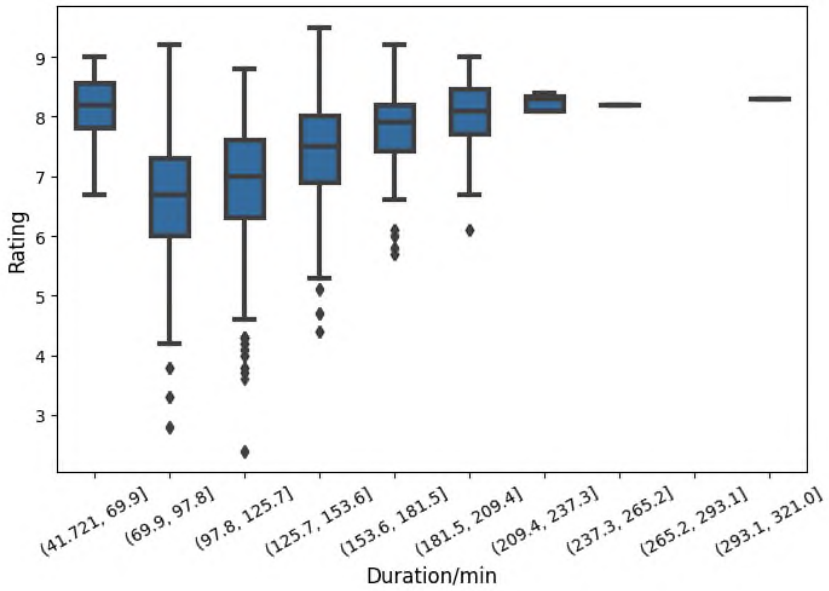

Figure 12. Box plot of average rating vs duration.

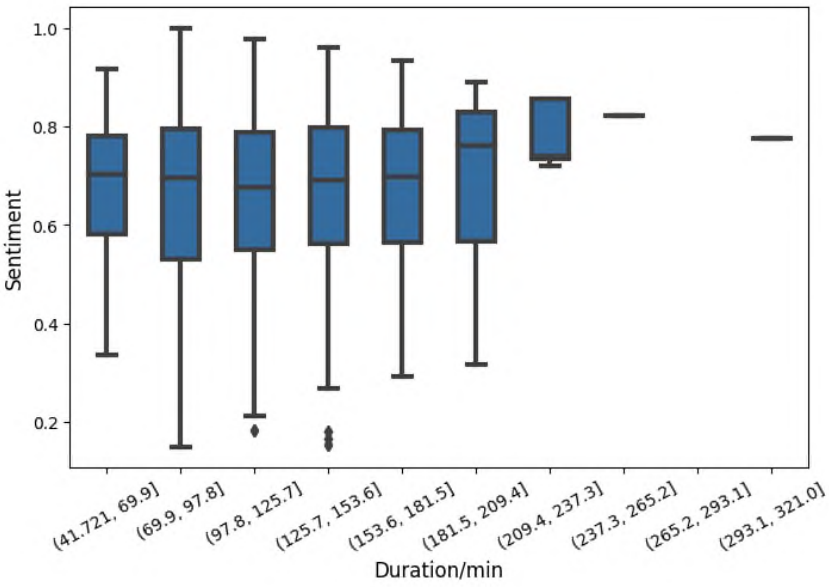

Figure 13. Box plot of average positive sentiment rate vs duration.

Figure 12 indicates the correlation between duration and IMDb rating. Short films have both high and low scores, but films that are over 200 minutes long basically get a score of 7 or more, and the duration of films with IMDb scores close to or below 4 are all less than 130 minutes. Movies with rich plots and complete storylines can effectively engage viewers and evoke their emotions. It was observed that movies with durations under 60 minutes, yet received high ratings, were predominantly from the Drama, Mystery, and Crime genres. These genres are characterized by their tightly-paced, dramatic storytelling, making them more appealing to audiences. However, **Figure 13** indicates the average positive sentiment rate is irrelevant with duration. This implies that the sentiment of audiences is primarily influenced by key elements rather than the quantity of information.

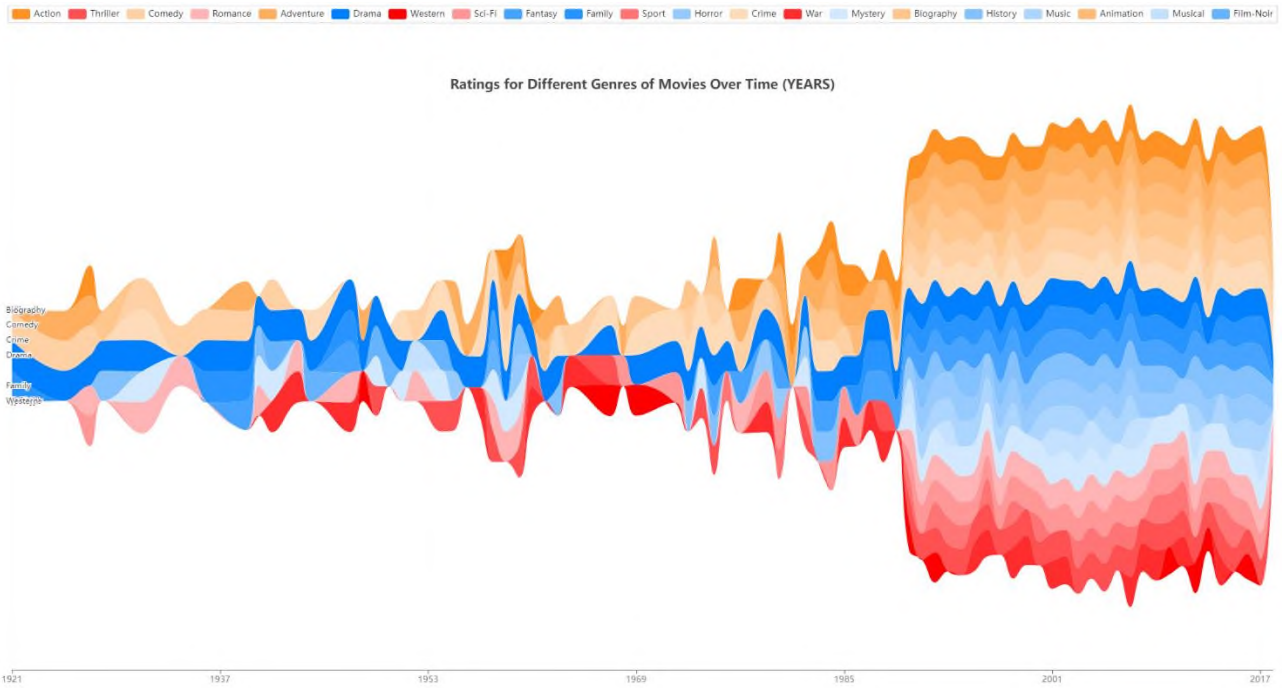

Figure 14. Thematic river plot of average rating among different genres over time.

The thematic river plot displays the average rating of genres of films over time, as shown in **Figure 14**. It can help identify significant changes and patterns in the evolution of genres and their ratings for each genre. The thematic river diagram shows that the film industry has seen a spurt of growth since the 1990s, and the trend of increasing numbers continues to get bigger, especially in the genres of action and comedy films which are more accepted by the public. Making films in these genres should get better feedback and reduce the risk at the box office. For the past 30 years, the genres and ratings of movies have remained largely stable.

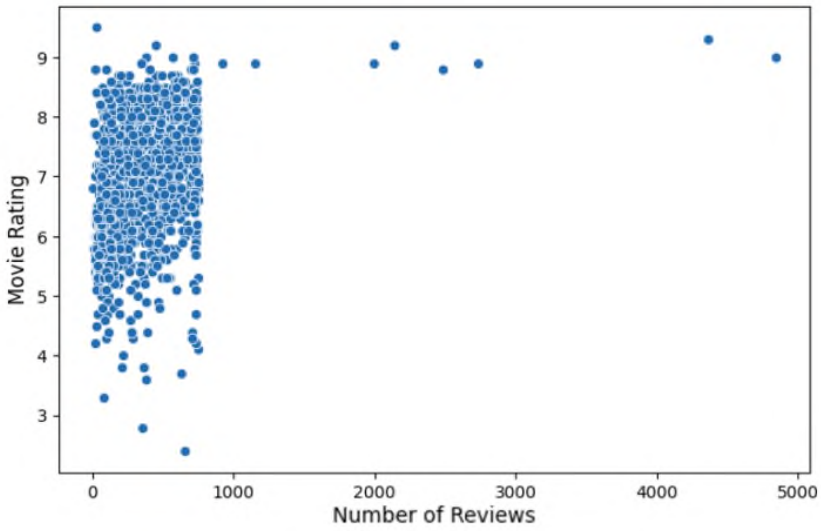

Figure 15. Scatter plot of the number of reviews and movie ratings.

Figure 15 shows the scatter plot of the number of reviews and movie ratings between the number of ratings and the final rating a movie received is weak; high rating movies does not mean a high number of reviews, and the low rating movies can have a large number of reviews.

4.3. Topic analysis

Topic models are evaluated based on their ability to describe documents well (i.e., with low confusion) and produce topics with coherent semantics[24]. This study selects the appropriate number of topics based on two parameters, perplexity and coherence. **Figure 16** shows the perplexity and coherence of LDA model with the increasing of topic. It can be seen that taking 7 topics is most appropriate when the perplexity and coherence keep a good balance. Set the number of iterations to 300 and take 15 subject terms in order to see the main idea of the subject.

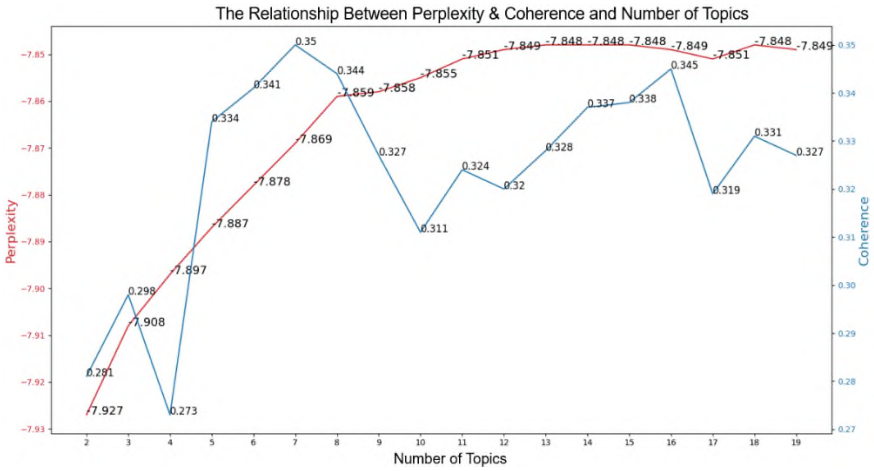

Figure 16. Perplexity and coherence vs number of topics.

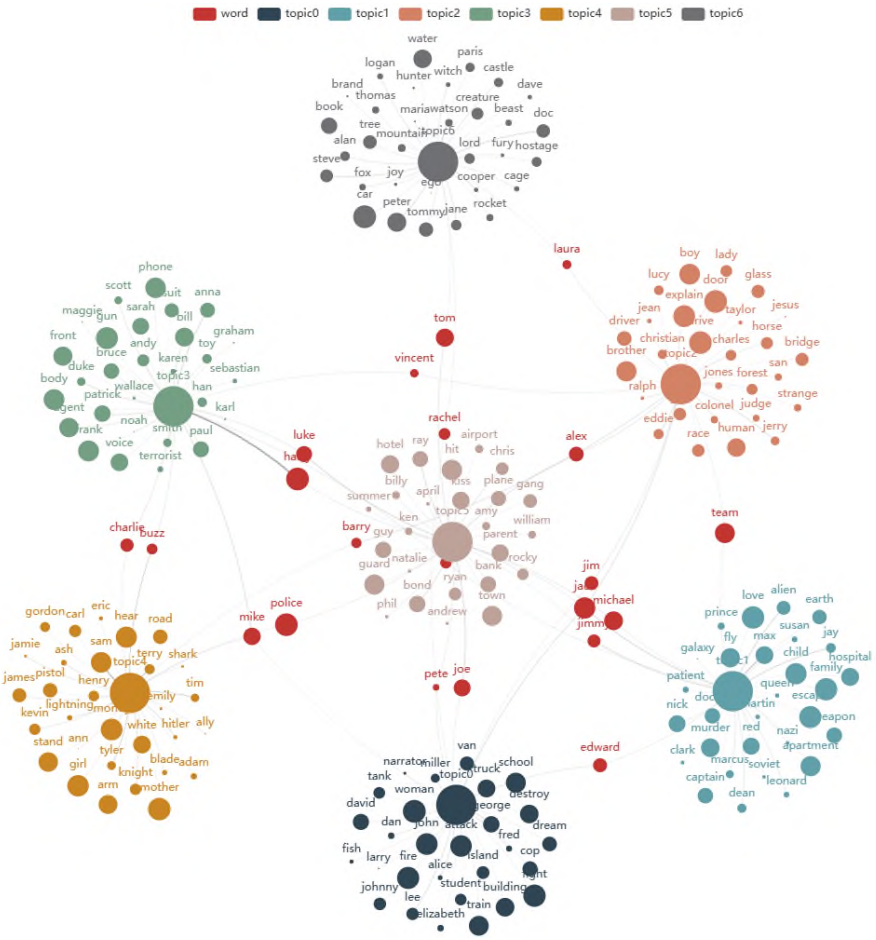

Figure 17. LDA model of seven topics.

Figure 17 shows the contents of the seven themes. Based on the topic words "lord", "castle", "cage", "mountain", "rocket", "tree", and "hostage", the topic 6 content could be a medieval or fantasy adventure movie with a sci-fi twist. The "lord" and "castle" suggest a medieval setting and a character of high status and power. The "mountain" and "tree" indicate that the story takes place mainly in the wilderness. "Cage" suggests a character being held captive or imprisoned, possibly a hostage. "Rocket" suggests a sci-fi element, possibly time travel or space travel. Overall, these movies could be a thrilling adventure with a mix of medieval and sci-fi elements, featuring a heroic quest, daring escapes, and epic battles.

For Topic 1, the topic word "love" suggests that the movie involves a romantic relationship between the characters. "Galaxy" and "alien" suggest that the movie is set in outer space and involves extraterrestrial life. "Escape" suggests that there is a plot involving characters trying to escape from something. "Hospital" and "patient" suggest that there is a medical or healthcare-related plot or setting. "Murder" suggests that there is a crime or murder plot involved. Finally, "earth" suggests that there is some connection to Earth in the movie. The topic of the movie is the story of the humans of Earth and the outsiders, mainly conveying love and peace, but also crime or fighting.

Based on the topic words "lady", "boy", "human", "brother", "horse", "forest", "bridge", and "strange", the movies with topic 2 content could be a fantasy adventure with a coming-of-age theme. Topic 3 could be thrilling and intense action stories with a plot involving espionage, terrorism, and political intrigue, featuring suspenseful action sequences and a high-stakes race against time, with the topic words like "agent", "terrorist", "bill", "body", "gun", and "rank". Based on the topic words "girl", "mother", "lightning", and "morning", the topic 4 content could be a heartfelt drama with a focus on family relationships and personal growth, featuring a sudden event that forces the characters to confront their own strengths and weaknesses. Based on the topic words "guard", "bank", "gang", "hit", "bond", and "airplane", topic 5 content could be a crime thriller with a plot involving a bank heist and a high-stakes chase. Based on the topic words "student", "school", "dream", "train", "fire", and "destroy", the topic 0 content could be a drama with a focus on personal growth, overcoming obstacles, and the power of education.

Topic 4 and Topic 5 share the same keywords: "police" and "milk".

Figure 18 shows the trend in high scoring films across different topics. There were large fluctuations in almost all themes, with theme 1 reaching its highest point in the most years. Topic 6 reached its highest point in 2014. The **Figure 19** shows more clearly the thematic distribution of high scoring films by year with heatmap. The graph shows that there are more films about topic 1 and topic 5. And topic 3 only appeared after the 1990s.

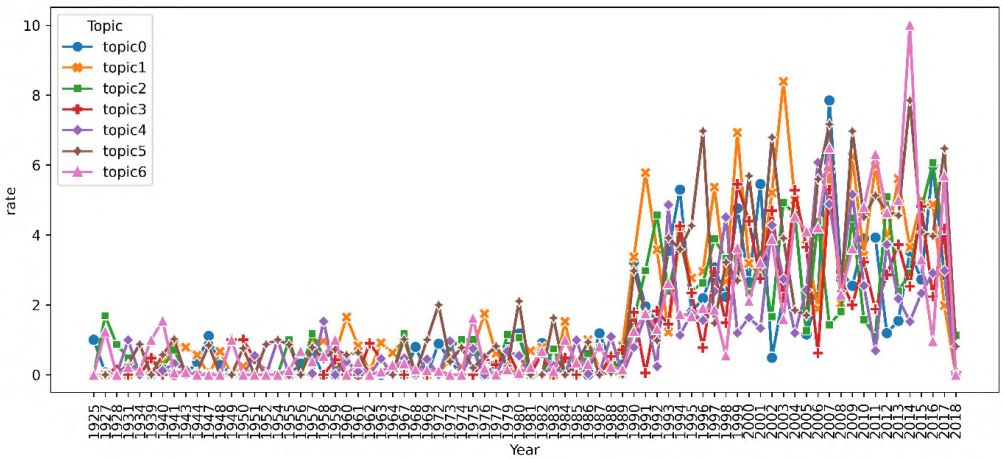

Figure 18. Line graph showing trends in the narrative themes of highly rated films.

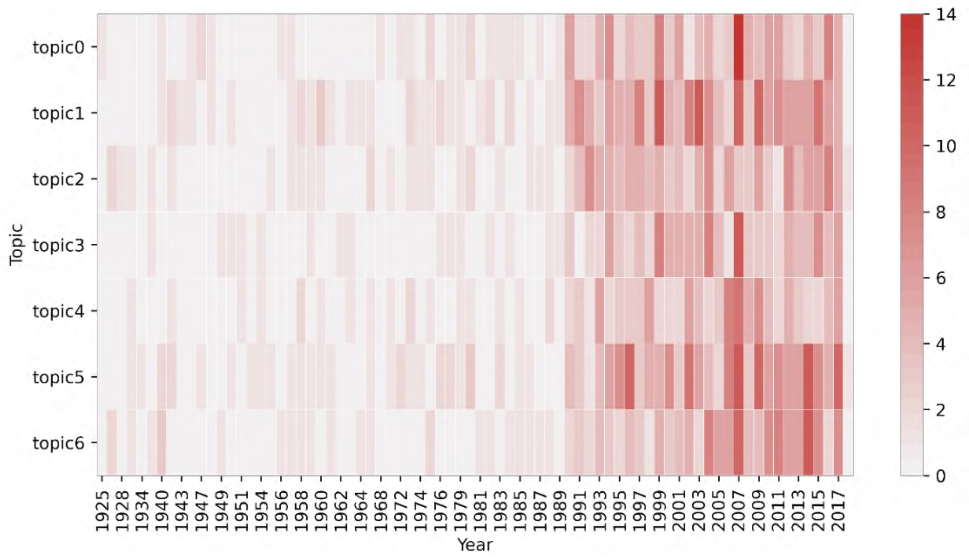

Figure 19. Heat map of the number of highly rated film themes.

5. Discussion

5.1. Discussion

Since the 1990s, the rapid development of high technology, centered around the information technology revolution, and the trend towards globalization have become mainstream. As a fusion of industry, technology, and art, the film industry has been greatly influenced by the overall economic environment of the world. Additionally, the popularity of the internet has also boosted the development of the film industry. IWOM for films has become increasingly important with the growth of the internet, and people can more easily access information online. From 1990 onwards, the film industry has experienced explosive growth (**Figure 8** and **Figure 14**). Although globalization and the popularity of the internet have provided a favorable development environment for the film industry, with the increasing aesthetic standards of audiences and more intense market competition, films that want to increase their chances of success not only need to seize current opportunities but, more importantly, focus on creating high-quality content.

Drama movies have been a popular and rapidly developing film genre, boasting a diverse range of themes and styles that are both innovative and firmly grounded. One of the key advantages of drama movies is their relatively low production cost, which allows them to be paired with other film genres as desired. Similarly, comedy movies have seen a significant increase in popularity and are now second only to drama movies in terms of production volume. This trend can be attributed to the relatively low production cost of both genres, which makes them an attractive investment for producers (**Figure 8**). The growing number of comedy movies being released is a reflection of the rising demand for lighthearted entertainment in the film market and underscores the commercial viability of comedy films as a profitable business model. The number of traditional music and western fantasy films has not increased rapidly, as their production costs are typically higher and require more investment in areas such as sets, costumes, music, cinematography, and sound effects. Additionally, these genres have a smaller audience and require a certain level of aesthetic taste.

The reason why movies are released more frequently in May to July, and November, December is mainly due to the fact that these two periods are when audiences are most active. During May, June, and July, schools have summer vacation, and many children and teenagers plan leisure activities, making movies a popular form of entertainment. Additionally, this period is an ideal time for movie studios to release films aimed at children and family audiences.

November and December mark the end-of-year holiday season and include major holidays such as Thanksgiving, Easter, and Christmas. People generally have more free time during this period and are looking for activities to enjoy with family and friends. As a result, movies become a popular choice for entertainment. However, while the end-of-year holiday season is a prime time for moviegoing, there does not appear to be a significant correlation between the time of a movie's release and its rating.

The differences in movie ratings and frequency counts across movie genres may Indicate that niche movies are more likely to receive high ratings, while under popular themes with high frequency counts, they may receive mediocre ratings. Innovation in movie elements may be the key to obtaining high ratings (**Figure 7** and **Figure 9**). Probably because they have some gore and violence in their plots. The reason for this is that these genres may have some gore and violence in their plots (**Figure 10**). Compared to action and horror films that rely on special effects and fighting, family films with a wide age range of audiences resulting in superficial content, and comedies with the same lack of profound content, the top five rated films are usually ambitious in their frameworks, have deeper connotations and often recreate history to provoke thought in the audience. Thus, profundity and thoughtfulness are an important factor in attracting audiences.

The study was given seven themes about what the content of high-rated films is talking about. Topics 1 and 5 are more predominant and they deal with people's love and curiosity for the unknown and exciting plots. (**Figures 14** and **15**) Specifically, people like content like aliens, galaxies and Earth because this theme deals with mysterious worlds and technological worlds that are unknown to the general public, providing audiences with great fantasy material. And violent plots such as gang battles, which ordinary people do not come across in their daily lives, are thrills that can only be experienced in the world of the screen. Escape, fight, murder and love, which deal with human existence and emotions, also bring tension and suspense.

5.2. Benefits and limitations

5.2.1. Benefits

1) The advantage of using Bi-LSTM for sentiment analysis modeling is that it can capture contextual information in movie reviews. Bi-LSTM deals with sequence data and takes historical information into account.

2) The method commonly used in LDA topic modeling. It is an unsupervised learning algorithm and does not rely on manual labeling.

3) This project uses an emotion-topic fusion analysis method based on Bi-LSTM and LDA to avoid the isolation of emotion analysis and topic mining and enable comprehensive text mining.

5.2.2. Limitations

1) The source of our data set is relatively limited. It is only the movie data of IMDb and cannot be generalized to the entire movie duration.

2) There will still be noise data in our emotion modeling. In order to focus more on the key information in movie reviews, we can consider the attention mechanism in the future.

6. Conclusions

In this study, statistical analysis and topic analysis were employed to explore the factors that contribute to popular movies. The following conclusions can be inferred from the computations and analysis included in this paper.

- The findings suggest that Film-Noir and War films have a greater likelihood of receiving high ratings, indicating their popularity among moviegoers.

- The results demonstrate a positive correlation between audience sentiment and ratings in most genres.
- High-rated movies are more likely to feature plot elements centered on interstellar and war themes, as well as include gun violence and terrorist activity.

This study investigates the various factors that contributed to high ratings for movies, with the aim of gaining deeper insights into the audience's perceptions. As this study only analyze datasets from the Kaggle platform, the findings and results may not be representative of the opinions and perceptions of the community as a whole. Therefore, a larger sample and a more scientific approach to data analysis would be worth adopting. By understanding why movies are received positive sentiment and high ratings, for future research into the fusion of artificial intelligence and digital media, this study offers useful reference points, this study provides valuable reference points for future research into the integration of artificial intelligence and digital media, as well as offering unique perspectives and experiences for the film industry. With advances in NLP, the benefits of employing NLP methodology in social media data analysis have been widely acknowledged. It is expected that the use of NLP methods for social media data analysis will continue to expand and revolutionize the field, presenting fresh chances for companies and researchers to draw insightful conclusions from massive amounts of unstructured text data.

The next step will be to expand data sources and achieve full acquisition of movie data, thereby ensuring the comprehensiveness of text mining results. In terms of topic mining, the performance of mining methods can be further improved; in terms of sentiment analysis, more fine-grained classification can also be performed.

Conflict of interest

The author declares no conflict of interest.

References

1. Zhang Y, Zhang L. Movie recommendation algorithm based on sentiment analysis and LDA. *Procedia Computer Science* 2022; 199: 871–878. doi: 10.1016/j.procs.2022.01.109
2. Bhuvaneshwari P, Rao AN, Robinson YH, Thippeswamy MN. Sentiment analysis for user reviews using Bi-LSTM self-attention based CNN model. *Multimedia Tools and Applications* 2022; 81(9): 12405–12419. doi: 10.1007/s11042-022-12410-4
3. Topal K, Ozsoyoglu G. Movie review analysis: Emotion analysis of IMDb movie reviews. In: Proceedings of the 2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM); 18–21 August 2016; San Francisco, CA, USA. pp. 1170–1176.
4. Sharma R, Morwal S, Agarwal B. Named entity recognition using neural language model and CRF for Hindi language. *Computer Speech & Language* 2022; 74: 101356. doi: 10.1016/j.csl.2022.101356
5. Trivedi SK, Dey S, Kumar A. Capturing user sentiments for online Indian movie reviews: A comparative analysis of different machine-learning models. *The Electronic Library* 2018; 36(4): 677–695. doi: 10.1108/EL-04-2017-0075
6. Kanani S, Patel S, Gupta RK, et al. An AI-enabled ensemble method for rainfall forecasting using long-short term memory. *Mathematical Biosciences and Engineering* 2023; 20(5): 8975–9002. doi: 10.3934/mbe.2023394
7. Rehman AU, Malik AK, Raza B, Ali W. A hybrid CNN-LSTM model for improving accuracy of movie reviews sentiment analysis. *Multimedia Tools and Applications* 2019; 78: 26597–26613. doi: 10.1007/s11042-019-07788-7
8. Hourrane O, Idrissi N, Benlahmar EH. Sentiment classification on movie reviews and twitter: An experimental study of supervised learning models. In: Proceedings of the 2019 1st International Conference on Smart Systems and Data Science (ICSSD); 3–4 October 2019; Rabat, Morocco. pp. 1–6.
9. Shaukat Z, Zulfiqar AA, Xiao C, et al. Sentiment analysis on IMDB using lexicon and neural networks. *SN Applied Sciences* 2020; 2(2): 1–10. doi: 10.1007/s42452-019-1926-x
10. Arora E, Mishra S, Kumar KV, Upadhyay P. Extending bidirectional language model for enhancing the performance of sentiment analysis. In: Gunjan V, Senatore S, Kumar A (editors). *Advances in Cybernetics, Cognition, and Machine Learning for Communication Technologies*. Springer; pp. 133–141.
11. Chirgaiya S, Sukheja D, Shrivastava N, Rawat R. Analysis of sentiment based movie reviews using machine learning techniques. *Journal of Intelligent & Fuzzy Systems* 2021; 41(5): 5449–5456. doi: 10.3233/JIFS-189866

12. Acikalin UU, Bardak B, Kutlu M. Turkish sentiment analysis using bert. In: Proceedings of the 2020 28th Signal Processing and Communications Applications Conference (SIU); 5–7 October 2020; Gaziantep, Turkey. pp. 1–4.
13. Wu J, Ye C, Zhou H. BERT for sentiment classification in software engineering. In: Proceedings of the 2021 International Conference on Service Science (ICSS); 14–16 May 2021; Xi'an, China. pp. 115–121.
14. Kaushik K, Parmar M. Sentiment analysis based on movie reviews using various classification techniques: A review. *International Journal of Scientific Research in Computer Science Engineering and Information Technology* 2021; 7(3): 197–208. doi: 10.32628/CSEIT217329
15. Hakim AA, Erwin A, Eng KI, et al. Automated document classification for news article in Bahasa Indonesia based on term frequency inverse document frequency (TF-IDF) approach. In: Proceedings of the 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE); 7–8 October 2014; Yogyakarta, Indonesia. pp. 1–4.
16. Yang Q. LDA-based topic mining research on China's government data governance policy. *Social Security and Administration Management* 2022; 3(2): 33–42. doi: 10.23977/socsam.2022.030205
17. Blei DM, Ng AY, Jordan MI. Latent dirichlet allocation. *Journal of Machine Learning Research* 2003; 3: 993–1022.
18. Newman D, Lau JH, Grieser K, Baldwin T. Automatic evaluation of topic coherence: *Proceedings of the Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*; 2–4 June 2010; Los Angeles, California, USA. pp. 100–108.
19. Musat CC, Velcin J, Trausan-Matu S, Rizoiu MA. Improving topic evaluation using conceptual knowledge. In: Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI); 16–22 July 2011; Barcelona, Catalonia, Spain. pp. 1866–1871.
20. Baroni M. Composition in distributional semantics. *Language and Linguistics Compass* 2013; 7(10): 511–522. doi: 10.1111/lnc3.12050
21. Roder M, Both A, Hinneburg A. Exploring the space of topic coherence measures. In: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining; 2–6 February 2015; Shanghai, China. pp. 399–408.
22. IMDB Movie Reviews with ratings. Available online: https://www.kaggle.com/datasets/nisargchodavadiya/imdb-movie-reviews-with-ratings-50k (accessed on 25 September 2023).
23. Tan KL, Lee CP, Lim KM. Roberta-Gru: A hybrid deep learning model for enhanced sentiment analysis. *Applied Sciences* 2023; 13(6): 3915. doi: 10.3390/app13063915
24. Ding R, Nallapati R, Xiang B. Coherence-aware neural topic modeling. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing; October–November 2018; Brussels, Belgium. pp. 830–836.