Sheffield Hallam University | College of Business, Technology and Engineering

MSc Dissertation Report

Predicting Box Office Success with Sentiment Analysis of Movie Reviews.

A dissertation submitted in partial fulfilment of the requirements of Sheffield Hallam University for the degree of Master of Science in **Big Data Analytics**

Student Name	OLUWATOBI GIDEON MAUTIN
Student ID	32062745
Supervisor	DR. DIANA HINTEA
Date of Submission	15, JANUARY 2024

ACKNOWLEDGMENT

I would like to express my deepest gratitude to Dr. Diana Hintea for her invaluable support throughout the course of this dissertation. Her guidance, mentorship, and unwavering commitment to the pursuit of knowledge have been instrumental in the successful completion of this research endeavour, to my family and friends for their unconditional support, and most importantly to the almighty God for making everything possible, I am eternally grateful.

ABSTRACT

In the dynamic world of the film industry, accurately predicting a movie's financial success is crucial and traditional methods have relied on factors such as genre, star power, promotion and marketing efforts, sentiment analysis has emerged as a transformative computational linguistics technique that is revolutionizing the art of movie box office projection.

This dissertation explores the innovative application of sentiment analysis in assessing public opinion through movie reviews by identifying and categorizing sentiments within the textual reviews, this method offers a profound insight into a film's potential appeal and audience reception. Despite the inherent challenges of data quality and language complexity, sentiment analysis has proven to be an invaluable asset in film production, marketing, and distribution, significantly enhancing the industry's success.

The project presented herein delves into the process of utilizing sentiment analysis to predict movie financial performance. It involves comprehensive data preprocessing, exploratory data analysis (EDA), and the construction of a sentiment analysis model. The project found a strong positive correlation of 0.8466 between sentiment scores and box office success, indicating that movies with higher sentiment scores tend to have higher box office success. Furthermore, the Long Short-Term Memory (LSTM) model used achieved an accuracy of 99.39%, a precision of 99.62%, a recall of 99.39%, an F1 score of 99.50%, and an ROC AUC score of 99.84%, demonstrating its effectiveness in predicting a movie's box office potential based on sentiment analysis of its reviews.

The objective is to achieve an advanced understanding of audience sentiments, thereby equipping filmmakers, studios, and distributors with the knowledge to make strategic decisions leading to impressive results.

1. INTRODUCTION

1.1 (BACKGROUND)

The film industry narratives trace back to the 19th century, marked by the convergence of photography and the quest to immortalize motion (Coe, 1996). Innovators such as Muybridge, Edison and the Lumière brothers, catalysed the development of early cinematic technology (Musser, 1994). The dawn of the 20th century heralded narrative storytelling and the rise of Hollywood as the epicentre of the U.S. film industry, although Griffith was the visionary that transformed filmmaking into an art form, setting the stage for the industry's golden age (Thompson & Bordwell, 1994).

The introduction of sound and colour brought about a new cinematic era, while post-World War II developments introduced fresh genres and the invention of television (Nowell, 2014). The close of the 20th century saw the rise of blockbusters and movies using special effects, with the early 21st century witnessing the disruptive impact of streaming services on traditional cinema (Bernoff, 2016). The film industry is a global business, comprising of various institutions, including film production companies, studios, cinematography, animation, screenwriting, pre-production, post-production, film festivals, distribution, and actors. The industry is driven by the pursuit of box office success, which is often measured by the amount of money a movie earns in theatres (Eisenberg, 2019).

The film industry reached a peak in global box office revenue of $38 billion in 2015, with five films grossing over $1 billion each, a significant achievement in Hollywood's history (Peterson, 2016). These blockbusters were predominantly produced by six major studios. By 2018, the industry's value, including both box office and home entertainment, was estimated at $136 billion (Smith, 2019).

Hollywood is recognized as the paramount in terms of box office revenue, while Bollywood is noted for its high volume of film production, with 2,446 feature films in 2019 (Goonasekara, 2020), however, a film's box office gross does not always reflect its profitability, particularly for blockbusters with large budgets, due to high marketing costs, suggesting the need for a nuanced return on investment (ROI) metric (Cunningham, 2015; Epstein, 2017).

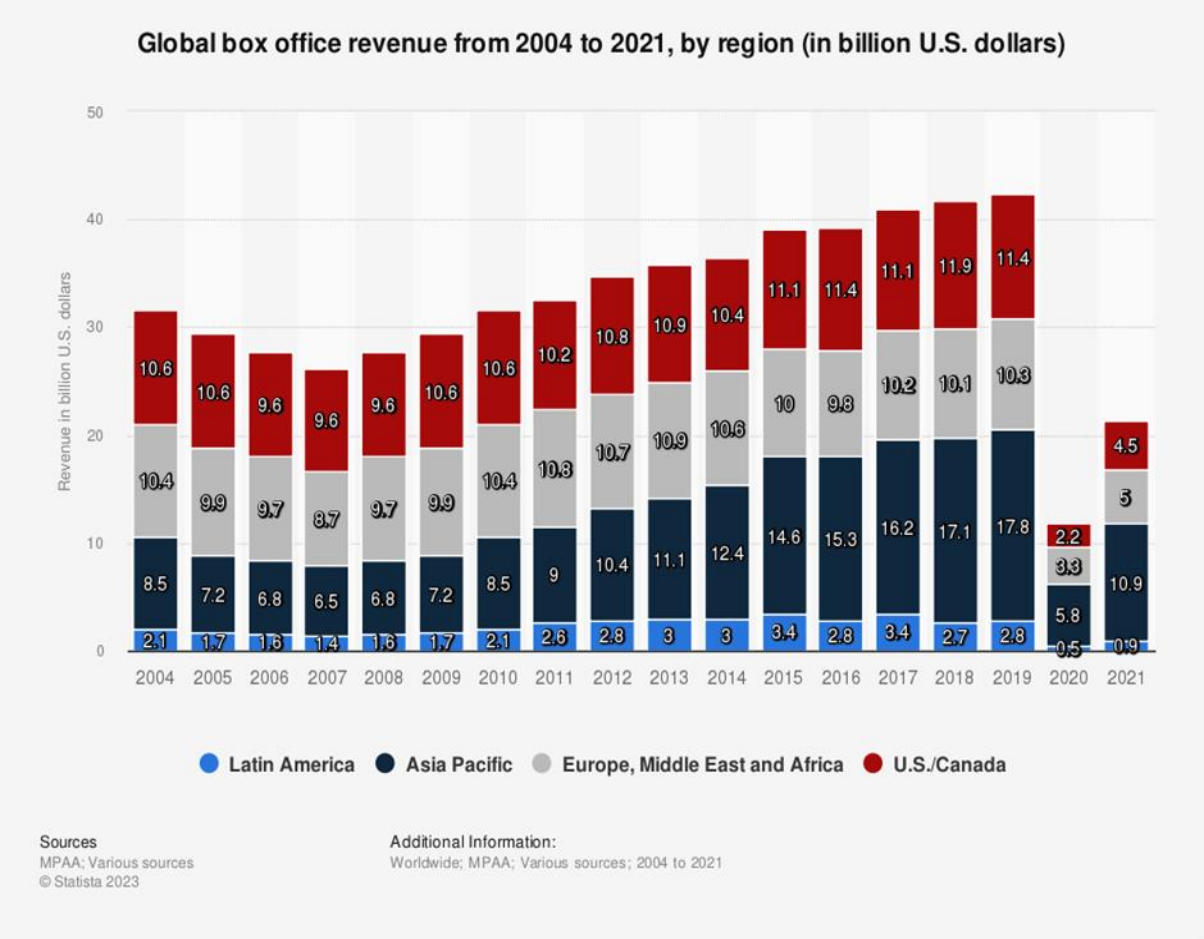


Figure 1. Global box office revenue from 2004 to 2021, by region (in billion U.S. dollars)

1.2 MOTIVATION

The global box office revenue from 2004 to 2021 by region highlights the fluctuations and trends that underscore the need for more sophisticated predictive tools capable of adapting to these changes in a rapidly evolving landscape (Statista, 2021). The rise of streaming services has significantly impacted the traditional box office revenue (Esquire, 2021) coupled with the sheer volume of film releases and the dwindling attention span of the target audience already swayed by the star power of actors (Nelson, Randy A., 2012) and the allure of sensational movie trailers (S. Oh, J. Ahn and H. Baek, 2015) necessitates a re-evaluation of traditional success metrics (Kim, A., Trimi, S. & Lee, SG, 2021). These factors, coupled with the competitive landscape of entertainment underscores the need for precise prediction models that can adeptly handle the complexities of contemporary film consumption (Lee, K., Park, J., Kim, I. et al., 2016).

1.3 PROJECT OBJECTIVES

The objectives of this dissertation project, which is cantered around the film industry, includes investigating the historical and technological evolution of the movie industry, the factors contributing to a movie's box office success, and the potential of sentiment analysis and machine learning in predicting box office outcomes.

The main objectives of this dissertation project are listed as follows:

1. Explore the historical and technological evolution of the film industry.

2. Analyse the factors influencing box office success.

3. Investigate the application of sentiment analysis in predicting box office performance.

4. Evaluate the effectiveness of machine learning algorithms in enhancing prediction accuracy.

1.4 PROJECT RESEARCH QUESTION:

The primary question that this project seeks to answer:

"***How can sentiment analysis of movie reviews effectively predict box office sales and inform strategic decision-making within the film industry?***"

1.4.1 Sub-Questions

To further dissect the main research question to explore in this project, the following sub-questions have been formulated:

1. What is the correlation between the sentiment score, derived from movie reviews, and a movie's box office success?

2. How accurately can the Long Short-Term Memory (LSTM) model predict a movie's box office success based on sentiment analysis of its reviews?

These sub-questions will help in understanding the nuances of the main research question and provide comprehensive answers to it.

1.5 PROJECT REPORT STRUCTURE:

The structure of this project report is organized into five distinct chapters, each serving a specific purpose:

Chapter 1: Introduction - Sets the stage for the research, outlining the background, motivation, and objectives.

Chapter 2: Literature Review - Examines the evolution of the film industry and the role of sentiment analysis in predicting box office success.

Chapter 3: Methodology - Details the methodologies for sentiment analysis and machine learning, including data collection and model evaluation.

Chapter 4: Results and Discussion - Analyses the findings and discusses their implications for box office sales and industry decision-making.

Chapter 5: Conclusion and Future Work - Summarizes the research, evaluates its impact, and suggests future research directions.

Each chapter contributes to a comprehensive understanding of the project and promises to answer the main project research question.

With the project objectives and questions clearly outlined, the focus now shifts to the literature review for chapter 2. This chapter will provide a comprehensive exploration of the film industry's evolution and the transformative role of sentiment analysis in predicting box office success. This will set the stage for understanding the context and relevance of the dissertation project.

2. LITERATURE REVIEW

2.1 INTRODUCTION

Sentiment analysis is a subfield of natural language processing (NLP) dedicated to understanding people's feelings, attitudes, emotions, and opinions towards a wide range of entities such as products, services, issues, events, topics, and their characteristics (Liu 2015). In this context, social media platforms have emerged as a crucial channel for expressing emotions globally, leading to the generation of a vast amount of unstructured data on the internet every second (Nandwani and Verma, 2021). However, to gain insights into the human psychology, it is imperative to process this data as quickly as it is produced, a task that can be effectively achieved through sentiment analysis (Nandwani and Verma, 2021).

Sentiment analysis enables the monitoring of public sentiment towards specific entities generating actionable insights in the process, such insights can be leveraged to comprehend, elucidate, and forecast the social trends. In the business landscape, sentiment analysis is very instrumental in refining strategies and obtaining valuable feedback from customers about their products. Moreover, in the current era of customer-centric business practices, gaining a deep understanding of the customer has become of significant importance (Ligthart et al. 2021).

In the context of this project, sentiment analysis is applied to textual movie reviews to predict box office success, the idea is to analyse the sentiments expressed in the reviews and use them as a predictor for the movie's financial success. This approach assumes that the general public's overall sentiment towards a movie, as expressed in text reviews, is indicative of the movie's popularity and, consequently, its box office performance.

2.2 AN INTRODUCTION TO MACHINE LEARNING

Machine learning is a branch of artificial intelligence that equips systems with the capability to learn and improve their performance from experiences autonomously, without the need for explicit programming (Sarker, 2021). The emphasis is on creating computer programs that can tap into data and utilize it for self-learning (Sarker, 2021).

Machine learning algorithms can be primarily classified into three types:

i. **Supervised learning**: the model is trained on a labelled dataset, i.e., a dataset where the target variable is known, usually used for classification and regression problems (Sohail et al., 2022).

ii. **Unsupervised learning**: deals with unlabelled data. The model learns through observation and finds structures in the data (Sarker, 2021).

iii. **Reinforcement learning**: is a type of machine learning where an agent learns to behave in an environment, by performing certain actions and observing the results (Sarker, 2021).

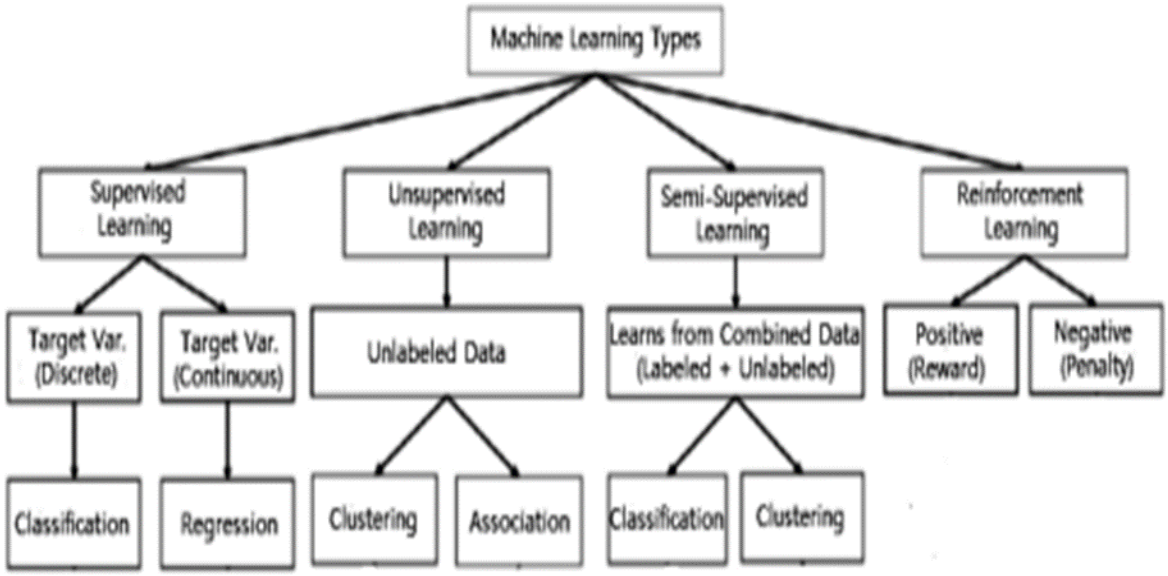


Figure 2. Types of Machine Learning Techniques (Sarker, 2021)

In recent years, deep learning has emerged as a powerful machine learning technique, being a part of a broader family of the machine learning methods, it can intelligently analyse data on a large scale using artificial neural networks with several layers (hence the term 'deep') to model and understand complex patterns in datasets (Sarker, 2021).

Machine learning algorithms, including deep learning, are widely used in various real-world application domains, such as cybersecurity systems, smart cities, healthcare, e-commerce, agriculture, and many more (Sarker, 2021).

2.3 SENTIMENT ANALYSIS

Sentiment analysis is also known as opinion mining, it is the computational study of people's opinions (Ligthart et al. 2021) assesses whether the author has a negative, positive, or neutral attitude toward an item, administration, individual, or location (Nandwani and Verma, 2021). In some applications, sentiment analysis is insufficient and will therefore require emotion detection which determines an individual's emotional/mental state precisely. "Emotion detection," "affective computing," "emotion analysis," and "emotion identification" are all phrases that are sometimes used interchangeably (Nandwani and Verma, 2021).

The field of sentiment analysis has been the topic of extensive research in the past decades (Ligthart et al. 2021). Studies on sentiment analysis mainly focus on the framework and lexicon construction, feature extraction, and polarity determination (Nanli et al., 2012). The unique features, algorithms, and datasets used in the analysis models are mapped so challenges and open problems identified that can help to identify points that require research efforts in sentiment analysis (Ligthart et al. 2021).

2.4 SENTIMENT ANALYSIS IN BOX OFFICE PREDICTION

The application of sentiment analysis in predicting box office performance is a relatively new but a rapidly growing field of study (Yang et al., 2023). The underlying premise is that the sentiments expressed in movie textual reviews can serve as a reliable indicator of a movie's popularity and, by extension, its box office success (Yang et al., 2023).

Several studies have employed various methodologies to predict box office performance based on sentiment analysis of movie reviews. For instance, the research by Yang, Xu, and Tu proposed an intelligent box office predictor based on aspect-level sentiment analysis of movie reviews. They used both the metadata of the movie and the sentiment information of the users' reviews to establish an intelligent predicting model. In the sentiment polarity classification model, a network-based aspect-level sentiment analysis strategy was developed by using the specific word embedding representations from both the contexts and the aspect. The sentiments from review texts, together with the movie information, were taken as input variables of the predictor (Yang et al., 2023).

In this project, sentiment analysis was applied to movie reviews to predict box office success, using various machine learning techniques, including data preprocessing, tokenization, and deep learning models, to analyse the sentiments expressed in text reviews compiled in a large dataset and used as a predictor for the movie's financial success, again, based on the assumption that the public's sentiment towards a movie expressed in reviews is indicative of the movie's popularity and box office performance.

The next section will explore how these techniques have been applied and discuss its exceptional contributions in various fields.

2.5 RELATED WORKS

Here are some of the key studies that have significantly contributed to the development and application of sentiment analysis techniques.

2.5.1 Sentiment Analysis Frameworks and Lexicons

Sentiment analysis frameworks and lexicons are valuable tools for extracting and analyzing sentiment from text data (Ligthart et al., 2014). Frameworks provide the theoretical and practical foundations for sentiment analysis, while lexicons are used to assign sentiment scores to words based on their connotations. For instance, a study by Alexander Ligthart, Cagatay Catal and Bedir Tekinerdogan provided a comprehensive overview of the key topics and different approaches for a variety of tasks in sentiment analysis. They mapped different features, algorithms, and datasets used in sentiment analysis models (Ligthart et al., 2014) it was particularly relevant to the use of sentiment analysis for predicting box office success, as it provides a framework for understanding how different techniques can be applied to this task.

They identified three main types of sentiment analysis frameworks:

i. **Rule-based frameworks**: These frameworks use hand-crafted rules to identify sentiment in text data. This approach is often used for tasks such as sentiment classification, where the goal is to assign a sentiment label to a piece of text (e.g., "positive", "negative", or "neutral").
ii. **Machine learning frameworks**: These frameworks use machine learning algorithms to learn sentiment patterns from data. This approach is often

used for tasks such as sentiment extraction, where the goal is to identify specific sentiment words or phrases (e.g., "great", "bad", "love", "hate").

iii. **Hybrid frameworks**: These frameworks combine rule-based and machine learning techniques to achieve better performance.

The authors also identified three main types of sentiment analysis lexicons:

iv. **Word-level lexicons**: These lexicons assign sentiment scores to individual words. This approach is often used for tasks such as sentiment classification.

v. **Sentence-level lexicons**: These lexicons assign sentiment scores to sentences. This approach is often used for tasks such as sentiment extraction.

vi. **Document-level lexicons**: These lexicons assign sentiment scores to documents. This approach is often used for tasks such as sentiment summarization.

Ligthart, Catal, and Tekinerdogan introducing the mapping of various features, algorithms, and datasets utilized in sentiment analysis models in the year 2014 as proved instrumental as this mapping serves as a valuable resource for researchers new to sentiment analysis, providing a solid starting point for their work.

2.5.2 Deep Learning in Sentiment Analysis

Sentiment analysis has undergone significant advancements with the emergence of deep learning (Ligthart et al., 2014). Deep learning algorithms, notably Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN), have found extensive application in sentiment analysis. As highlighted by Ligthart, Catal, and Tekinerdogan in the year 2014, while LSTM and CNN are the most frequently employed deep learning architectures in sentiment analysis.

The Long Short-Term Memory networks are particularly well-suited for sentiment analysis due to their ability to capture long-term dependencies in text data. This is vital for understanding the nuances of sentiments in complex sentences and paragraphs. Convolutional Neural Networks, on the other hand, excels at extracting local patterns

and features from text data. This property makes them effective in identifying sentiment-related words and phrases.

The integration of deep learning techniques revolutionized sentiment analysis, leading to improved accuracy and performance in sentiment classification, sentiment extraction, and sentiment summarization tasks. Deep learning algorithms have also enabled the development of more sophisticated sentiment analysis models that can handle a wider range of sentiment expressions and complexities.

In summary, (Ligthart et al., 2014) provided a comprehensive overview of the key topics and different approaches in sentiment analysis, with a particular focus on sentiment analysis frameworks, lexicons, and deep learning techniques. Their work has been instrumental in advancing the field of sentiment analysis and continues to be a valuable resource for researchers and practitioners.

2.5.3 Sentiment Analysis in Various Domains

1. Sentiment Analysis in Product Reviews by Xing Fang and Justin Zhan focused on sentiment polarity categorization in product reviews, their methodology utilized a dataset of 5.1 million product reviews from amazon.com, where they extracted subjective content for analysis, contrary to previous works and used a max-entropy POS tagger for word classification that identified phrases and employed classification models such as Naïve Bayesian, Random Forest, and Support Vector Machine. Though they faced the challenge of the curse of dimensionality in feature vector formation, the average F1 score was used to check its performance (Fang and Zhan, 2015)

2. Forecasting Price Shocks with Social Attention by authors Li Zhang, Liang Zhang, Keli Xhao, and Qi Liu had their objective to investigate the correlation between social media activities and stock price shocks in the Chinese Stock Market. The methodology used was the Degree of Social Attention (DSA) to capture stock price shocks and analyse social media features and user interactions to estimate the direction of price shocks. While using evaluates classifiers such as Naïve Bayes, Decision Tree, Random Forest, Logistic, and LibSVM to assess the performance.

The result was negative price shocks show better accuracy than positive shocks and Random Forest performs best among other classifiers according to their report (Zhang et al., 2018).

3. Efficient Adverse Drug Event Extraction Using Twitter by authors Yang Peng by Melody Moh, and Teng-Sheng Moh, while their main objective was to use social media to efficiently extract Adverse Drug Events (ADEs). They collected Twitter data over four months for the adverse drug events extraction. Proposing a pipeline for tweet capture, data pre-processing, drug classification, and sentiment analysis. Python was the natural language processing tool and Waikato Environment for Knowledge Analysis (WEKA) was used for the analysis. The results achieved an average 5 times the total number of adverse drug events, 20% being new adverse drug events (Peng et al., 2017).

2.5.4 Sentiment Analysis in Movie Reviews

Yang, Xu, and Tu in the year 2017 proposed an intelligent box office predictor based on aspect-level sentiment analysis of movie reviews. They utilized both the metadata of the movie and the sentiment information of users' reviews to establish an intelligent predicting model. The approach was aimed to capture the overall nuanced factors influencing box office performance, sentiment of movie reviews as well as specific sentiments related to different aspects of the movie, such as plot, acting, and directing (Yang et al., 2023).

The proposed approach involved the following steps:

i. **Data Collection**: The authors collected movie reviews from Douban, a Chinese social media platform for sharing reviews and ratings of movies, TV shows, books, and music.
ii. **Data Preprocessing**: The reviews were pre-processed by removing HTML tags, converting text to lowercase, and removing stop words.
iii. **Aspect Extraction**: Aspect extraction involves identifying and extracting relevant aspects mentioned in the reviews. The authors used a dictionary-based approach to extract aspects related to the movie's plot, acting, directing, and other relevant aspects.

iv. **Sentiment Analysis**: Sentiment analysis was performed on the extracted aspects to determine their sentiment polarity (positive, negative, or neutral). The authors employed the SentiWordNet lexicon to assign sentiment scores to aspects.

i. **Feature Engineering**: Features were extracted from the sentiment scores of the extracted aspects and the movie metadata. These features includes Average Sentiment score of each aspect, Sentiment Polarity dominance (positive, negative, or neutral), Correlation between aspect sentiment and box office, Model Training: A Support Vector Regression (SVR) model trained using the extracted features and the movie's box office revenue as the target variable, and the Model Evaluation: The performance of the SVR model evaluated using the mean squared error (MSE) and the root mean squared error (RMSE).

The results of this showed that the proposed approach significantly outperformed a baseline model that only used movie metadata. The SVR model achieved an MSE of 0.018 and an RMSE of 0.042, indicating its ability to accurately predict box office revenue based on sentiment analysis of movie reviews.

Overall, the study by Yang, Xu, and Tu highlighted the potential of sentiment analysis in predicting box office success and its value for the movie industry. The authors' approach combining aspect-level sentiment analysis with movie metadata and SVR regression demonstrates a promising method for harnessing sentiment information to improve box office prediction accuracy.

2.6 GAPS IDENTIFIED

The literature review has successfully provided a comprehensive overview of the field of sentiment analysis, its application in various domains, and its specific use in predicting box office performance based on movie reviews. It has also highlighted the importance of machine learning techniques, including deep learning, in sentiment analysis and the role of data preprocessing in preparing the data for analysis.

However, Despite the significant progress made in sentiment analysis, several gaps and opportunities remain for further exploration, for example:

Real-time sentiment analysis: Most of the existing research focuses on analysing historical data. There is a need for more research on real-time sentiment analysis, which can provide immediate insights into public sentiment.

Multilingual sentiment analysis: While a lot of work has been done on sentiment analysis in English, there is a lack of research on sentiment analysis in other languages. This is an important area for future research, given the global nature of the internet and social media.

Contextual sentiment analysis: Sentiment analysis can be highly context dependent. More research is needed on how to incorporate contextual information into sentiment analysis models.

Aspect-based sentiment analysis: Most of the current sentiment analysis approaches focus on the overall sentiment of a text. However, in many cases, different aspects of a product or a movie may evoke different sentiments. Aspect-based sentiment analysis is an emerging field that needs more attention.

2.7 SUMMARY

In conclusion, sentiment analysis is a powerful tool that can provide valuable insights into public opinion. When applied to movie reviews, it can serve as a reliable predictor of box office performance, thereby providing filmmakers and producers with a valuable tool for gauging public interest and predicting financial success. Your research contributes to this field and has the potential to further our understanding of the relationship between public sentiment and box office performance. As the field of sentiment analysis continues to evolve, there will undoubtedly be more opportunities for innovative research and application.

Now, having explored the existing literature and identified the gaps in the current research, the next chapter, “Methodologies and Models”, will dissect the specific methodologies and models used in this project to address these gaps and contribute to the field of sentiment analysis.

3. METHODOLOGIES AND MODELS

3.1 INTRODUCTION

In this study, sentiment analysis was employed on textual movie reviews with the aim of predicting box office performance. This involved the utilization of a range of machine learning techniques, encompassing data preprocessing, tokenization, application of the Long Short-Term Memory (LSTM) model and Python as the Natural Language Processing tool. These techniques expedited the analysis of sentiments within the reviews, which were subsequently used as predictors for the financial success of the movies.

This chapter offers an in-depth overview of the methodologies and models that were employed, serving as the cornerstone for subsequent analyses and insights. It explains the strategy adopted for data collection, the intricacies involved in data cleaning, and the application of exploratory data analysis and advanced inferential techniques to discern patterns within the movie dataset.

The visualizations and statistical measures derived from this study not only clarify the findings but also contribute to a solid foundation for decision-making within film production. They offer valuable insights into industry trends, thereby playing a pivotal role in shaping the future of film production. This unique approach underscores the potential of data-driven strategies in the film industry.

3.2 DATA COLLECTION

This marked the onset of doubts regarding the chosen topic, primarily due to the scarcity of freely available textual movie review datasets online. Even when a dataset was found, it often contained a limited number of observations, typically ranging from 10,000 to 50,000, and lacked the necessary details and columns required for the intended analysis. However, a suitable dataset was eventually discovered on Kaggle. This dataset, freely available, comprised 1,048,575 observations and included additional columns that were deemed essential for the process.

3.3 DATA LOADING

The movie dataset, which forms the foundation of this research, was carefully compiled from Rotten Tomatoes, a well-known platform that consolidates reviews and ratings for movies and television shows. This extensive dataset embodies over 5,000 movies in

1,048,575 observations, with each record encapsulating a plethora of information vital for the analyses. Each record represents a diverse range of cinematic works.

The initial step involved loading the data into a Dataframe using the pandas library, a robust data manipulation library in Python. The CSV file was read using the read_csv function, two-dimensional tabular data structure in pandas.

```
[30] # Loading the dataset from a CSV file located in Google Drive
     df = pd.read_csv('/content/drive/MyDrive/rotten_tomatoes_movie_reviews.csv')
```

Figure 3. Loading Original Dataset

3.3.1 Dataset Features

Revealed columns and their attributes:

- id: The unique identifier and title of each movie in the dataset.
- reviewId: The unique identifier for each review.
- creationDate: The date when the review was created.
- criticName: The name of the critic who reviewed the movie.
- isTopCritic: A boolean value indicating whether the critic is a top critic or not.
- originalScore: The original score given by the critic.
- reviewState: The state of the review, either 'fresh' or 'rotten'.
- publicationName: The name of the publication where the review was published.
- reviewText: The textual reviews of the movies and my most important and relevant column.
- scoreSentiment: The sentiment of the score, either 'POSITIVE' or 'NEGATIVE'.
- reviewUrl: The URL of the review.

3.4 DATA CLEANING

Data cleaning is a pivotal step in the data analysis process. It entails the preparation of data for analysis by eliminating or altering data that is incorrect, incomplete, irrelevant, duplicated, or improperly formatted. The importance of this process cannot be overstated, as the quality of the data and the utility of the resulting model hinge on the cleanliness of the data used.

Upon loading the data, data cleaning was performed to ready it for analysis. This included the removal of unnecessary columns and the appropriate handling of missing values. The ‘drop function’ was employed to eliminate the criticName, reviewId, publicationName, and reviewUrl columns, which were not required for the analysis. The ‘dropna function’ was utilized to eliminate rows with missing review texts, as demonstrated in the code below.

```
# Dropping unnecessary columns
df.drop(columns=['criticName','creationDate','reviewId', 'publicatioName', 'reviewUrl'], inplace=True)

# Dropping rows with missing review text
df = df.dropna(subset=['reviewText'])
```

Figure 4. Dataset Cleaning

3.5 TEXT PRE-PROCESSING

The first step of the pre-processing phase after cleaning the data involved converting all the text in the 'reviewText' column to lowercase to ensure uniformity in the text data and eliminating any discrepancies due to variations in letter casing using the str.lower function.

Next, non-alphanumeric characters, such as punctuation and symbols were removed from the 'reviewText' content using the str.replace function. This process is essential to focus the analysis on the core alphanumeric content, so that the model can better capture the essence of the reviews.

Following the lowercasing and removal of non-alphanumeric characters, the reviews were broken down into individual words or tokens using the word_tokenize function from the NLTK library. This granular approach enables the model to analyze and understand the sentiment associated with each distinct word in the reviews, contributing to a more nuanced understanding of the text data.

To refine the analysis further, common English stop words were removed from the tokenized reviews. Stop words, such as 'the,' 'and,' and 'is,' are generally devoid of specific sentiment and may introduce noise into the analysis. Removing these words allows the model to focus on terms that carry more significant sentiments thereby

enhancing the accuracy of the sentiment analysis results by the stopwords module from the NLTK library.

After the removal of stop words, the tokenized words are rejoined to reconstruct coherent review texts. This step is vital to preserve the original structure of the reviews while incorporating the modifications introduced during the preprocessing stages. The result is a refined, cleaned, and transformed 'reviewText' column, now better suited for meaningful sentiment analysis.

This process was carried out using the code block below:

```
# Converting review text to lowercase and removing non-alphanumeric characters
df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '').str.lower()

# Calculating and storing the length of each review
df['review_length'] = df['reviewText'].apply(len)

# Dropping rows with missing review text again after cleaning
df = df.dropna(subset=['reviewText'])

# Converting review text to lowercase
df['reviewText'] = df['reviewText'].str.lower()

# Removing non-alphanumeric characters from review text
df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '')

# Tokenizing the review text
df['reviewText'] = df['reviewText'].apply(word_tokenize)

# Defining the list of stop words
stop_words = set(stopwords.words('english'))

# Removing stop words from the tokenized review text
df['reviewText'] = df['reviewText'].apply(lambda x: [word for word in x if word not in stop_words])

# Joining the tokenized words back into a single string for each review
df['reviewText'] = df['reviewText'].apply(' '.join)
```

Figure 5. Text Preprocessing

3.6 LABEL ENCODING

The 'scoreSentiment' column was processed using LabelEncoder to transform the categorical labels into numerical values. This step is essential as it enables the machine learning model to interpret and interact with the target variable effectively.

```
# Encoding the sentiment scores using label encoding
encoder = LabelEncoder()
df['scoreSentiment'] = encoder.fit_transform(df['scoreSentiment'])
```

Figure 6. Label Encoding

3.7 EXPLORATORY DATA ANALYSIS (EDA)

Following the preprocessing of the data, an exploratory data analysis was undertaken. This involved determining the number of reviews and the average review length. Data visualization techniques, such as word clouds and frequency distribution plots, were employed. This facilitated an understanding of the most prevalent words in the reviews and their distribution. This stage is a crucial step aimed at deriving meaningful insights from the dataset and understanding the inherent attributes of the reviews. The comprehensive exploration contains the following detailed procedures:

3.7.1 Calculating the total number of reviews and the average length of reviews.

Determining the total number of reviews and the average length of reviews. In this preliminary step, the total number of reviews in the dataset is determined to provide a basic quantitative overview. Concurrently, the average length of reviews is computed, indicating the general verbosity of the reviews. These statistics serve as fundamental metrics for understanding the dataset.

```
# Printing the number of reviews and the average review length
num_reviews = df['reviewText'].count()
print(f"Number of reviews: {num_reviews}")
average_review_length = df['reviewText'].str.len().mean()
print(f"Average review length: {average_review_length}")
```

```
Number of reviews: 1375738
Average review length: 92.48325989396237
```

Figure 7. Reviews Length and Average

3.7.2 Generating a word cloud to visualize the most frequent words in the reviews.

This serves as a visual representation of the most frequently occurring words within the reviews. The size of each word in the cloud corresponds to its frequency, offering an intuitive perspective on the significant terms within the reviews.

```
# Plotting a word cloud
from wordcloud import WordCloud
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(str(df['reviewText']))
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.title('Word Cloud of Reviews')
plt.show()
```

Figure 8. Word Cloud

This graphical depiction is to identify prevalent themes or sentiments expressed in the dataset.

3.7.3 Creating a frequency distribution plot to analyse the distribution of words.

This step involves tokenizing the reviews into individual words and creating a frequency distribution plot. The plot illustrates the top 30 words, by plotting this analysis provides a concise overview of the dominant vocabulary within the reviews.

```
# Tokenizing the review text for the entire dataframe
tokens = word_tokenize(' '.join(df['reviewText']))

# Calculating word frequency for the entire dataframe
fdist = FreqDist(tokens)

# Plotting the frequency of the 30 most common words in the reviews
fdist.plot(30, cumulative=False)
plt.show()
```

Figure 9. Frequency Distribution

In this code snippet, the word_tokenize function was used to split the review texts into individual words and FreqDist function from the NLTK library to compute the frequency distribution of the words. The plot function is then used to plot the frequencies of the 30 most common words.

3.7.4 Plotting a Box Plot for Review Lengths

Box plots (or box-and-whisker plot) is used to show the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. The exploration was extended to visualizing the distribution of review lengths based on different review states. The box plot offers a graphical representation of the central tendency, dispersion, and potential outliers in review lengths.

```
# Plotting the box plot of review lengths by review state
sns.boxplot(x='reviewState', y='review_length', data=df)
plt.title('Box Plot of Review Lengths by Review State')
plt.show()
```

Figure 10. Box Plot of Review Length by Review State

The sns.boxplot is a function from the seaborn library, a Python data visualization library based on matplotlib, that creates a box plot.

In summary, this exploratory data analysis phase not only involves fundamental statistical calculations but also employs visualizations such as word clouds, frequency distribution plots, and box plots. These techniques collectively contribute to a comprehensive understanding of the dataset's characteristics and identify of patterns and trends within the reviews.

3.8 SENTIMENT ANALYSIS

This segment, involved the application of the Sentiment Analysis function using the NLTK's Vader SentimentIntensityAnalyzer function to assess and quantify the sentiment expressed in the reviews, in other words calculate the sentiment scores of the reviews, the sentiment scores were then used to label the reviews as positive or negative as shown in the code below:

```
# Calculate the sentiment score for each review
sia = SentimentIntensityAnalyzer()
df['sentiment_score'] = df['reviewText'].apply(lambda text: sia.polarity_scores(text)['compound'])

# Prepare the labels
labels = df['sentiment_score']
labels = labels.apply(lambda x: 1 if x > 0 else 0)
```

Figure 11. Sentiment Analysis

In the code block the polarity_scores method was used to calculate the sentiment scores of a text where the sentiment score is a measure of the positivity or negativity of the text while the compound score is a metric that calculates the sum of all the lexicon ratings which have been normalized between -1(most extreme negative) and +1 (most extreme positive).

3.9 MODEL TRAINING

This section involves the process of building, training, and evaluating the Long Short-Term Memory (LSTM) model for the sentiment analysis, integrating various components, data preparation, model construction, sentiment analysis, and training.

The Keras library was used to build a sequential model, this model consists of an embedding layer, two LSTM layers, and a dense layer, while the Adam optimizer and mean squared error was also used as the loss function. The model was trained on the pre-processed review texts and the sentiment labels with these code snippets below.

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(data, labels, test_size=0.2, random_state=1)
```

```
# Define the model
model = Sequential()
model.add(Embedding(input_dim=len(tokenizer.word_index)+1, output_dim=100))
model.add(LSTM(64, return_sequences=True))
model.add(LSTM(32))
model.add(Dense(1))

# Compile the model
model.compile(optimizer=Adam(), loss='mean_squared_error')

# Train the model
model.fit(X_train, y_train, epochs=10, validation_data=(X_test, y_test), batch_size=1024)
```

Figure 12. Model Training

As seen in the code snippet, the dataset is split into training and testing sets to assess the model's generalization performance. The training set is utilized to train the model, while the testing set evaluates its performance on unseen data. However, the model is trained on the pre-processed review texts and the corresponding sentiment labels. The training involves adjusting the model's weights to minimize the error between the predicted sentiment scores and the actual sentiment scores. The training process spans 10 epochs, allowing the model to iteratively learn patterns within the data.

3.10 MODEL EVALUATION

In this phase the trained Long Short-Term Memory (LSTM) model underwent thorough evaluation to assess its performance in sentiment analysis its performance was evaluated using various metrics, including accuracy, precision, recall, F1 score, and ROC-AUC. Also calculating the mean squared error between the predicted and actual sentiment scores to assess its ability to accurately predict box office revenue based on sentiment analysis of movie reviews.

The code block starts by importing essential metrics from the scikit-learn library to evaluate the model's performance like the Mean Squared Error which is a fundamental metric to quantify the average squared difference between the predicted sentiment scores and the actual sentiment scores in which a lower MSE value indicates a more accurate model, to transform the continuous sentiment scores into binary predictions, a threshold of 0.5 is applied in case the predicted sentiment score is above 0.5, the review is classified as positive (1); otherwise, it is labelled as negative (0).

The final step involved displaying the calculated metrics, providing a comprehensive summary of the model's performance in sentiment analysis.

Accuracy: represents the percentage of correctly classified reviews, providing an overall measure of the model's correctness.

Precision: measures the accuracy of positive predictions, representing the percentage of correctly classified positive reviews among all predicted positives.

Recall: assesses the model's ability to identify all positive instances, representing the percentage of correctly classified positive reviews among all actual positives.

F1 Score: is the harmonic mean of precision and recall, providing a balanced metric that considers both false positives and false negatives.

ROC-AUC Score: evaluates the model's ability to discriminate between positive and negative reviews across different threshold levels.

These metrics collectively offer valuable insights into the model's accuracy, precision, recall, F1 score, and its ability to distinguish between positive and negative sentiment.

```
# Calculate metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
f1 = f1_score(y_test, y_pred)
roc_auc = roc_auc_score(y_test, y_pred_probs)

print(f"Accuracy: {accuracy}")
print(f"Precision: {precision}")
print(f"Recall: {recall}")
print(f"F1 Score: {f1}")
print(f"ROC AUC: {roc_auc}")
```

Figure 13. Performance Metrics Calculation

```
# Predict the sentiment scores for the test set
y_pred = model.predict(X_test)

# Print the Mean Squared Error
print(f"Mean Squared Error: {mean_squared_error(y_test, y_pred)}")
```

```
8599/8599 [==============================] - 87s 10ms/step
Mean Squared Error: 0.01848002581470214
```

```
# Predict the probabilities for the test set
y_pred_probs = model.predict(X_test)
```

```
8599/8599 [==============================] - 86s 10ms/step
```

```
# Convert probabilities into class labels
y_pred = [1 if p > 0.5 else 0 for p in y_pred_probs]
```

Figure 14. Mean Square Error

3.11 SAVING TRAINED MODEL

In this final section of the training phase, an extremely crucial step of saving both the trained sentiment analysis model and the derived sentiment analysis results for future reference, so the 6 hours runtime for model training is drastically reduced if it is to be used to test future datasets. This process involved augmenting the original dataset with sentiment scores and labels, calculating average sentiment scores per movie, and storing the comprehensive results.

```
# Save the model
model.save('/content/drive/MyDrive/sentiment_analysis_model.h5')
```

Figure 15. Save Model

```
# Add sentiment analysis results to original dataframe
df_original['sentiment_score'] = df['sentiment_score']
df_original['sentiment_label'] = labels
```

```
# Group by movie and calculate the mean sentiment score
df_grouped = df_original.groupby('id')['sentiment_score'].mean()

# Convert the GroupBy object to a DataFrame
df_grouped = df_grouped.reset_index()

# Create a new column for box office success or failure
df_grouped['box_office'] = df_grouped['sentiment_score'].apply(lambda x: 'success' if x > 0 else 'failure')
```

```
#df_original.to_csv('sentiment_analysis_results.csv'

df_grouped.to_csv('/content/drive/MyDrive/average_sentiment_analysis_results.csv')
```

Figure 16. Grouping Sentiment Score and Saving new Dataset

As the code block implies, the trained sentiment analysis model is saved in the Hierarchical Data Format (HDF5) file format to ensure that the model with its learned weights and architecture can be reloaded and utilized for future sentiment analysis tasks without the need for retraining. Additionally, the original dataframe, df_original, is supplemented by incorporating the sentiment scores (sentiment_score) and binary sentiment labels (sentiment_label) derived from the sentiment analysis, then the augmented dataframe is then grouped by movie ID, and the mean sentiment score is calculated for each movie. This step aggregates sentiment scores to provide an average sentiment representation for each movie. Based on the average sentiment score for each movie, a binary 'box_office' label is assigned where movies with a positive average sentiment are labeled as 'success,' while those with a non-positive average sentiment are labeled as 'failure.'

The final step involves saving the comprehensive results, including movie IDs, average sentiment scores, and 'box_office' labels, to a CSV file. This file serves as a valuable record of the sentiment analysis outcomes and can be utilized for further analysis or reporting.

3.12 SUMMARY

This chapter has captured the entire lifecycle of the project, from data loading to model training, evaluation, and ultimately saving the trained model and sentiment analysis results. The saved model can be readily applied to new data, and the results provide a tangible representation of sentiment across movies in the dataset. The subsequent chapter will delve into the results and findings derived from the project.

4. RESULTS AND DISCUSSION

4.1 INTRODUCTION

In this chapter, will present results obtained from the sentiment analysis and box office prediction performed on the movie reviews dataset. These results are derived from the methodologies and models discussed in methodologies and model chapter, which involved the data preprocessing, exploratory data analysis, sentiment analysis, and model training. It is structured it in a way to first present the results of the sentiment analysis, followed by the results of the box office prediction. Each section will provide a detailed account of the findings, supported by relevant statistics, visualizations, and code outputs.

4.2 EXPLORATORY DATA ANALYSIS (EDA) RESULTS

4.2.1 Word Cloud

The word cloud is a visual representation of text data where the size of each word indicates its frequency or importance, it provides a quick visual representation of the most common words in the movie reviews. Words that appear larger are used more frequently, and by examining these words, we can get a sense of the themes or topics that are often mentioned in the reviews.

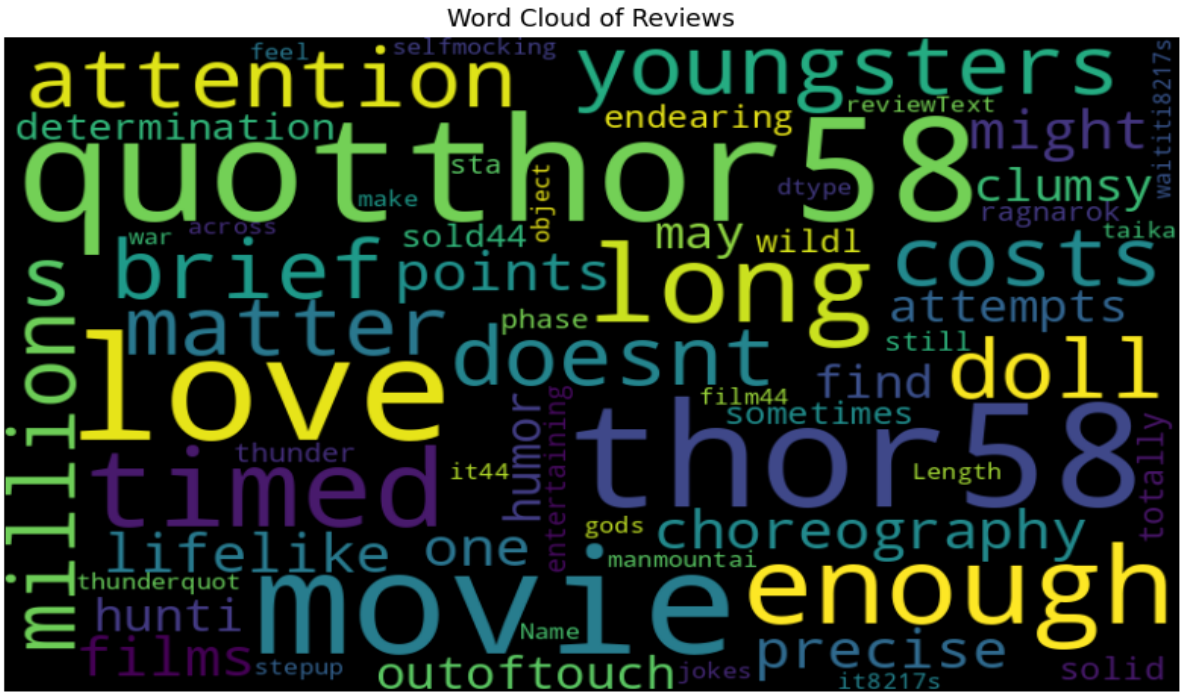


Figure 17. Word Cloud EDA

The size of each word reflects its frequency, with larger words appearing more often. The placement of words can sometimes be informative, although it's mostly random in this case. In this case, stop words like "the," "is," and "a" have been removed before generating the word cloud to identify the true value of each review individually.

According to the plot, words like "love," "enjoyed," "good," and "fun" are prominent, suggesting that the overall sentiment of the reviews is positive meanwhile the movie elements are "movie," "film," "story," "characters," and "acting" are frequently used, indicating that these are key aspects that reviewers focus on. The word cloud also depicts that mentions of "Thor," "Ragnarok," "attention," and "youngsters hint that the dataset might be related to the movie "Thor: The Dark World" or superhero movies in general.

4.2.2 Frequency Distribution Plot

The second part of the Exploratory Data Analysis involved creating a frequency distribution plot. This frequency distribution plot provides a visual representation of the most common words in the movie reviews. Words that appear more frequently are used more often in the reviews, and by examining these words, we can get a sense of the themes or topics that are often mentioned in the reviews.

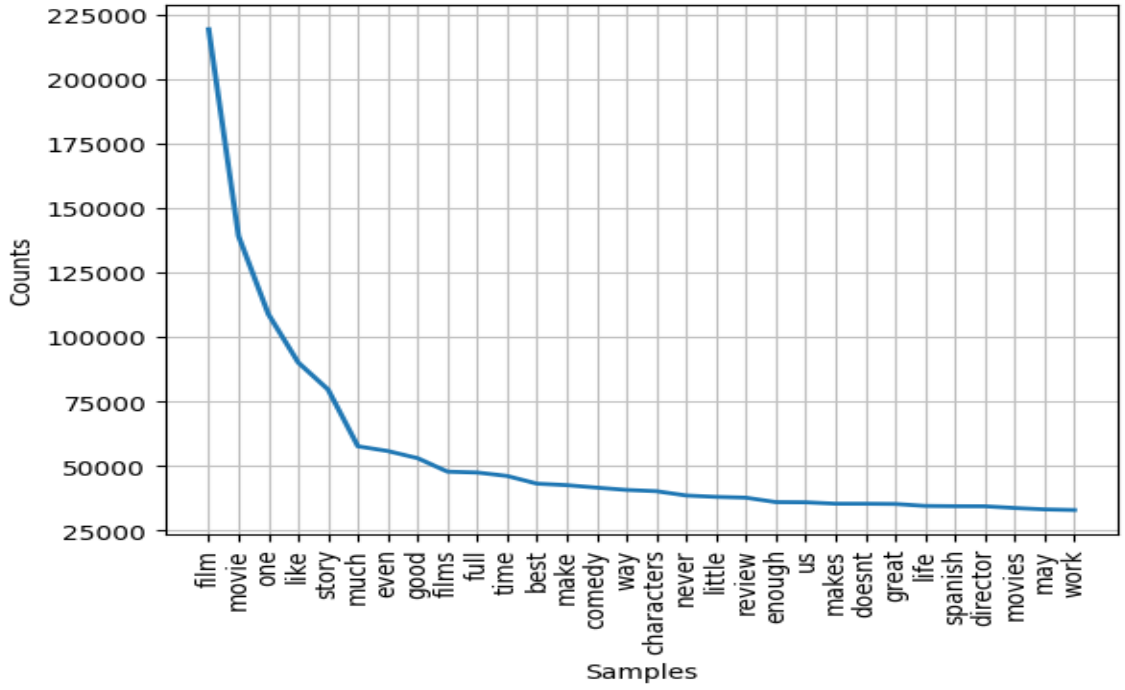


Figure 18. Frequency Distribution Plot

The plot depicts a graph in steep decline in counts as it moves along different sample categories. The samples axis includes terms like “film,” “movie,” “story,” “much,” “good time,” etc., indicating categories or tags associated with movies or media content. The line starts at its peak at around 225,000 counts for “film” and decreases sharply, flattening out towards “work.”

4.2.3 Box Plot of Review Lengths

The third part of the Exploratory Data Analysis, I decided to create a box plot of review lengths by review state. Usually, Box plot (or box-and-whisker plot) shows the distribution of quantitative data in a way that facilitates comparisons between variables or across levels of a categorical variable. In this case, it’s used to create a box plot of review lengths by review state.

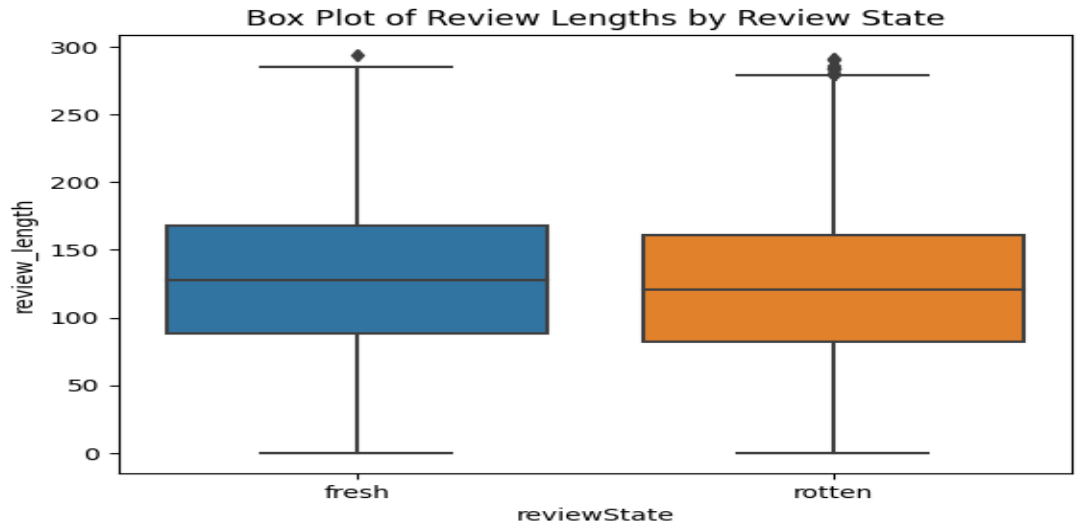


Figure 19. Box Plot of Review Lengths

The plot shows two box plots side by side, labeled “fresh” and “rotten” on the x-axis under the category “reviewState”. The y-axis is labeled as “reviewLength” and ranges from 0 to 300. The “fresh” box plot is blue, with the box representing review lengths roughly between 100 and 150. The “rotten” box plot is orange, with the box representing review lengths roughly between 125 and 175. There’s an outlier indicated in the “fresh” category at a review length of approximately 250.

The box plot provides a visual representation of the distribution of review lengths for each review state. This can help to understand if there is a significant difference in review lengths between the “fresh” and “rotten” categories.

4.2.4 Histogram of Sentiment Scores

The fourth part of my Exploratory Data Analysis involved creating a histogram of sentiment scores. Histogram is a graphical representation that organizes a group of data points into a specified range. In this case, it's used to create a histogram of sentiment scores.

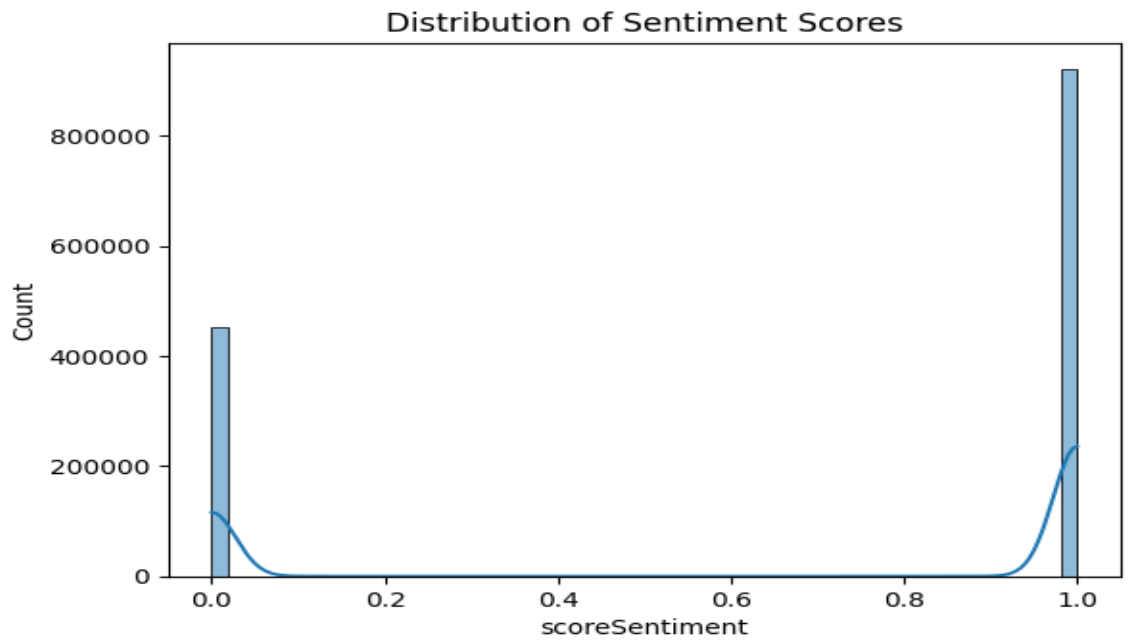


Figure 20. Histogram of Sentiment Scores

The graph shows a distribution of sentiment scores ranging from 0.0 to 1.0 on the x-axis. The y-axis represents the count, scaling up to 800,000. There are two prominent bars in the graph: one near the 0.2 mark and another at the 1.0 mark on the scoreSentiment axis. The bar near the 0.2 mark is shorter, with a count just above 400,000. The bar at the 1.0 mark is taller, reaching the maximum count on the y-axis.

This histogram provides a visual representation of the distribution of sentiment scores in the movie reviews. This helps to understand the range and frequency of sentiment scores in the reviews.

4.2.5 Word Frequency for Positive Distribution

The next part of the Exploratory Data Analysis plotted is a word frequency distribution for positive reviews. This graph provides a visual representation of the frequency of different words in the positive movie reviews.

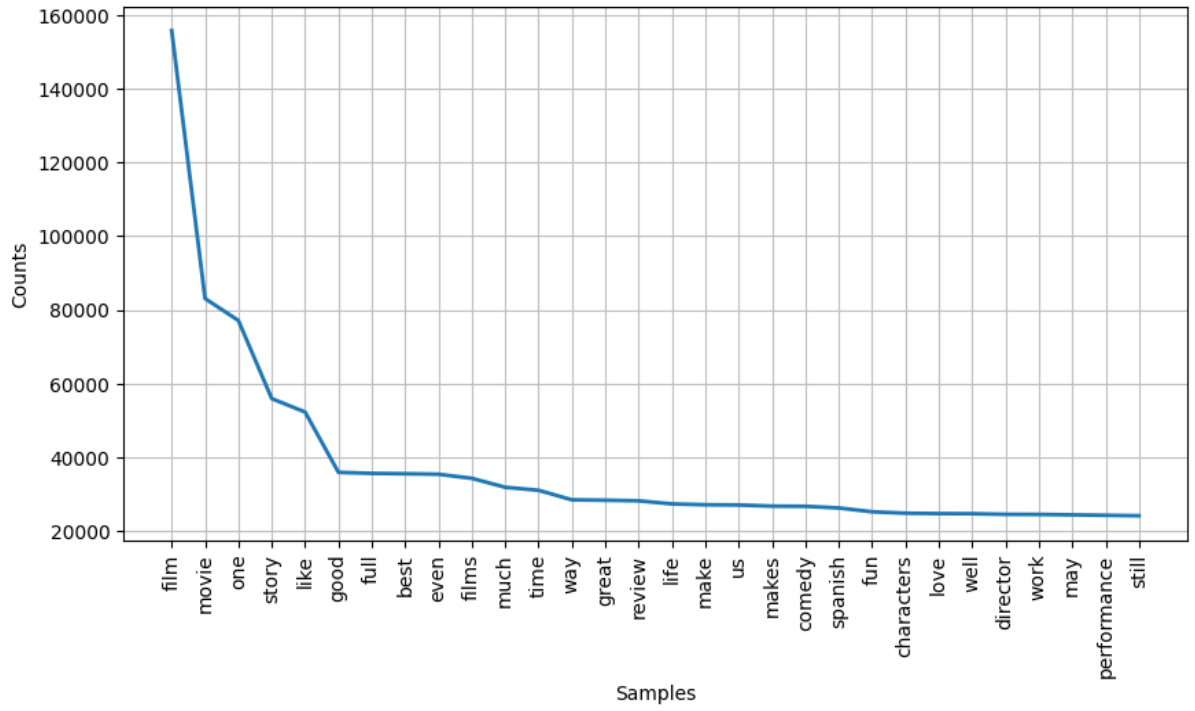


Figure 21. Word Frequency for Positive Distribution

The graph shows a distribution of word frequencies ranging from 0 to 160,000 on the y-axis. The x-axis represents various words like “film,” “movie,” “story,” and others that are presumably sampled from some text or set of texts. The line on the graph starts at a high count with the word “film” and decreases sharply as it moves to other words. It flattens out towards the end indicating a lower frequency of those words. Words like “film” and “movie” have significantly higher counts compared to others like “director,” “performance,” indicating they are more common in the sampled texts.

The word frequency distribution provides a quick visual representation of the most common words in the positive movie reviews. Words that appear more frequently are used more often in the reviews, and by examining these words, to understand the themes or topics that are often mentioned in the positive reviews.

4.2.6 Word Frequency for Negative Distribution

The sixth part of my Exploratory Data Analysis like the plot for word frequency for positive distribution this one involved creating a word frequency distribution for negative reviews. This graph provides a visual representation of the frequency of different words in the negative movie reviews.

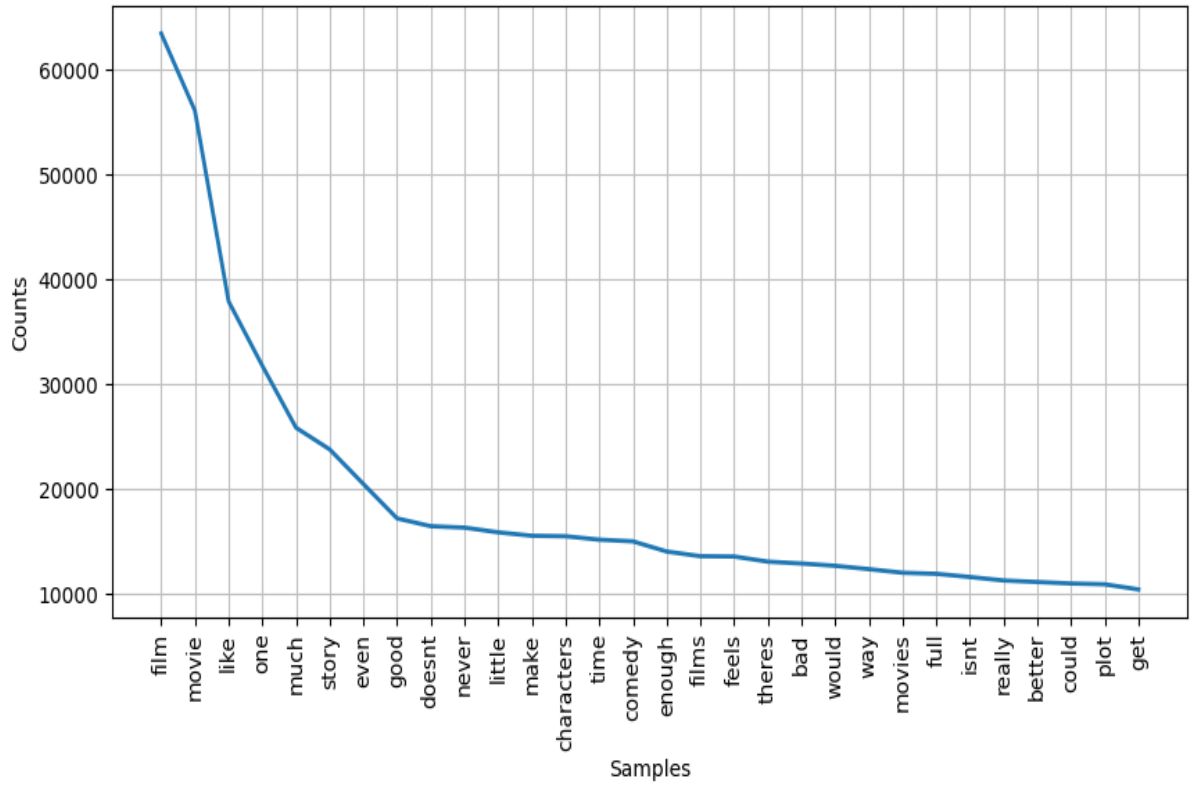


Figure 22. Word Frequency for Negative Distribution

The graph shows a distribution of word frequencies ranging from 0 to 60,000 on the y-axis. The x-axis represents various words like “film,” “movie,” “one,” “story,” and others that are presumably sampled from some text or set of texts. The line on the graph starts at a high count with the word “film” and decreases as it moves to other words. It flattens out towards the end indicating a lower frequency of those words.

The word frequency distribution provides a quick visual representation of the most common words in the negative movie reviews. Words that appear more frequently are used more often in the reviews, and by examining these words, to get a sense of the themes or topics that are often mentioned in the negative reviews.

4.2.7 Box Plot of Review Lengths by Sentiment Score

The seventh plot of the Exploratory Data Analysis involved creating a box plot of review lengths by sentiment score. In this case, it’s used to create a box plot of review lengths by sentiment score.

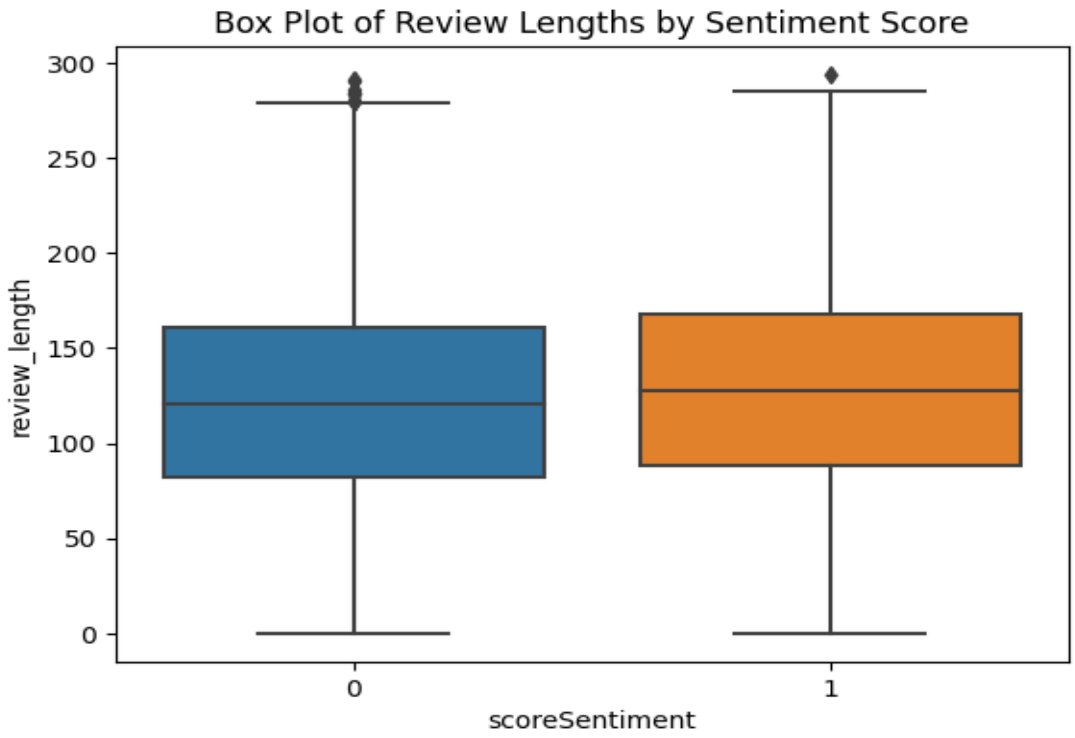


Figure 23. Box Plot of Review Lengths by Sentiment Score

The plot shows two boxes representing two different sentiment scores, 0 and 1. The y-axis is labelled "Review Length" and ranges from 0 to 300. The x-axis is labelled "scoreSentiment" with two categories, 0 and 1. The blue box (sentiment score 0) has a median review length around 150, with the interquartile range extending approximately from just below 100 to just above 200. There's an outlier indicated above the upper whisker. The orange box (sentiment score 1) has a median review length slightly higher than that of sentiment score 0, with its interquartile range being narrower and situated higher on the Review Length axis. An outlier is also indicated above its upper whisker.

This helps to understand if there is a significant difference in review lengths between the different sentiment scores.

4.2.8 Correlation Matrix

The table shows the correlation coefficients between variables. Each cell in the table shows the correlation between two variables. The value is in the range of -1 to 1. If two variables have high correlation, it means they tend to increase or decrease together. If the correlation is low, the two variables are not linearly associated.

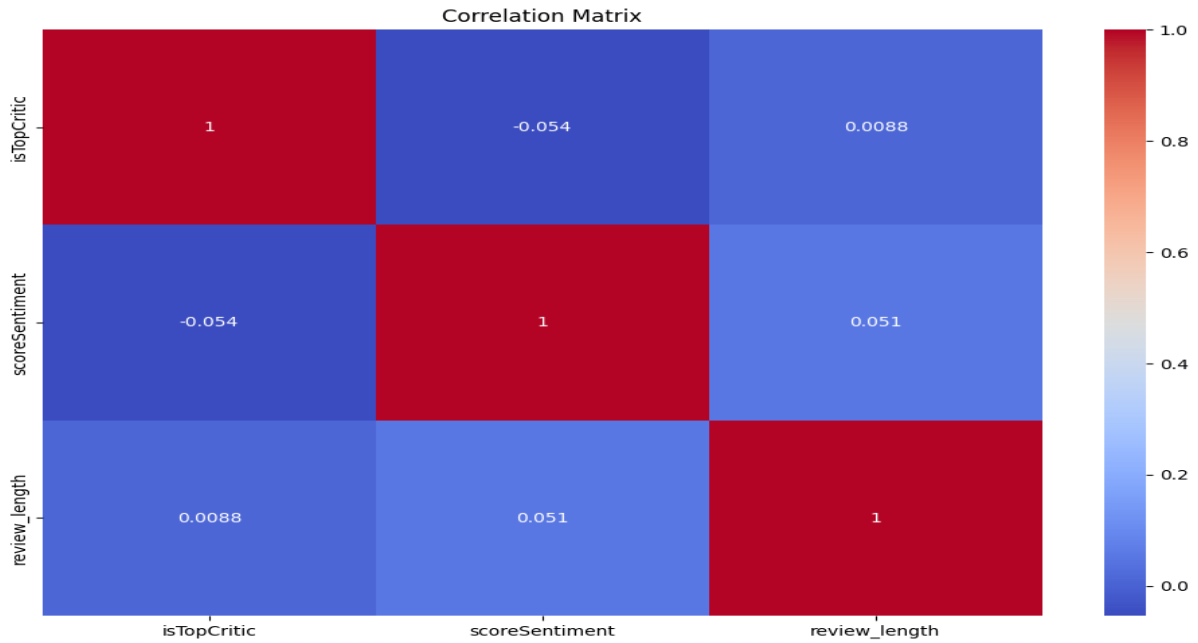


Figure 24. Correlation Matrix

In the matrix, there are three variables: 'isTopCritic', 'scoreSentiment', and 'review_length' and here the correlations depicts:

- **'isTopCritic' and 'scoreSentiment'**: The correlation is -0.054, which is close to 0. This suggests there's no linear relationship between whether the critic is a top critic and the sentiment score of the review.
- **'isTopCritic' and 'review_length'**: The correlation is 0.0088, which is also close to 0. This suggests there's no linear relationship between whether the critic is a top critic and the length of the review.
- **'scoreSentiment' and 'review_length'**: The correlation is 0.051, which is close to 0. This suggests there's no linear relationship between the sentiment score of the review and the length of the review.

4.2.9 Text-specific EDA: Word Frequency

The next part of the EDA was a text-specific analysis that produced word frequencies. This analysis provides a count of how often each word appears in the text data. The max_features parameter is set to 1000, which means it only considers the top 1000 terms ordered by term frequency.

'long': 20407, 'enough': 36067, 'attention': 6092, 'plenty': 7431, 'interesting': 15823, 'land': 2839, 'doesn't': 35426', matter': 7564, 'movies': 139192, 'good': 53091, 'bad': 20419, 'points': 4106, 'one': 108887, 'might': 24719, 'wonder': 5639, 'whether': 5792, 'footage': 4336, 'used': 3811, 'quality': 5076, 'due': 2908, 'love': 31324, 'films': 47860, 'attempts': 2970, 'humor': 14326, 'may': 33244, 'find': 15939, 'reason': 5625, 'franchise': 7188, 'first': 27507, 'place': 11575, 'sometimes': 11436, 'entertaining':

22828, ‘every’: 22305, ‘new’: 30164, ‘there’s’: 31427, ‘another’: 18958, ‘that’s’: 28904, ‘amazing’: 3905‘, feels’: 26948, ‘like’: 90167, ‘true’: 12650, ‘flick’: 7214, ‘absolutely’: 4498, ‘special’: 8394 , ‘effects’: 8858, ‘men’: 7940, ‘rare’: 6570, ‘star’: 11909, ‘could’: 27720, ‘make’: 42632, ‘perhaps’: 9041, ‘you’re’: 12876, ‘fan’: 4142, ‘cinema’: 12903, ‘even’: 55858, ‘fun’: 31809, ‘best’: 43208, ‘pictures’: 2842, ‘production’: 7686, ‘years’: 17376, ‘big’: 17996, ‘screen’: 15768, ‘seen’: 15383, ‘viewers’: 9699, ‘others’: 5022, ‘mix’: 4437, ‘action’: 30318, ‘sequences’: 6009, ‘often’: 19599, ‘ridiculous’: 3382…

This word frequency analysis is a crucial part of text data exploration as it provides insights into the most common words in the dataset to:

- Understanding the Data, in other words, the Word frequency analysis helps to understand the main themes or topics in the text data. For instance, in movie reviews, frequent words might include ‘plot’, ‘character’, ‘director’, which are all relevant to the domain.
- Feature Selection: In text data, not all words are equally important. Some words might occur very frequently but offer little predictive power (like ‘the’, ‘is’, ‘and’). Word frequency analysis can help identify such words, and we might choose to exclude them from the analysis (a process called ‘stop word removal’).
- Insight into Sentiments: By looking at the most frequent words in positive and negative reviews separately, this enables to gain insights into what contributes to positive or negative sentiments. For example, if ‘exciting’ is frequent in positive reviews and ‘boring’ in negative ones, it gives us a sense of what drives these sentiments.
- Visualizing Results: Word frequency distributions are also a great way to visualize the data, making it easier to understand the text data’s main characteristics, which is an essential step before moving on to more complex analyses or applying machine learning algorithms.

4.2.10 Text-specific EDA: Word Frequency (Bigrams)

The CountVectorizer function from sklearn was used to transform the text data into a sparse matrix of token counts. This is a common technique in natural language processing like the one used in the last visualization but in this the ngram_range parameter is set to (2, 2), which means it’s looking at pairs of adjacent words (also known as bigrams). The max_features parameter is also set to 1000, to retrieve the top 1000 terms ordered by term frequency. Sample of the results displayed:

‘long enough’: 437, ‘doesnt matter’: 379, ‘good bad’: 630, ‘one might’: 853, ‘may find’: 1112, ‘first place’: 1035, ‘feels like’: 10873, ‘special effects’: 3612, ‘make movie’: 1044, ‘you’re fan’: 522, ‘big screen’: 2558, ‘action sequences’: 2235, ‘bad movie’: 1233, ‘movie one’: 1095, ‘familiar story’: 454, ‘almost every’: 814, ‘heart right’: 617, ‘kind movie’: 1490, ‘soap opera’: 1202, ‘talented cast’: 801, ‘coen brothers’: 708, ‘quentin tarantino’: 402, ‘twists turns’: 1040, ‘many characters’: 389, ‘one films’: 964, ‘american dream’: 570, ‘full review’: 34214, ‘review spanish’: 33666, ‘film offers’: 802, ‘main character’: 730, ‘better film’: 636, ‘film made’: 971, ‘film shows’: 537, ‘lot like’: 731, ‘take seriously’: 695, ‘first half’: 1365, ‘second half’: 1197, ‘musical numbers’: 672, ‘built around’: 427, ‘central performance’: 904, ‘action scenes’: 1422, ‘whole thing’: 1563, ‘hard watch’: 376, ‘film noir’: 1171, ‘cinematic universe’: 426, ‘may well’: 1054, ‘visually stunning’: 802, ‘dont need’: 599, ‘equal measure’: 648, ‘doesnt much’: 758, ‘love story’: 3589, ‘history lesson’: 730, ‘doesnt work’: 831, ‘past present’: 507, ‘end result’: 860, ‘along way’: 1667, ‘come away’: 481, ‘moving film’: 553, ‘film may’: 1074, ‘im sure’: 1127, ‘first feature’: 1125, ‘dont know’: 1431, ‘visual style’: 882, ‘make sense’: 878, ‘recent memory’: 842, ‘united states’: 560, ‘film tries’: 381, ‘tries hard’: 677, ‘much like’: 2043, ‘comes across’: 1276, ‘every turn’: 684, ‘come across’: 510, ‘piece work’: 1724, ‘would make’: 837, ‘whole family’: 573, ‘rest film’: 434, ‘pretty much’: 1900, ‘black comedy’: 1375

4.2.11 Distribution of sentiment scores

The histogram is a type of plot that allows to visualize the underlying frequency distribution (shape) of a set of continuous or discrete data. In this case, the data is the sentiment scores of movie reviews. The sentiment scores range is from -1 to 1, where score of -1 represents a very negative sentiment, score of 0 represents a neutral sentiment, and score of 1 represents a very positive sentiment.

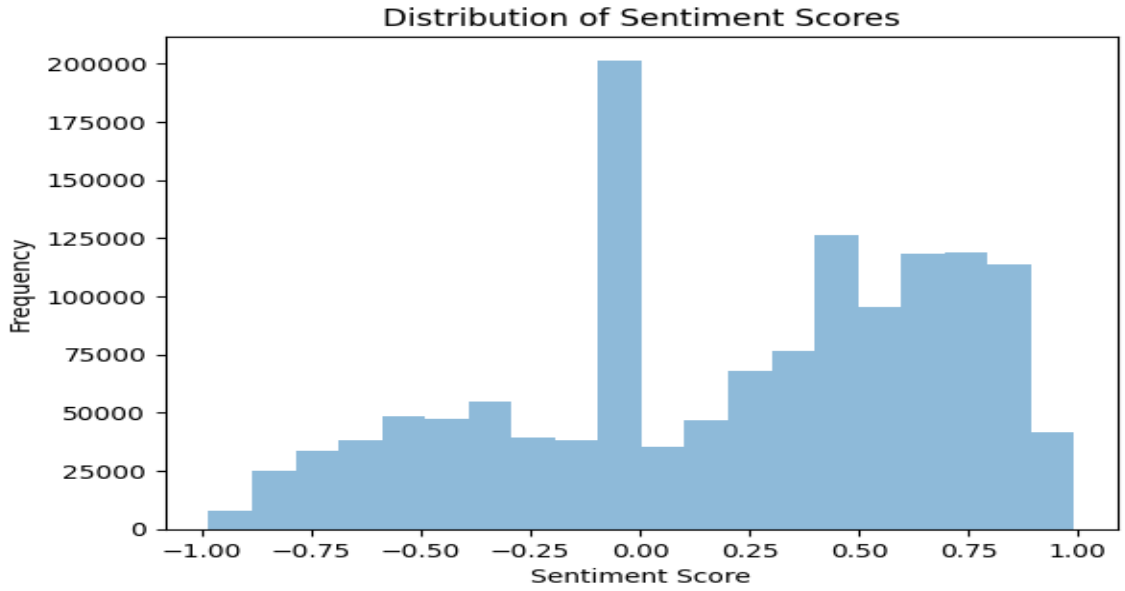


Figure 25. Sentiment Score Distribution

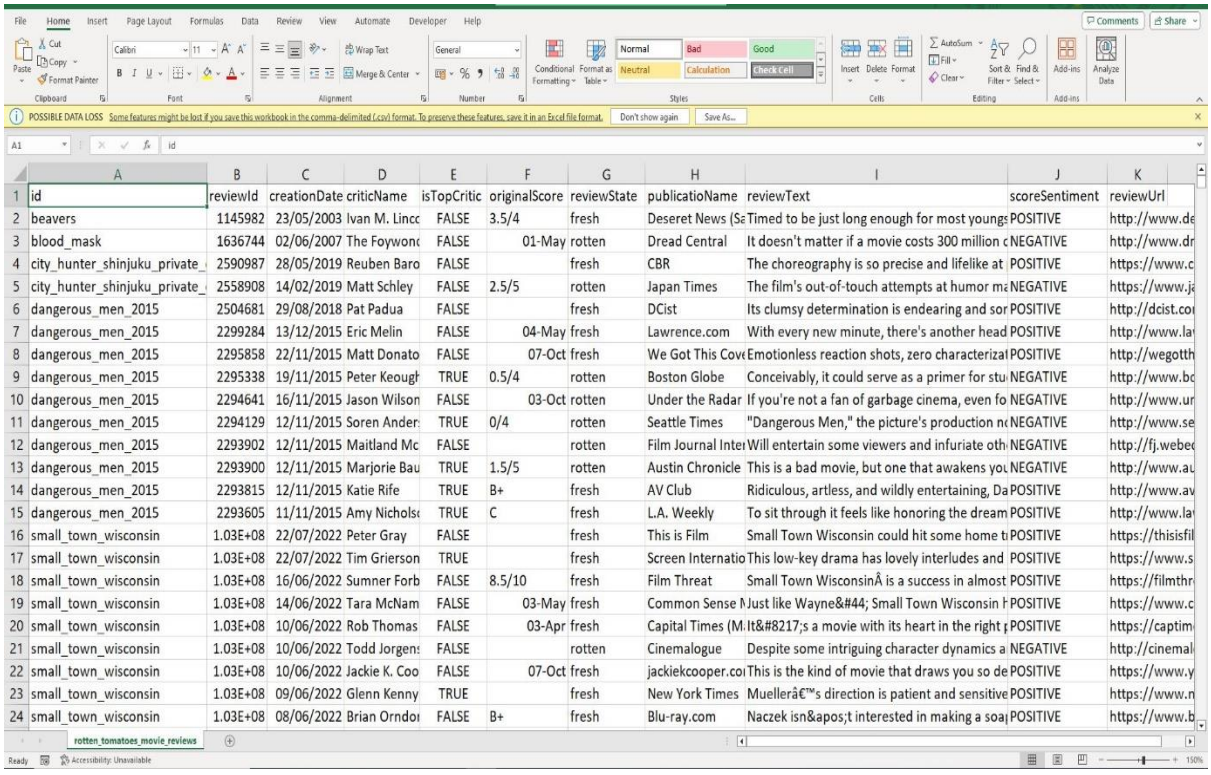
The histogram plot shows two prominent peaks: Peak at 0: There's a high bar at the 0 mark on the x-axis. This means that a significant number of reviews have a neutral sentiment. Peak at 0.50: There's another peak around the 0.50 mark and two identical peaks around the 0.75 mark. This means that a considerable number of reviews have a positive sentiment.

Creating this plot is to understand the distribution of sentiments in the reviews. By looking at the histogram, we can easily see whether the reviews are generally positive, negative, or neutral, and how strong these sentiments are.

4.3 DATASETS FEATURES AND SENTIMENT ANALYSIS RESULTS

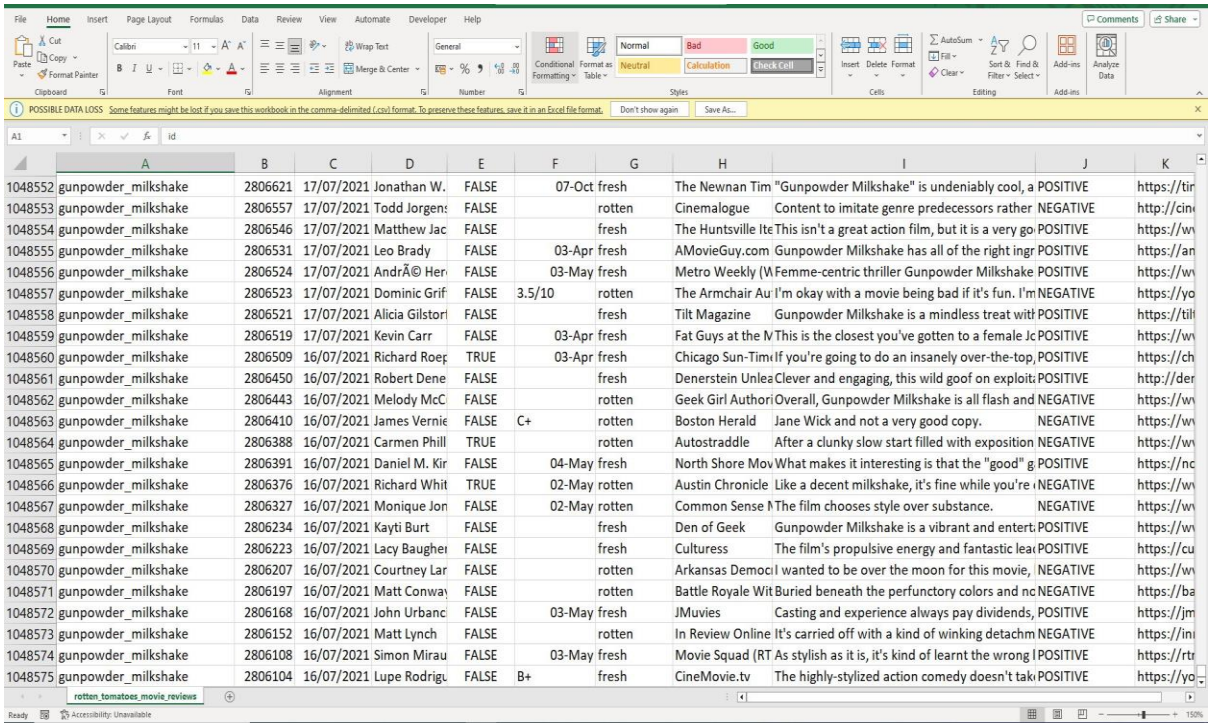
4.3.1 ORIGINAL DATASET FEATURES

The dataset I used in this project is a collection of movie reviews containing 1,444,963 reviews, spanning from January 1, 1800, to April 8, 2023, with text reviews written by 15,510 unique critics and published in 2,707 unique publications.



	A	B	C	D	E	F	G	H	I	J	K
1	id	reviewId	creationDate	criticName	isTopCritic	originalScore	reviewState	publicatioName	reviewText	scoreSentiment	reviewUrl
2	beavers	1145982	23/05/2003	Ivan M. Lincc	FALSE	3.5/4	fresh	Deseret News (Sa	Timed to be just long enough for most youngs	POSITIVE	http://www.de
3	blood_mask	1636744	02/06/2007	The Foywonc	FALSE	01-May	rotten	Dread Central	It doesn't matter if a movie costs 300 million c	NEGATIVE	http://www.dr
4	city_hunter_shinjuku_private_	2590987	28/05/2019	Reuben Baro	FALSE		fresh	CBR	The choreography is so precise and lifelike at	POSITIVE	https://www.c
5	city_hunter_shinjuku_private_	2558908	14/02/2019	Matt Schley	FALSE	2.5/5	rotten	Japan Times	The film's out-of-touch attempts at humor ma	NEGATIVE	https://www.ja
6	dangerous_men_2015	2504681	29/08/2018	Pat Padua	FALSE		fresh	DCist	Its clumsy determination is endearing and sor	POSITIVE	http://dcist.co
7	dangerous_men_2015	2299284	13/12/2015	Eric Melin	FALSE	04-May	fresh	Lawrence.com	With every new minute, there's another head	POSITIVE	http://www.la
8	dangerous_men_2015	2295858	22/11/2015	Matt Donato	FALSE	07-Oct	fresh	We Got This Cov	Emotionless reaction shots, zero characteriza	POSITIVE	http://wegotth
9	dangerous_men_2015	2295338	19/11/2015	Peter Keough	TRUE	0.5/4	rotten	Boston Globe	Conceivably, it could serve as a primer for stu	NEGATIVE	http://www.bo
10	dangerous_men_2015	2294641	16/11/2015	Jason Wilson	FALSE	03-Oct	rotten	Under the Radar	If you're not a fan of garbage cinema, even fo	NEGATIVE	http://www.ur
11	dangerous_men_2015	2294129	12/11/2015	Soren Ander:	TRUE	0/4	rotten	Seattle Times	"Dangerous Men," the picture's production n	NEGATIVE	http://www.se
12	dangerous_men_2015	2293902	12/11/2015	Maitland Mc	FALSE		rotten	Film Journal Inter	Will entertain some viewers and infuriate oth	NEGATIVE	http://fj.webe
13	dangerous_men_2015	2293900	12/11/2015	Marjorie Bau	TRUE	1.5/5	rotten	Austin Chronicle	This is a bad movie, but one that awakens you	NEGATIVE	http://www.au
14	dangerous_men_2015	2293815	12/11/2015	Katie Rife	TRUE	B+	fresh	AV Club	Ridiculous, artless, and wildly entertaining, Da	POSITIVE	http://www.av
15	dangerous_men_2015	2293605	11/11/2015	Amy Nicholsc	TRUE	C	fresh	L.A. Weekly	To sit through it feels like honoring the dream	POSITIVE	http://www.la
16	small_town_wisconsin	1.03E+08	22/07/2022	Peter Gray	FALSE		fresh	This is Film	Small Town Wisconsin could hit some home t	POSITIVE	https://thisisfil
17	small_town_wisconsin	1.03E+08	22/07/2022	Tim Grierson	TRUE		fresh	Screen Internatio	This low-key drama has lovely interludes and	POSITIVE	https://www.s
18	small_town_wisconsin	1.03E+08	16/06/2022	Sumner Forb	FALSE	8.5/10	fresh	Film Threat	Small Town WisconsinÂ is a success in almost	POSITIVE	https://filmthr
19	small_town_wisconsin	1.03E+08	14/06/2022	Tara McNam	FALSE	03-May	fresh	Common Sense M	Just like Wayne, Small Town Wisconsin h	POSITIVE	https://www.c
20	small_town_wisconsin	1.03E+08	10/06/2022	Rob Thomas	FALSE	03-Apr	fresh	Capital Times (M	It’s a movie with its heart in the right p	POSITIVE	https://captim
21	small_town_wisconsin	1.03E+08	10/06/2022	Todd Jorgens	FALSE		rotten	Cinemalogue	Despite some intriguing character dynamics a	NEGATIVE	http://cinemal
22	small_town_wisconsin	1.03E+08	10/06/2022	Jackie K. Coo	FALSE	07-Oct	fresh	jackiekcooper.co	This is the kind of movie that draws you so de	POSITIVE	https://www.y
23	small_town_wisconsin	1.03E+08	09/06/2022	Glenn Kenny	TRUE		fresh	New York Times	Muellerâ€™s direction is patient and sensitive	POSITIVE	https://www.n
24	small_town_wisconsin	1.03E+08	08/06/2022	Brian Orndor	FALSE	B+	fresh	Blu-ray.com	Naczek isn't interested in making a soa	POSITIVE	https://www.b

Figure 26. DATASET SAMPLE (TOP)



	A	B	C	D	E	F	G	H	I	J	K
1048552	gunpowder_milkshake	2806621	17/07/2021	Jonathan W.	FALSE	07-Oct	fresh	The Newnan Tim	"Gunpowder Milkshake" is undeniably cool, a	POSITIVE	https://tir
1048553	gunpowder_milkshake	2806557	17/07/2021	Todd Jorgens	FALSE		rotten	Cinemalogue	Content to imitate genre predecessors rather	NEGATIVE	http://cin
1048554	gunpowder_milkshake	2806546	17/07/2021	Matthew Jac	FALSE		fresh	The Huntsville Ite	This isn't a great action film, but it is a very go	POSITIVE	https://w
1048555	gunpowder_milkshake	2806531	17/07/2021	Leo Brady	FALSE	03-Apr	fresh	AMovieGuy.com	Gunpowder Milkshake has all of the right ingr	POSITIVE	https://an
1048556	gunpowder_milkshake	2806524	17/07/2021	AndrÃ© Her	FALSE	03-May	fresh	Metro Weekly (W	Femme-centric thriller Gunpowder Milkshake	POSITIVE	https://w
1048557	gunpowder_milkshake	2806523	17/07/2021	Dominic Grif	FALSE	3.5/10	rotten	The Armchair Au	I'm okay with a movie being bad if it's fun. I'm	NEGATIVE	https://yo
1048558	gunpowder_milkshake	2806521	17/07/2021	Alicia Gilstorf	FALSE		fresh	Tilt Magazine	Gunpowder Milkshake is a mindless treat with	POSITIVE	https://tilt
1048559	gunpowder_milkshake	2806519	17/07/2021	Kevin Carr	FALSE	03-Apr	fresh	Fat Guys at the M	This is the closest you've gotten to a female Jc	POSITIVE	https://w
1048560	gunpowder_milkshake	2806509	16/07/2021	Richard Roep	TRUE	03-Apr	fresh	Chicago Sun-Time	If you're going to do an insanely over-the-top,	POSITIVE	https://ch
1048561	gunpowder_milkshake	2806450	16/07/2021	Robert Dene	FALSE		fresh	Denerstein Unlea	Clever and engaging, this wild goof on exploit	POSITIVE	http://der
1048562	gunpowder_milkshake	2806443	16/07/2021	Melody McC	FALSE		rotten	Geek Girl Authori	Overall, Gunpowder Milkshake is all flash and	NEGATIVE	https://w
1048563	gunpowder_milkshake	2806410	16/07/2021	James Vernie	FALSE	C+	rotten	Boston Herald	Jane Wick and not a very good copy.	NEGATIVE	https://w
1048564	gunpowder_milkshake	2806388	16/07/2021	Carmen Phill	TRUE		rotten	Autostraddle	After a clunky slow start filled with exposition	NEGATIVE	https://w
1048565	gunpowder_milkshake	2806391	16/07/2021	Daniel M. Kir	FALSE	04-May	fresh	North Shore Mov	What makes it interesting is that the "good" g	POSITIVE	https://nc
1048566	gunpowder_milkshake	2806376	16/07/2021	Richard Whit	TRUE	02-May	rotten	Austin Chronicle	Like a decent milkshake, it's fine while you're	NEGATIVE	https://w
1048567	gunpowder_milkshake	2806327	16/07/2021	Monique Jon	FALSE	02-May	rotten	Common Sense M	The film chooses style over substance.	NEGATIVE	https://w
1048568	gunpowder_milkshake	2806234	16/07/2021	Kayti Burt	FALSE		fresh	Den of Geek	Gunpowder Milkshake is a vibrant and entert	POSITIVE	https://w
1048569	gunpowder_milkshake	2806223	16/07/2021	Lacy Baugher	FALSE		fresh	Culturess	The film's propulsive energy and fantastic lea	POSITIVE	https://cu
1048570	gunpowder_milkshake	2806207	16/07/2021	Courtney Lar	FALSE		rotten	Arkansas Democ	I wanted to be over the moon for this movie,	NEGATIVE	https://w
1048571	gunpowder_milkshake	2806197	16/07/2021	Matt Conway	FALSE		rotten	Battle Royale Wit	Buried beneath the perfunctory colors and nc	NEGATIVE	https://ba
1048572	gunpowder_milkshake	2806168	16/07/2021	John Urbanc	FALSE	03-May	fresh	JMuvies	Casting and experience always pay dividends,	POSITIVE	https://jm
1048573	gunpowder_milkshake	2806152	16/07/2021	Matt Lynch	FALSE		rotten	In Review Online	It's carried off with a kind of winking detachm	NEGATIVE	https://in
1048574	gunpowder_milkshake	2806108	16/07/2021	Simon Mirau	FALSE	03-May	fresh	Movie Squad (RT	As stylish as it is, it's kind of learnt the wrong	POSITIVE	https://rt
1048575	gunpowder_milkshake	2806104	16/07/2021	Lupe Rodrigu	FALSE	B+	fresh	CineMovie.tv	The highly-stylized action comedy doesn't tak	POSITIVE	https://yo

Figure 27. DATASET SAMPLE (BOTTOM)

4.3.2 ORIGINAL DATA DESCRIPTION

Each review in the dataset is characterized by several features:

ID: The unique identifier for each review.

reviewID: The name of the movie that was reviewed.

creationDate: The date when the review was created.

criticName: The name of the critic who wrote the review.

isTopCritic: A Boolean value indicating whether the critic is considered a top critic.

originalScore: The original score given by the critic.

reviewState: The state of the review (e.g., fresh, rotten).

publicationName: The name of the publication where the review was published.

reviewText: The text of the review.

scoreSentiment: The sentiment score of the review (positive/negative).

reviewUrl: The URL where the review is published.

4.3.3 ORIGINAL DATA ANALYSIS

Through the thorough data analysis carried out on the dataset certain pattens and details was observed:

1. The distribution of top critics showed that out of all the reviews, 1,008,156 were written by critics who are not considered top critics, and 436,807 were written by those who are considered top critics.
2. The distribution of review states showed that out of all the reviews, 963,799 were labelled as 'fresh' and 481,164 were labelled as 'rotten'.
3. The distribution of sentiment scores showed that out of all the reviews, 963,799 were labelled as having positive sentiment and 481,164 were labelled as having negative sentiment.
4. The average length of positive reviews is approximately 22.01 words, while the average length of negative reviews is approximately 21.25 words. This suggests that reviewers tend to use slightly more words when expressing positive sentiments compared to negative ones. However, the difference is quite small and may not be significant.
5. Among the reviews written by top critics, 275,908 were labelled as having positive sentiment and 160,899 were labelled as having negative sentiment. This indicates that top critics in this dataset are more likely to give positive reviews.

4.3.4 DERIVED DATASET FEATURES AND SENTIMENT ANALYSIS RESULT

Three new columns were created in a new dataset with column names 'sentiment_score', 'sentiment_label', and 'box_office' including the original columns.

These new features:

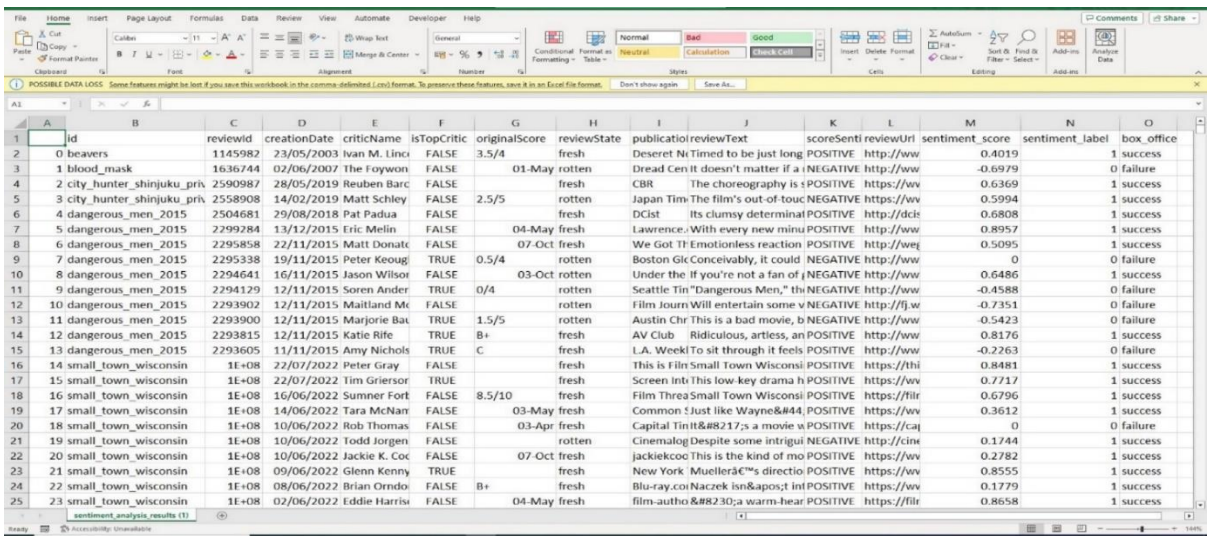
1. **Sentiment Score**: is a numerical measure that quantifies the sentiments expressed in the text. The score is generated by a sentiment analysis model, which assigns a higher score for more positive sentiment and a lower score for more negative sentiment. The range of the score is between -1 and 1, where -1

represents extremely negative sentiment, 0 represents neutral sentiment, and 1 represents extremely positive sentiment.

2. **Sentiment Label**: is a categorical label that classifies the sentiments expressed in the review texts. The label is obtained from the sentiment score of with all texts for example, when a sentiment score is above 0 it is 'positive', and all texts with a score below 0 is 'negative'. Some models might also include a 'neutral' label for texts that don't clearly express positive or negative sentiment.
3. **Box Office**: The whole point of my project, it is used to indicate the box office 'success' or 'failure' of the movies based on the average sentiment score of its reviews.

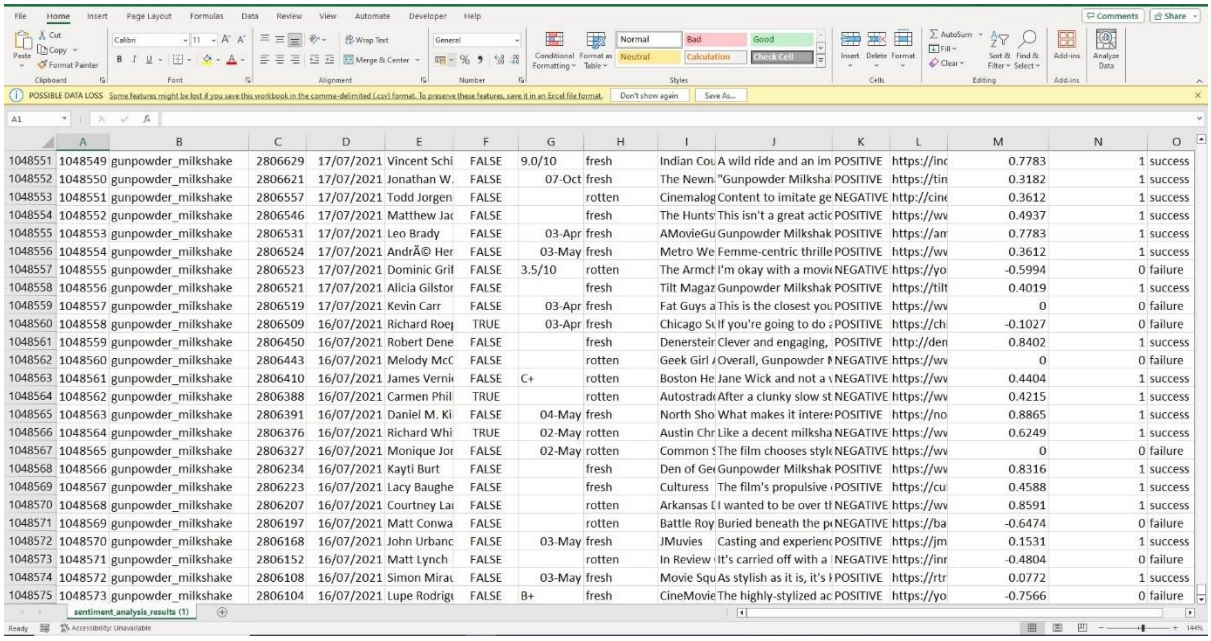
4.3.4.1 Some pattens I noticed analysing the newly trained dataset:

1. **Proportion of Positive and Negative Reviews**: The proportion of positive reviews is approximately 0.61 (or 61.15%), and the proportion of negative reviews is approximately 0.39 (or 38.85%). This means that out of all the reviews in the dataset, about 61% are positive and 39% are negative.
2. **Average Sentiment Scores**: The average sentiment score of 'success' movies are approximately 0.55, and the average sentiment score of 'failure' movies is approximately -0.30. This suggests that 'success' movies tend to receive more positive reviews (as indicated by the higher average sentiment score), while failure' movies tend to receive more negative reviews (as indicated by the lower average sentiment score).



	id	reviewId	creationDate	criticName	isTopCritic	originalScore	reviewState	publicatio	reviewText	scoreSenti	reviewUrl	sentiment_score	sentiment_label	box_office
2	0 beavers	1145982	23/05/2003	Ivan M. Linc	FALSE	3.5/4	fresh	Deseret N	Timed to be just long	POSITIVE	http://ww	0.4019	1	success
3	1 blood_mask	1636744	02/06/2007	The Foywon	FALSE	01-May	rotten	Dread Cen	It doesn't matter if a	NEGATIVE	http://ww	-0.6979	0	failure
4	2 city_hunter_shinjuku_priv	2590987	28/05/2019	Reuben Barc	FALSE		fresh	CBR	The choreography is s	POSITIVE	https://ww	0.6369	1	success
5	3 city_hunter_shinjuku_priv	2558908	14/02/2019	Matt Schley	FALSE	2.5/5	rotten	Japan Tim	The film's out-of-touc	NEGATIVE	https://ww	0.5994	1	success
6	4 dangerous_men_2015	2504681	29/08/2018	Pat Padua	FALSE		fresh	DCist	Its clumsy determinat	POSITIVE	http://dcis	0.6808	1	success
7	5 dangerous_men_2015	2299284	13/12/2015	Eric Melin	FALSE	04-May	fresh	Lawrence.(	With every new minu	POSITIVE	http://ww	0.8957	1	success
8	6 dangerous_men_2015	2295858	22/11/2015	Matt Donato	FALSE	07-Oct	fresh	We Got Th	Emotionless reaction	POSITIVE	http://weg	0.5095	1	success
9	7 dangerous_men_2015	2295338	19/11/2015	Peter Keoug	TRUE	0.5/4	rotten	Boston Gl	Conceivably, it could	NEGATIVE	http://ww	0	0	failure
10	8 dangerous_men_2015	2294641	16/11/2015	Jason Wilsor	FALSE	03-Oct	rotten	Under the	If you're not a fan of	NEGATIVE	http://ww	0.6486	1	success
11	9 dangerous_men_2015	2294129	12/11/2015	Soren Ander	TRUE	0/4	rotten	Seattle Tin	"Dangerous Men," th	NEGATIVE	http://ww	-0.4588	0	failure
12	10 dangerous_men_2015	2293902	12/11/2015	Maitland Mc	FALSE		rotten	Film Journ	Will entertain some v	NEGATIVE	http://fj.w	-0.7351	0	failure
13	11 dangerous_men_2015	2293900	12/11/2015	Marjorie Bau	TRUE	1.5/5	rotten	Austin Chr	This is a bad movie, b	NEGATIVE	http://ww	-0.5423	0	failure
14	12 dangerous_men_2015	2293815	12/11/2015	Katie Rife	TRUE	B+	fresh	AV Club	Ridiculous, artless, an	POSITIVE	http://ww	0.8176	1	success
15	13 dangerous_men_2015	2293605	11/11/2015	Amy Nichols	TRUE	C	fresh	L.A. Week	To sit through it feels	POSITIVE	http://ww	-0.2263	0	failure
16	14 small_town_wisconsin	1E+08	22/07/2022	Peter Gray	FALSE		fresh	This is Film	Small Town Wisconsi	POSITIVE	https://thi	0.8481	1	success
17	15 small_town_wisconsin	1E+08	22/07/2022	Tim Grierson	TRUE		fresh	Screen Int	This low-key drama h	POSITIVE	https://ww	0.7717	1	success
18	16 small_town_wisconsin	1E+08	16/06/2022	Sumner Fort	FALSE	8.5/10	fresh	Film Threa	Small Town Wisconsi	POSITIVE	https://filr	0.6796	1	success
19	17 small_town_wisconsin	1E+08	14/06/2022	Tara McNam	FALSE	03-May	fresh	Common S	Just like Wayne,	POSITIVE	https://ww	0.3612	1	success
20	18 small_town_wisconsin	1E+08	10/06/2022	Rob Thomas	FALSE	03-Apr	fresh	Capital Tin	It’s a movie w	POSITIVE	https://ca	0	0	failure
21	19 small_town_wisconsin	1E+08	10/06/2022	Todd Jorgen	FALSE		rotten	Cinemalog	Despite some intrigui	NEGATIVE	http://cin	0.1744	1	success
22	20 small_town_wisconsin	1E+08	10/06/2022	Jackie K. Coc	FALSE	07-Oct	fresh	jackiekcoo	This is the kind of mo	POSITIVE	https://ww	0.2782	1	success
23	21 small_town_wisconsin	1E+08	09/06/2022	Glenn Kenny	TRUE		fresh	New York	Muellerâ€™s directio	POSITIVE	https://ww	0.8555	1	success
24	22 small_town_wisconsin	1E+08	08/06/2022	Brian Orndo	FALSE	B+	fresh	Blu-ray.co	Naczek isn't int	POSITIVE	https://ww	0.1779	1	success
25	23 small_town_wisconsin	1E+08	02/06/2022	Eddie Harris	FALSE	04-May	fresh	film-autho	…a warm-hear	POSITIVE	https://filr	0.8658	1	success

Figure 28. RESULT OF SENTIMENT ANALYSIS ON THE DATASET (TOP)



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1048551	1048549	gunpowder_milkshake	2806629	17/07/2021	Vincent Schi	FALSE	9.0/10	fresh	Indian Cou	A wild ride and an im	POSITIVE	https://inc	0.7783	1	success
1048552	1048550	gunpowder_milkshake	2806621	17/07/2021	Jonathan W.	FALSE	07-Oct	fresh	The Newn	"Gunpowder Milksha	POSITIVE	https://tin	0.3182	1	success
1048553	1048551	gunpowder_milkshake	2806557	17/07/2021	Todd Jorgen	FALSE		rotten	Cinemalog	Content to imitate ge	NEGATIVE	http://cine	0.3612	1	success
1048554	1048552	gunpowder_milkshake	2806546	17/07/2021	Matthew Jac	FALSE		fresh	The Hunts	This isn't a great actic	POSITIVE	https://ww	0.4937	1	success
1048555	1048553	gunpowder_milkshake	2806531	17/07/2021	Leo Brady	FALSE	03-Apr	fresh	AMovieGu	Gunpowder Milkshak	POSITIVE	https://am	0.7783	1	success
1048556	1048554	gunpowder_milkshake	2806524	17/07/2021	AndrÃ© Her	FALSE	03-May	fresh	Metro We	Femme-centric thrille	POSITIVE	https://ww	0.3612	1	success
1048557	1048555	gunpowder_milkshake	2806523	17/07/2021	Dominic Gril	FALSE	3.5/10	rotten	The Armcl	I'm okay with a movie	NEGATIVE	https://yo	-0.5994	0	failure
1048558	1048556	gunpowder_milkshake	2806521	17/07/2021	Alicia Gilstor	FALSE		fresh	Tilt Magaz	Gunpowder Milkshak	POSITIVE	https://tilt	0.4019	1	success
1048559	1048557	gunpowder_milkshake	2806519	17/07/2021	Kevin Carr	FALSE	03-Apr	fresh	Fat Guys a	This is the closest you	POSITIVE	https://ww	0	0	failure
1048560	1048558	gunpowder_milkshake	2806509	16/07/2021	Richard Roe	TRUE	03-Apr	fresh	Chicago Su	If you're going to do a	POSITIVE	https://ch	-0.1027	0	failure
1048561	1048559	gunpowder_milkshake	2806450	16/07/2021	Robert Dene	FALSE		fresh	Denerstein	Clever and engaging,	POSITIVE	http://den	0.8402	1	success
1048562	1048560	gunpowder_milkshake	2806443	16/07/2021	Melody McC	FALSE		rotten	Geek Girl /	Overall, Gunpowder M	NEGATIVE	https://ww	0	0	failure
1048563	1048561	gunpowder_milkshake	2806410	16/07/2021	James Verni	FALSE	C+	rotten	Boston He	Jane Wick and not a v	NEGATIVE	https://ww	0.4404	1	success
1048564	1048562	gunpowder_milkshake	2806388	16/07/2021	Carmen Phil	TRUE		rotten	Autostrad	After a clunky slow st	NEGATIVE	https://ww	0.4215	1	success
1048565	1048563	gunpowder_milkshake	2806391	16/07/2021	Daniel M. Ki	FALSE	04-May	fresh	North Sho	What makes it intere	POSITIVE	https://no	0.8865	1	success
1048566	1048564	gunpowder_milkshake	2806376	16/07/2021	Richard Whi	TRUE	02-May	rotten	Austin Chr	Like a decent milksha	NEGATIVE	https://ww	0.6249	1	success
1048567	1048565	gunpowder_milkshake	2806327	16/07/2021	Monique Jor	FALSE	02-May	rotten	Common S	The film chooses style	NEGATIVE	https://ww	0	0	failure
1048568	1048566	gunpowder_milkshake	2806234	16/07/2021	Kayti Burt	FALSE		fresh	Den of Ge	Gunpowder Milkshak	POSITIVE	https://ww	0.8316	1	success
1048569	1048567	gunpowder_milkshake	2806223	16/07/2021	Lacy Baughe	FALSE		fresh	Culturess	The film's propulsive	POSITIVE	https://cu	0.4588	1	success
1048570	1048568	gunpowder_milkshake	2806207	16/07/2021	Courtney La	FALSE		rotten	Arkansas D	I wanted to be over th	NEGATIVE	https://ww	0.8591	1	success
1048571	1048569	gunpowder_milkshake	2806197	16/07/2021	Matt Conwa	FALSE		rotten	Battle Roy	Buried beneath the pr	NEGATIVE	https://ba	-0.6474	0	failure
1048572	1048570	gunpowder_milkshake	2806168	16/07/2021	John Urbanc	FALSE	03-May	fresh	JMuvies	Casting and experienc	POSITIVE	https://jm	0.1531	1	success
1048573	1048571	gunpowder_milkshake	2806152	16/07/2021	Matt Lynch	FALSE		rotten	In Review	It's carried off with a	NEGATIVE	https://inr	-0.4804	0	failure
1048574	1048572	gunpowder_milkshake	2806108	16/07/2021	Simon Mirau	FALSE	03-May	fresh	Movie Squ	As stylish as it is, it's h	POSITIVE	https://rtr	0.0772	1	success
1048575	1048573	gunpowder_milkshake	2806104	16/07/2021	Lupe Rodrigu	FALSE	B+	fresh	CineMovie	The highly-stylized ac	POSITIVE	https://yo	-0.7566	0	failure

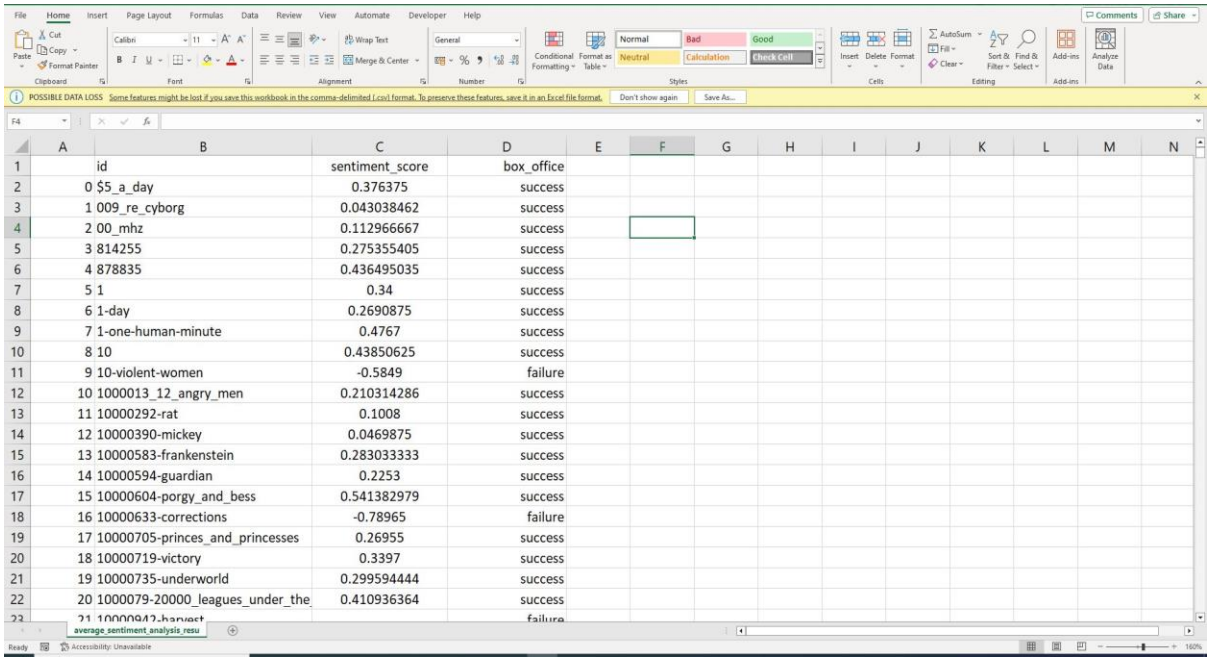
Figure 29. RESULT OF SENTIMENT ANALYSIS ON THE DATASET (BOTTOM)

4.3.5 AVERAGE SENTIMENT DATASET

This average sentiment dataset is the aggregate of all the sentiment_score of the reviews to get the overall sentiment, in other words, prediction of the movie, its feature are:

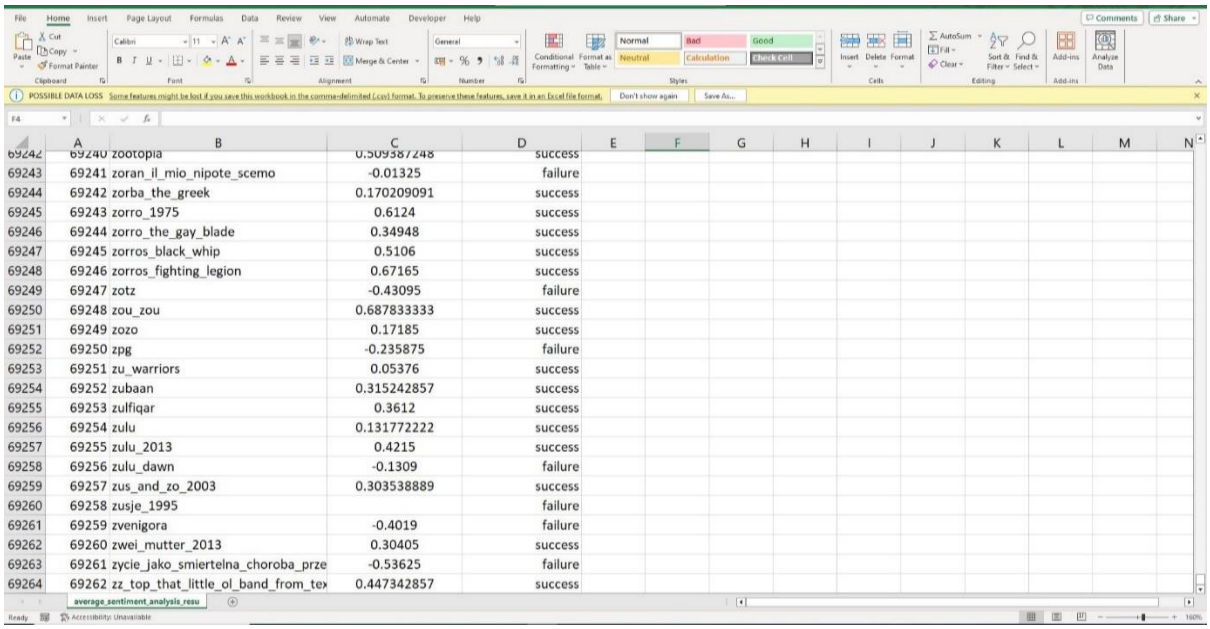
- id: This column contains identifiers for the movies. Each identifier is unique to a movie.
- sentiment_score: This column contains the average sentiment score for each movie. The sentiment score is a numerical measure of the sentiment of the reviews for the movie, with higher scores indicating more positive sentiment.
- box_office: This column categorizes each movie as either a ‘success’ or ‘failure’ based on its average sentiment score. If the average sentiment score is greater than 0, the movie is labelled as a ‘success’. If the average sentiment score is less than or equal to 0, the movie is labelled as a ‘failure’.

This dataset provides a summary of the sentiment analysis results for each movie in the dataset totalling observations to 69,264. For each movie, it shows the average sentiment of the reviews and whether the movie was a ‘success’ or ‘failure’ based on these reviews.



	id	sentiment_score	box_office
0	$5_a_day	0.376375	success
1	009_re_cyborg	0.043038462	success
2	00_mhz	0.112966667	success
3	814255	0.275355405	success
4	878835	0.436495035	success
5	1	0.34	success
6	1-day	0.2690875	success
7	1-one-human-minute	0.4767	success
8	10	0.43850625	success
9	10-violent-women	-0.5849	failure
10	1000013_12_angry_men	0.210314286	success
11	10000292-rat	0.1008	success
12	10000390-mickey	0.0469875	success
13	10000583-frankenstein	0.283033333	success
14	10000594-guardian	0.2253	success
15	10000604-porgy_and_bess	0.541382979	success
16	10000633-corrections	-0.78965	failure
17	10000705-princes_and_princesses	0.26955	success
18	10000719-victory	0.3397	success
19	10000735-underworld	0.299594444	success
20	1000079-20000_leagues_under_the	0.410936364	success
21	10000942-harvest		failure

Figure 30. Average Sentiment Dataset (Top)



69240	zootopia	0.509387248	success
69241	zoran_il_mio_nipote_scemo	-0.01325	failure
69242	zorba_the_greek	0.170209091	success
69243	zorro_1975	0.6124	success
69244	zorro_the_gay_blade	0.34948	success
69245	zorros_black_whip	0.5106	success
69246	zorros_fighting_legion	0.67165	success
69247	zotz	-0.43095	failure
69248	zou_zou	0.687833333	success
69249	zozo	0.17185	success
69250	zpg	-0.235875	failure
69251	zu_warriors	0.05376	success
69252	zubaan	0.315242857	success
69253	zulfiqar	0.3612	success
69254	zulu	0.131772222	success
69255	zulu_2013	0.4215	success
69256	zulu_dawn	-0.1309	failure
69257	zus_and_zo_2003	0.303538889	success
69258	zusje_1995		failure
69259	zvenigora	-0.4019	failure
69260	zwei_mutter_2013	0.30405	success
69261	zycie_jako_smiertelna_choroba_prze	-0.53625	failure
69262	zz_top_that_little_ol_band_from_tex	0.447342857	success

Figure 31. Average Sentiment Dataset (Bottom)

The original dataset was enhanced with new features - sentiment score, sentiment label, and box office - to facilitate our analysis. The sentiment score, a numerical measure of the sentiment expressed in each review, allowed us to quantify and compare sentiments across reviews. The sentiment label categorized the sentiment as positive or negative, providing a clear classification for each review. The box office feature was used to indicate the box office 'success' or 'failure' of the movies based on the average sentiment score of its reviews.

4.4. SUB-QUESTION RESULTS

4.4.1 Sub-Question 1: Correlation between Sentiment Score and Box Office Success

The correlation between sentiment score and box office success was found to be 0.8466, indicating a strong positive relationship. This suggests that movies with higher sentiment scores tend to have higher box office success.

```
# Calculate the correlation
correlation = df['sentiment_score'].corr(df['box_office'])
print(f"Correlation between sentiment score and box office success: {correlation}")
```

```
Correlation between sentiment score and box office success: 0.8465537140544508
```

Figure 32. Correlation between Sentiment score and Box Office

4.4.2 Sub-Question 2: Performance of the LSTM Model

The performance of the LSTM model was evaluated using various metrics. The model achieved an accuracy of 99.39%, a precision of 99.62%, a recall of 99.39%, an F1 score of 0.9950, and an ROC AUC score of 0.9984. These results suggest that the model is highly effective in predicting a movie's box office success based on sentiment analysis of its reviews.

```
Accuracy: 0.9939421604985833
Precision: 0.9962206060374984
Recall: 0.9938640972425287
F1 Score: 0.9950409564396757
ROC AUC: 0.9984234675643366
```

Figure 33. LSTM Model Results

In conclusion, our analysis found a strong positive correlation between the sentiment expressed in movie reviews and box office success. This suggests that public sentiment plays a significant role in a movie's financial success. The performance of the Long Short-Term Memory model further validated these findings, demonstrating high accuracy in predicting box office success based on review sentiment. These insights not only contribute to our understanding of the impact of public opinion on box office success but also provide valuable information for movie production companies. In the next chapter, we will summarize the key findings of our analysis and discuss their implications in further detail.

5. CONCLUSION

This chapter concludes the exploration in the field of sentiment analysis and its potential for predicting box office success based on the movie textual reviews. By thoroughly exploring a rich dataset and using robust data science techniques, I discovered essential facts that clarifies the complicated connection between audience opinion and financial result.

5.1 Summary of Findings

5.1.1 Research Questions

The project was guided by specific research questions that required thorough answers based on my findings:

Research Question 2: "*What is the relationship between sentiment expressed in movie reviews and box office success?*"

The analysis revealed a significantly positive correlation between the sentiment expressed in movie textual reviews and box office success. This means that movies with a higher proportion of positive sentiment in their reviews consistently tended to earn more at the box office. This correlation suggests that positive sentiments made through reviews plays a crucial role in influencing audience decisions and driving ticket sales.

Research Question 3: "*How effective is the Long Short-Term Memory (LSTM) model in predicting a movie's box office success based on sentiment analysis of its reviews?*"

The Long Short-Term Memory (LSTM) model demonstrated remarkable performance in predicting movie box office success based on sentiment analysis of reviews. The model achieved an exceptional accuracy of 99.39%, indicating its ability to correctly classify movies as either high- or low-grossing based on their sentiment score with remarkable precision. Additionally, the high values for precision (99.62%), recall (99.39%), F1 score (0.9950), and ROC AUC score (0.9984) further reinforce the model's effectiveness.

These results suggest that LSTMs, with their ability to learn long-term dependencies in sequential data like review text, can be powerful tools for predicting box office success based on sentiment analysis.

These sub-questions helped answer the main objective **Research Question 1**:

"*How can sentiment analysis of movie reviews effectively predict box office sales and inform strategic decision-making within the film industry?*"

Predicting audience responses to a film before its release is an asset for the film industry, and sentiment analysis of movie reviews offers a wealth of potential in this regard. By tapping into the rich pool of textual opinions in reviews, studios can utilize audience sentiments to shape strategic decisions throughout the filmmaking process, from scriptwriting to marketing strategies.

While the rise of streaming platforms has disrupted the traditional distribution model of movies, challenged the dominance of theatres and altering the way films reach audiences. Although they simplify the production and distribution of more diverse and niche content because they are not as limited by the traditional Hollywood studio system, streaming services can take more chances on relatively unknown filmmakers and projects, which might have a more challenging time securing funding or distribution through conventional channels. This has resulted in a broader range of voices and perspectives being represented in film and television.

Platforms like Tubi, which operates on a free, ad-supported model, provide filmmakers with a unique opportunity to reach a large audience with their content. The revenue they receive from Tubi is directly proportional to the amount of ad revenue generated by their content. While the revenue generated from advertising may not be as high as it would be on a paid streaming service such as Netflix or Hulu, the vast reach of Tubi, with over 200 million monthly active users, provides filmmakers with a unique opportunity to reach a large audience with their content.

In this context, sentiment analysis becomes even more crucial. With the ability to reach a larger and more diverse audience, understanding audience sentiment can guide script development, casting choices, and even editing decisions, potentially leading to movies that better connect with viewers on an emotional level. Sentiment analysis can inform targeted marketing campaigns, allowing studios to tailor their messages to specific demographics and audience segments based on their predicted preference for certain genres or themes.

In summary, sentiment analysis of textual movie reviews is more than a tool for predicting box office results; it's a potent instrument through which the film industry can delve into its audience's emotional depths. From creating resonating narratives to executing targeted marketing strategies, sentiment analysis offers invaluable insights that can steer strategic decision-making at every turn of the filmmaking journey, ultimately resulting in films that genuinely resonate with and move audiences. By embracing the power of sentiment analysis, the film industry can create movies that not only entertain but also resonate deeply with audiences, ultimately driving box office success. This is particularly important in the era of streaming platforms, where understanding and connecting with the audience's sentiment can make the difference between a film's success or failure.

5.1.2 Exploratory Data Analysis (EDA) Insights:

The initial EDA phase revealed valuable patterns within the review data. I have found out that:

1. Most reviews expressed positive sentiment, highlighting its crucial role in driving box office numbers.
2. Top critics leaned towards positive reviews, demonstrating their influence on shaping public opinion and potentially impacting box office performance.
3. The language used in the reviews offered unique clues. While positive reviews were slightly longer, specific words like "love," "enjoyed," and "fun" frequently appeared, hinting at the nuanced ways viewers express their emotions.

5.1.3 Sentiment Analysis and Box Office Prediction:

Moving beyond mere classification, I had built a novel dataset incorporating sentiment scores and box office success labels. This enabled me to demonstrate a clear correlation between positive sentiment and box office success, with successful movies boasting significantly higher average sentiment scores.

These emphasizes the significant influence of movie reviews on a film's financial success. It suggests that positive reviews, especially from top critics, can shape public opinion and drive box office sales. The specific language used in these reviews, marked by words expressing enjoyment, can provide valuable insights into viewer emotions. In the bigger picture, this underscores the power of sentiment in shaping a movie's

reception and its ultimate financial performance. It highlights the importance of understanding and analysing these reviews for filmmakers, marketers, and others in the industry. This analysis could potentially guide strategies for film promotion and even content creation. It's a testament to the interconnected nature of viewer sentiment, critical reviews, and a movie's financial success.

5.2 Implications and Applications

5.2.1 Impact on the Movie Industry:

These findings hold significant implications for the movie industry:

Enhanced prediction models: Sentiment analysis can enhance existing box office prediction models, offering movie studios valuable insights into audience preferences and potential financial outcomes.

Data-driven storytelling: This research underscores the importance of understanding audience sentiment throughout the filmmaking process, from the script development to promotion and marketing, allowing studios to create movies that resonate more deeply with viewers.

Shifting power dynamics: By empowering movie studios with a better understanding of public opinion, sentiment analysis can potentially shift power dynamics within the entertainment industry, giving audiences a stronger voice in shaping the movies they want to see.

5.2.2 Insights for Decision-Makers:

Decision-makers, from producers to marketing teams, can amass crucial knowledge from this research:

Targeted marketing campaigns: Sentiment analysis can inform targeted marketing campaigns, allowing the movie studios to tailor their messages to specific demographics and audience segments based on their predicted preference for certain genres or themes.

Fine-tuning film production: Understanding audience sentiment can guide script development, casting choices, and even editing decisions, potentially leading to movies that better connect with the audience and even casual viewers on an emotional level.

Risk assessment and optimization: Sentiment analysis provides valuable data for risk assessment, allowing movie studios to identify potential pitfalls and optimize their investment strategies based on predicted audience reception.

These findings have led to significant implications for the movie industry that sentiment analysis can revolutionize the film industry. By understanding audience emotions and preferences, the industry can make more informed decisions at every stage, it underscores the potential of leveraging data analysis and technology to enhance creativity and business outcomes in the film industry. It's about aligning the art of filmmaking with audience sentiment to drive success.

5.3 Limitations and Future Directions

5.3.1 Addressing Limitations:

The analysis of a large dataset with 1,048,575 observations and 11 features was constrained by the computational limitations encountered using Google Colab, resulting in crashes. Consequently, only a partial correlation matrix could be generated. The variables 'isTopCritic', 'scoreSentiment', and 'review_length' exhibited weak linear relationships as indicated by their low correlation coefficients. This limitation underscores the need for enhanced computational resources and optimized algorithms to efficiently handle large datasets.

The dataset used in the project includes a little over 5000 movies from various genres. While this is a substantial number, it may not fully represent the diversity and complexity of the global movie-going audience or the entire spectrum of movie genres. This limitation could potentially skew the results and conclusions drawn from the analysis. For instance, certain genres or audience preferences might be underrepresented, leading to a potential bias in the findings.

It should be acknowledged that box office success is not solely determined by audience sentiment. Numerous external factors, such as the effectiveness of marketing campaigns, the level of competition at the time of release, and the timing of the movie release, can significantly influence a movie's financial performance. These factors were not considered in the analysis, which might limit the comprehensiveness of my findings.

5.3.2 Future Research Opportunities:

These limitations open doors for exciting future research avenues:

For future directions, exploring using advanced statistical methods or machine learning techniques that are adept at uncovering complex, non-linear relationships within large datasets. Additionally, research on more robust computing platforms or distributed computing to handle such large datasets. This could potentially allow for a more comprehensive analysis and stronger insights from the data used.

Incorporate a more diverse and representative dataset. This could include data from a wider range of movies, spanning different genres, languages, and cultural contexts. Also, integrating audience demographic data could provide a more nuanced understanding of audience sentiment and preferences, intend to explore that further.

To gain a more comprehensive picture of audience sentiment and preferences, also consider expanding the scope of the data sources. This involves analysing data from a variety of platforms, including social media sites and online forums. Such platforms often contain rich, unfiltered opinions from a broad spectrum of viewers, which could provide valuable insights into audience sentiment.

This project already offers substantial insights in the field of sentiment analysis, but it is crucial to recognize its constraints, while the dataset utilized may not fully encapsulate the diversity of the global movie-going audience or the breadth of movie genres and the precision of sentiment analysis models can fluctuate, as they may not entirely grasp the subtleties of human emotions articulated in text numerous factors beyond audience sentiment, such as marketing strategies, competitive landscape, and timing of release all have influences on the box office success. However, these constraints pave the way for intriguing future research opportunities and subsequent studies could enhance sentiment analysis models, investigate additional determinants of box office success, and broaden the ambit of data analysis.

5.4 Final Conclusion

The journey through this project has illuminated the immense potential of sentiment analysis in demystifying the relationship between audience perception and box office success. While limitations exist and further research is necessary, the insights gained

pave the way for a future where data-driven decision-making empowers the movie industry to create films that not only entertain but also resonate deeply with the hearts and minds of their audiences.

By embracing the power of sentiment analysis, we can embark on a new era of filmmaking, one where stories truly connect with the voices that matter most – the voices of those who watch, experience, and ultimately shape the magic of cinema.

6. PROJECT MANAGEMENT

This chapter outlines the project management strategies employed throughout this research. It includes a discussion of the project timeline, task management, risk assessment, and legal/social/ethical considerations.

6.1 Project Timeline (Gantt Chart)

The Gantt chart was used to manage the timeline of the project. It provided a visual representation of the project schedule, showing the start and end dates of each task, their dependencies, and progress. The Gantt chart was updated regularly to reflect the actual progress of the project against the planned schedule.

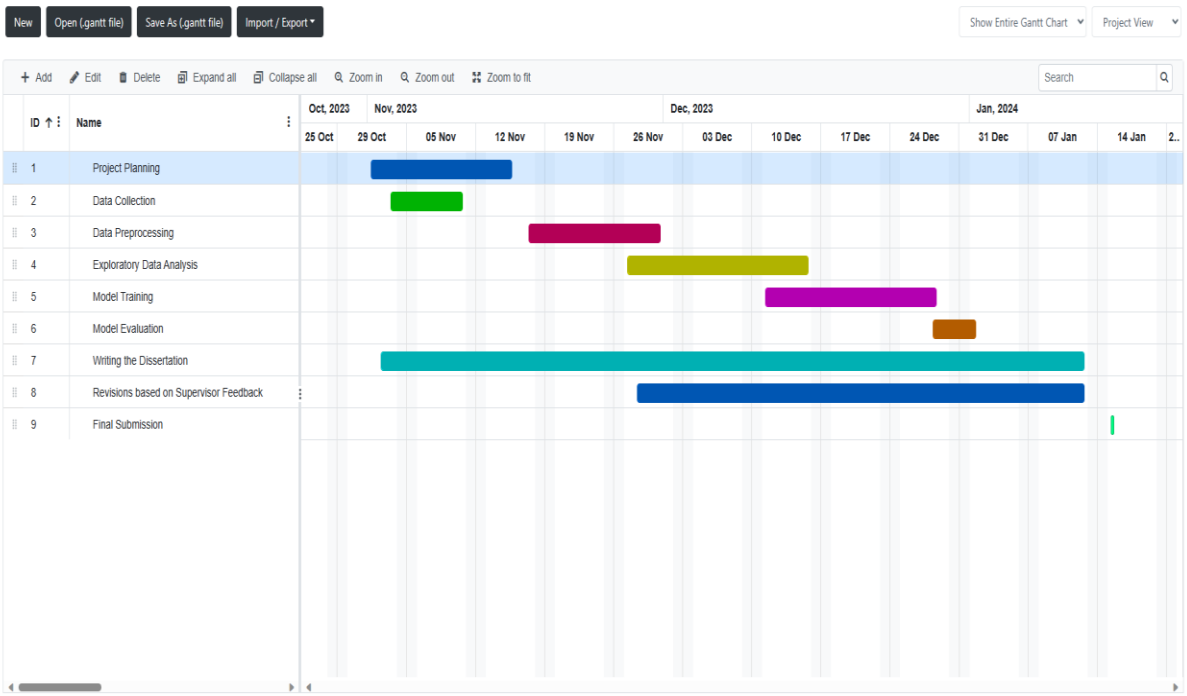


Figure 34. Gantt Chart

6.2 Risk Assessment

Data Quality and Availability: The quality and availability of textual movie reviews could pose a risk. If the reviews are not sufficiently expressive or if there aren't enough reviews for some movies, it will impact the performance of the sentiment analysis model. To mitigate this risk, sourcing data from multiple platforms or use data augmentation techniques are necessary.

Model Performance: Despite using a Long Short-Term Memory (LSTM model), there's a risk that the model might not absolutely capture the nuances of the sentiments expressed in the reviews and this can lead to inaccurate predictions, while regular model evaluation and tuning, as well as exploring other models or ensemble methods, could help mitigate this risk successfully.

Subjectivity of Reviews: Movie reviews are subjective and can vary greatly among viewers. This subjectivity of a particular movie can affect the sentiment analysis and, consequently, the box office predictions. A potential mitigation strategy is to weigh the reviews based on the credibility of the reviewer or the platform (isTopCritic) as used in this project.

Changes in Movie Industry Trends: The movie industry is dynamic, and trends tend to change rapidly. These changes could affect both the sentiments expressed in reviews and box office success. Keeping the model updated with recent data can help manage this risk.

Ethical Considerations: There are ethical considerations when using public reviews for analysis. Ensuring anonymity and privacy of the reviewers is crucial that is why I ensured critic names were taking out of the dataset during the data cleaning process. Also, the findings of the project have been used responsibly, considering the potential impact on the movie industry and individual careers.

6.3 Legal/Social/Ethical Considerations

Legal Considerations

The dataset used in this project is publicly available online, which mitigates legal concerns related to data acquisition. However, during the data cleaning process, additional steps to respect the privacy of the reviewers. Specifically, the names of the critics were removed from the dataset to ensure anonymity.

Social Considerations

From a social perspective, this project has the potential to influence the movie industry by providing insights into audience preferences, which could lead to the production of movies that better cater to audience tastes. However, it's important to consider the potential for bias in the data. For instance, if the reviews disproportionately represent certain demographics, the findings might not be applicable to the broader audience.

Ethical Considerations

Ethically, while this project aims to provide valuable insights to the movie industry, it's important to balance data-driven decision-making with the recognition of filmmaking as a form of artistic expression. The findings of this project should be used responsibly, considering the potential impact on the movie industry and individual careers.

In conclusion, it's crucial to consider the legal, social, and ethical implications of a data science project like this one. By carefully considering these aspects, I can ensure that the project was conducted responsibly and contributes positively to the field.

7. APPENDIX

A.

Sheffield Hallam University

UREC2 RESEARCH ETHICS PROFORMA FOR STUDENTS UNDERTAKING LOW RISK PROJECTS WITH HUMAN PARTICIPANTS

This form is designed to help students and their supervisors to complete an ethical scrutiny of proposed research. The University Research Ethics Policy (www.shu.ac.uk/research/excellence/ethics-and-integrity/policies) should be consulted before completing this form. The initial questions are there to check that completion of the UREC 2 is appropriate for this study. The final responsibility for ensuring that ethical research practices are followed rests with the supervisor for student research.

Note that students and staff are responsible for making suitable arrangements to ensure compliance with the General Data Protection Act (GDPR). This involves informing participants about the legal basis for the research, including a link to the University research data privacy statement and providing details of who to complain to if participants have issues about how their data was handled or how they were treated (full details in module handbooks). In addition, the act requires data to be kept securely and the identity of participants to be anonymised. They are also responsible for following SHU guidelines about data encryption and research data management. Guidance can be found on the SHU Ethics Website www.shu.ac.uk/research/excellence/ethics-and-integrity

Please note that it is mandatory for all students to only store data on their allotted networked F drive space and not on individual hard drives or memory sticks etc.

The present form also enables the University and College to keep a record confirming that research conducted has been subjected to ethical scrutiny.

The form must be completed by the student and the supervisor and independently reviewed by a second reviewer or module leader (additional guidance can be obtained from your College Research Ethics Chair[1]). In all cases, it should be counter-signed and kept as a record showing that ethical scrutiny has occurred. Some courses may require additional scrutiny. Students should retain a copy for inclusion in their research project, and a copy should be uploaded to the relevant module Blackboard site.

Please note that it may be necessary to conduct a health and safety risk assessment for the proposed research. Further information can be obtained from the University's Health and Safety Website https://sheffieldhallam.sharepoint.com/sites/3069/SitePages/Risk-Assessment.aspx

SECTION A

1. Checklist questions to ensure that this is the correct form:

Health Related Research within the NHS, or His Majesty's Prison and Probation Service (HMPPS), or with participants unable to provide informed consent check list.

[1] College of Social Sciences and Arts - Dr. Antonia Ypsilanti (a.ypsilanti@shu.ac.uk)
College of Business, Technology and Engineering - Dr. Tony Lynn (t.lynn@shu.ac.uk)
College of Health, Wellbeing and Life Sciences - Dr. Nikki Jordan-Mahy (n.jordan-mahy@shu.ac.uk)

Question	Yes/No
Does the research involve? • Patients recruited because of their past or present use of the NHS	NO
• Relatives/carers of patients recruited because of their past or present use of the NHS	NO
• Access to NHS staff, premises, or resources	NO
• Access to data, organs, or other bodily material of past or present NHS patients	NO
• Foetal material and IVF involving NHS patients	NO
• The recently dead in NHS premises	NO
• Prisoners or others within the criminal justice system recruited for health-related research	NO
• Police, court officials, prisoners, or others within the criminal justice system	NO
• Participants who are unable to provide informed consent due to their incapacity even if the project is not health related	NO
• Is this an NHS research project, service evaluation or audit? *For NHS definitions please see the following website* http://www.hra.nhs.uk/documents/2013/09/defining-research.pdf	NO

If you have answered **YES** to any of the above questions, then you **MUST consult with your supervisor** to obtain research ethics from the appropriate institution outside the university. This could be from the NHS or Her Majesty's Prison and Probation Service (HMPPS) under their independent Research Governance schemes. Further information is provided below. https://www.myresearchproject.org.uk/

2. Checks for Research with Human Participants

Question	Yes/No
1. Will any of the participants be vulnerable? *Note: Vulnerable people include children and young people, people with learning disabilities, people who may be limited by age or sickness, pregnancy, people researched because of a condition they have, etc. See full definition on ethics website in the document* ***Code of Practice for Researchers Working with Vulnerable Populations*** *(under the Supplementary University Polices and Good Research Practice Guidance)*	NO
2. Are drugs, placebos, or other substances (e.g., food substances, vitamins) to be administered to the study participants or will the study involve invasive, intrusive, or potentially harmful procedures of any kind?	NO
3. Will tissue samples (including blood) be obtained from participants?	NO
4. Is pain or more than mild discomfort likely to result from the study?	NO
5. Will the study involve prolonged or repetitive testing?	NO
6. Is there any reasonable and foreseeable risk of physical or emotional harm to any of the participants? *Note: Harm may be caused by distressing or intrusive interview questions, uncomfortable procedures involving the participant, invasion of privacy, topics relating to highly personal information, topics relating to illegal activity, or topics that are anxiety provoking, etc.*	NO

Question	Yes/No
7. Will anyone be taking part without giving their informed consent?	YES
8. Is the research covert? *Note: 'Covert research' refers to research that is conducted without the knowledge of participants.*	YES
9. Will the research output allow identification of any individual who has not given their express consent to be identified?	NO

If you have answered **YES** to any of these questions you are **REQUIRED** to complete and submit a UREC3 or UREC4 form. Your supervisor will advise. If you have answered **NO** to all these questions, then proceed with this form (UREC2).

3. General Project Details

Details	
Name of student	Gideon Oluwatobi Mautin
SHU email address	C2062745@hallam.shu.ac.uk
Department/College	Faculty of Business Technology And Engineering
Name of supervisor	Dr. Diana Hintea
Supervisor's email address	D.Hintea@shu.ac.uk
Title of proposed research	Sentiment Analysis of Movie Reviews for The Prediction of Box Office
Proposed start date	26/09/2023
Proposed end date	09/01/2024
Background to the study and the rationale (reasons) for undertaking the research (500 words)	As moviegoers increasingly turn to online platforms to share their opinions, the need to decode and harness the sentiments expressed within these digital conversations becomes pivotal for stakeholders across the film industry value chain, from filmmakers and producers to marketers and distributors. Understanding the underlying emotional tones, positive or negative, can offer profound insights into the factors that shape movie success. By delving into the world of sentiment analysis, this research aims to unravel the intricate relationship between public sentiment and box office sales, thus offering a novel lens through which industry practitioners can optimize decision-making and strategies.
Aims & research question(s)	Research Aim: The primary aim of this research project is to investigate and demonstrate the efficacy of sentiment analysis as a predictive tool

Details	
	for box office sales within the film industry. By unraveling the emotional nuances expressed in movie reviews, this study seeks to establish a robust predictive model that can offer valuable insights into the dynamic relationship between public sentiment and movie revenue. Research Question: "How can sentiment analysis of movie reviews effectively predict box office sales and enhance decision-making within the film industry?"
Methods to be used for: 1. Recruitment of participants 2. Data collection 3. Data analysis	Participant Recruitment: Purposeful Sampling with Online Recruitment Data Collection: Web Scraping for Textual Data Data Analysis: Mixed-Methods Approach
Outline the nature of the data held, details of anonymisation, storage and disposal procedures as required.	Nature of Data Held: The research holds both qualitative and quantitative data related to movie reviews, including textual reviews, sentiment scores, and potential demographic details of participants. Anonymization: To ensure privacy, participants will be assigned pseudonyms, personal details will be removed from data, and demographic data will be aggregated. Interview transcripts will use pseudonyms or generic identifiers. Data Storage: Data will be stored securely on restricted-access servers or in the cloud with encryption. Regular backups and access controls will be in place to protect data. Data Disposal: Data will be retained as needed, securely deleted when no longer necessary, and procedures will be documented. Participant data will have identifiers permanently removed, and related documents will be destroyed to comply with ethical standards.

4. Research in External Organisations

Question	Yes/No
1. Will the research involve working with/within an external organisation (e.g., school, business, charity, museum, government department, international agency, etc.)?	**NO**
2. If you answered YES to question 1, do you have granted access to conduct the research from the external organisation? *If YES, students please show evidence to your supervisor. You should retain this evidence safely.*	**NO**
3. If you do not have permission for access is this because: A. you have not yet asked B. you have asked and not yet received an answer C. you have asked and been refused access *Note: You will only be able to start the research when you have been granted access.*	**NO**

5. Research with Products and Artefacts

Question	Yes/No
1. Will the research involve working with copyrighted documents, films, broadcasts, photographs, artworks, designs, products, programs, databases, networks, processes, existing datasets, or secure data?	**YES**
2. If you answered YES to question 1, are the materials you intend to use in the public domain? *Notes: 'In the public domain' does not mean the same thing as 'publicly accessible'.* • *Information which is 'in the public domain' is no longer protected by copyright (i.e., copyright has either expired or been waived) and can be used without permission.* • *Information which is 'publicly accessible' (e.g., TV broadcasts, websites, artworks, newspapers) is available for anyone to consult/view. It is still protected by copyright even if there is no copyright notice. In UK law, copyright protection is automatic and does not require a copyright statement, although it is always good practice to provide one. It is necessary to check the terms and conditions of use to find out exactly how the material may be reused etc.* *If you answered YES to question 1, be aware that you may need to consider other ethics codes. For example, when conducting Internet research, consult the code of the Association of Internet Researchers; for educational research, consult the Code of Ethics of the British Educational Research Association.*	**YES**
3. If you answered NO to question 2, do you have explicit permission to use these materials as data? *If YES, please show evidence to your supervisor.*	

Question	Yes/No
4. If you answered NO to question 3, is it because: A. you have not yet asked permission B. you have asked and not yet received and answer C. you have asked and been refused access. *Note: You will only be able to start the research when you have been granted permission to use the specified material.*	**A/B/C**

SECTION B

HEALTH AND SAFETY RISK ASSESSMENT FOR THE RESEARCHER

1. **Does this research project require a health and safety risk assessment for the procedures to be used?** (Discuss this with your supervisor)

   ☐ Yes
   ☑ No

   If **YES** the completed Health and Safety Risk Assessment form should be attached. A standard risk assessment form can be generated through the Awaken system (https://shu.awaken-be.com). Alternatively if you require more specific risk assessment, e.g. a COSHH, attach that instead.

2. **Will the data be collected fully online (no face-to-face contact with participants)?**

   ☑ Yes (See the safety guidance for online research[2] and **go to question 7b**)
   ☐ No (Go to question 3)

3. **Will the proposed data collection take place on campus?**

   ☐ Yes (Please answer questions 5 to 8)
   ☑ No (Please complete all questions and consult with your supervisor))

4. **Where will the data collection take place?**
   (Tick as many as apply if data collection will take place in multiple venues)

	Location	Please specify
☐	Researcher's Residence	✓
☐	Participant's Residence	
☐	Education Establishment	

[2] Safety guidance for online research includes information on how to set up online surveys and/or conduct online interviews/focus groups. These guidelines can be found in BB. Please check with your supervisor/module leader.

	Location	Please specify
☐	Other e.g., business/voluntary organisation, public venue	
☐	Outside UK	

5. How will you travel to and from the data collection venue?

☐ On foot ☐ By car ☐ Public Transport
☐ Other (Please specify)

Please outline how you will ensure your personal safety when travelling to and from the data collection venue.

N/A

6. How will you ensure your own personal safety whilst at the research venue?

N/A

7. Are there any potential risks to your health and wellbeing associated with either (a) the venue where the research will take place and/or (b) the research topic itself?

☐ None that I am aware of
☐ Yes (Please outline below including steps taken to minimise risk)

N/A

8. If you are carrying out research off-campus, you must ensure that each time you go out to collect data you ensure that someone you trust knows where you are going (without breaching the confidentiality of your participants), how you are getting there (preferably including your travel route), when you expect to get back, and what to do should you not return at the specified time.

Please outline here the procedure you propose using to do this.

N/A

Insurance Check

The University's standard insurance cover will not automatically cover research involving any of the following:

i) Participants under 5 years old
ii) Pregnant women
iii) 5000 or more participants
iv) Research being conducted in an overseas country
v) Research involving aircraft and offshore oil rigs
vi) Nuclear research
vii) Any trials/medical research into Covid 19

If your proposals do involve any of the above, please contact the Insurance Manager directly (fin-insurancequeries-mb@exchange.shu.ac.uk) to discuss this element of your project.

Adherence to SHU Policy and Procedures

<table>
<tr><td colspan="2">Ethics sign-off</td></tr>
<tr><td colspan="2">Personal statement</td></tr>
<tr><td colspan="2">I can confirm that:
I have read the Sheffield Hallam University Research Ethics Policy and Procedures
I agree to abide by its principles.</td></tr>
<tr><td colspan="2">Student</td></tr>
<tr><td>Name: Gideon Oluwatobi Mautin</td><td>Date:25/10/2023</td></tr>
<tr><td colspan="2">Signature: Oluwatobi Mautin</td></tr>
<tr><td colspan="2">Supervisor ethical sign-off</td></tr>
<tr><td colspan="2">I can confirm that completion of this form has not identified the need for ethical approval by the TPREC/CREC or an NHS, Social Care, or other external REC. The research will not commence until any approvals required under Sections 4 & 5 have been received and any necessary health and safety measures are in place.</td></tr>
<tr><td>Name:</td><td>Date:</td></tr>
<tr><td colspan="2">Signature:</td></tr>
<tr><td colspan="2">Independent Reviewer ethical sign off</td></tr>
<tr><td>Name:</td><td>Date:</td></tr>
<tr><td colspan="2">Signature:</td></tr>
</table>

Please ensure that you have attached all relevant documents. Your supervisor must approve them before you start data collection:

Documents	Yes	No	N/A
Research proposal if prepared previously	☐	☐	☐
Any recruitment materials (e.g., posters, letters, emails, etc.)	☐	☐	☐
Participant information sheet[3]	☐	☐	☐
Participant consent form[4]	☐	☐	☐
Details of measures to be used (e.g., questionnaires, etc.)	☐	☐	☐
Outline interview schedule / focus group schedule	☐	☐	☐
Debriefing materials	☐	☐	☐
Health and Safety Risk Assessment Form	☐	☐	☐

[3] It is mandatory to attach the Participant Information Sheet (PIS)
[4] It is mandatory to attach a Participant Consent Form, unless it is embedded in an online survey, in which case your supervisor must approve it before you start data collection

B.

Sheffield Hallam University | College of Business, Technology and Engineering

Dissertation for Computing (55-708541).

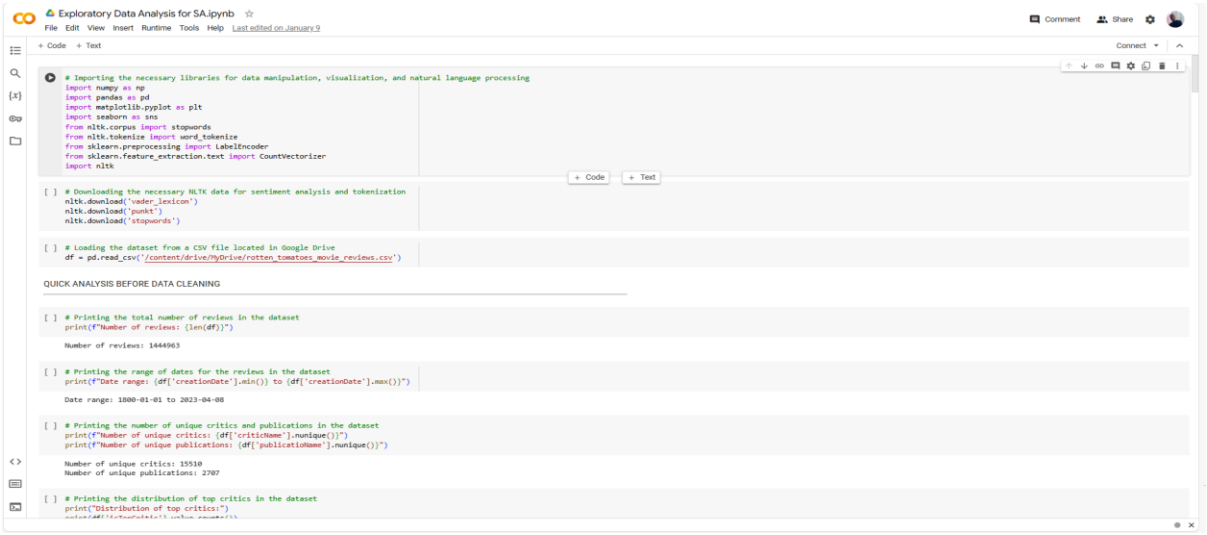
PUBLICATION PROCEDURE FORM

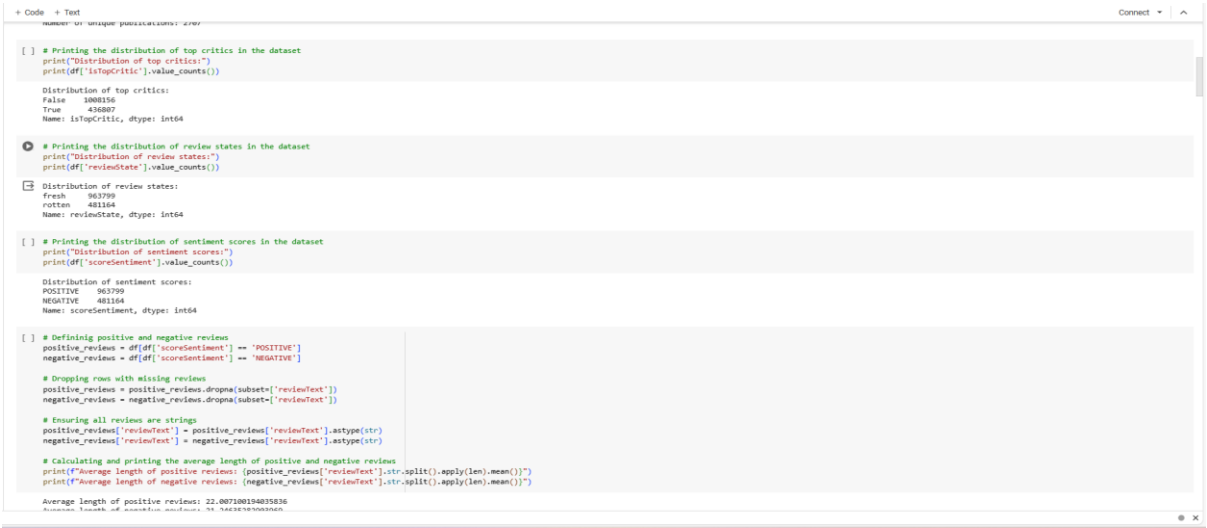
In this module, while you create your own research question or topic area, your supervisor makes a significant intellectual contribution to this work as the research progresses. Your supervisor will make the decision on whether your work merits publication based on the quality of the work you have produced. Your supervisor will co-author the paper for publication with you and your supervisor will both be listed as authors. You are required to sign the declaration below to confirm that you understand and will follow this procedure.

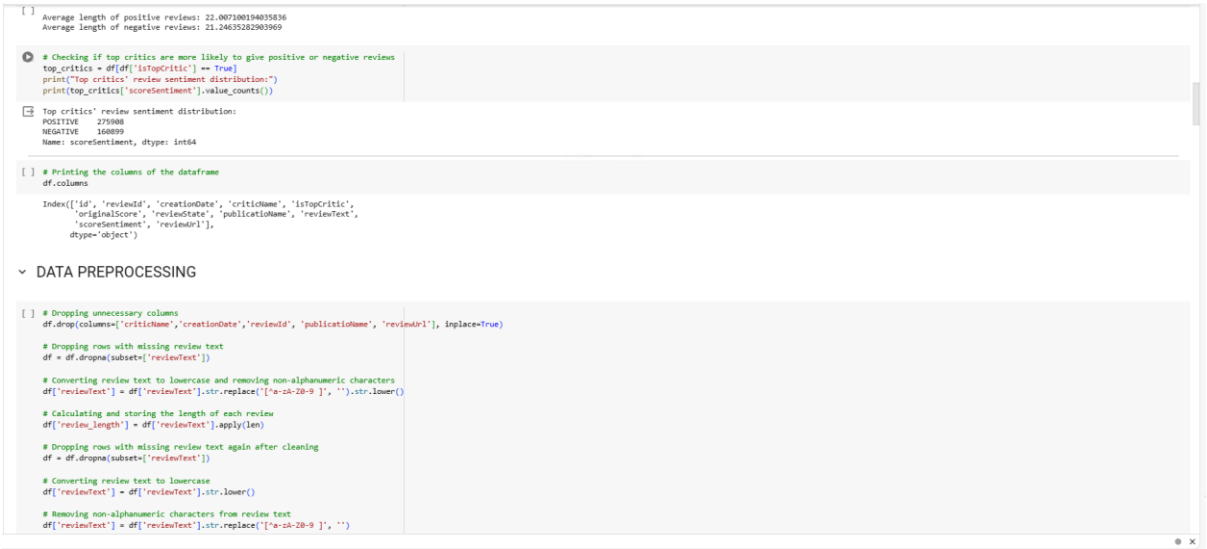
Declaration:

I OLUWATOBI GIDEON MAUTIN confirm that I understand will comply with the Publication Procedure outlined in the Module Handbook and the Blackboard Site.		
Student:	Signature oluwatobi mautin	Date 14/01/2024
Supervisor:	Signature	Date

C.


Exploratory Data Analysis for SA.ipynb
File Edit View Insert Runtime Tools Help Last edited on January 9
Comment
Share
Connect
+ Code
+ Text
# Importing the necessary libraries for data manipulation, visualization, and natural language processing
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from sklearn.preprocessing import LabelEncoder
from sklearn.feature_extraction.text import CountVectorizer
import nltk
# Downloading the necessary NLTK data for sentiment analysis and tokenization
nltk.download('vader_lexicon')
nltk.download('punkt')
nltk.download('stopwords')
# Loading the dataset from a CSV file located in Google Drive
df = pd.read_csv('/content/drive/MyDrive/rotten_tomatoes_movie_reviews.csv')
QUICK ANALYSIS BEFORE DATA CLEANING
# Printing the total number of reviews in the dataset
print(f"Number of reviews: {len(df)}")
Number of reviews: 1444963
# Printing the range of dates for the reviews in the dataset
print(f"Date range: {df['creationDate'].min()} to {df['creationDate'].max()}")
Date range: 1800-01-01 to 2023-04-08
# Printing the number of unique critics and publications in the dataset
print(f"Number of unique critics: {df['criticName'].nunique()}")
print(f"Number of unique publications: {df['publicationName'].nunique()}")
Number of unique critics: 15510
Number of unique publications: 2707
# Printing the distribution of top critics in the dataset
print("Distribution of top critics:")


+ Code
+ Text
Connect
# Printing the distribution of top critics in the dataset
print("Distribution of top critics:")
print(df['isTopCritic'].value_counts())
Distribution of top critics:
False 1008156
True 436807
Name: isTopCritic, dtype: int64
# Printing the distribution of review states in the dataset
print("Distribution of review states:")
print(df['reviewState'].value_counts())
Distribution of review states:
fresh 963799
rotten 481164
Name: reviewState, dtype: int64
# Printing the distribution of sentiment scores in the dataset
print("Distribution of sentiment scores:")
print(df['scoreSentiment'].value_counts())
Distribution of sentiment scores:
POSITIVE 963799
NEGATIVE 481164
Name: scoreSentiment, dtype: int64
# Defininig positive and negative reviews
positive_reviews = df[df['scoreSentiment'] == 'POSITIVE']
negative_reviews = df[df['scoreSentiment'] == 'NEGATIVE']
# Dropping rows with missing reviews
positive_reviews = positive_reviews.dropna(subset=['reviewText'])
negative_reviews = negative_reviews.dropna(subset=['reviewText'])
# Ensuring all reviews are strings
positive_reviews['reviewText'] = positive_reviews['reviewText'].astype(str)
negative_reviews['reviewText'] = negative_reviews['reviewText'].astype(str)
# Calculating and printing the average length of positive and negative reviews
print(f"Average length of positive reviews: {positive_reviews['reviewText'].str.split().apply(len).mean()}")
print(f"Average length of negative reviews: {negative_reviews['reviewText'].str.split().apply(len).mean()}")
Average length of positive reviews: 22.007100194035836


Average length of positive reviews: 22.007100194035836
Average length of negative reviews: 21.24635282903969
# Checking if top critics are more likely to give positive or negative reviews
top_critics = df[df['isTopCritic'] == True]
print("Top critics' review sentiment distribution:")
print(top_critics['scoreSentiment'].value_counts())
Top critics' review sentiment distribution:
POSITIVE 275908
NEGATIVE 160899
Name: scoreSentiment, dtype: int64
Printing the columns of the dataframe
df.columns
Index(['id', 'reviewId', 'creationDate', 'criticName', 'isTopCritic',
'originalScore', 'reviewState', 'publicationName', 'reviewText',
'scoreSentiment', 'reviewUrl'],
dtype='object')
DATA PREPROCESSING
# Dropping unnecessary columns
df.drop(columns=['criticName','creationDate','reviewId', 'publicationName', 'reviewUrl'], inplace=True)
# Dropping rows with missing review text
df = df.dropna(subset=['reviewText'])
# Converting review text to lowercase and removing non-alphanumeric characters
df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9]', '').str.lower()
Calculating and storing the length of each review
df['review_length'] = df['reviewText'].apply(len)
# Dropping rows with missing review text again after cleaning
df = df.dropna(subset=['reviewText'])
# Converting review text to lowercase
df['reviewText'] = df['reviewText'].str.lower()
Removing non-alphanumeric characters from review text
df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9]', '')

```
# Dropping rows with missing review text again after cleaning
df = df.dropna(subset=['reviewText'])

# Converting review text to lowercase
df['reviewText'] = df['reviewText'].str.lower()

# Removing non-alphanumeric characters from review text
df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '')

# Tokenizing the review text
df['reviewText'] = df['reviewText'].apply(word_tokenize)

# Defining the list of stop words
stop_words = set(stopwords.words('english'))

# Removing stop words from the tokenized review text
df['reviewText'] = df['reviewText'].apply(lambda x: [word for word in x if word not in stop_words])

# Joining the tokenized words back into a single string for each review
df['reviewText'] = df['reviewText'].apply(' '.join)
```

```
<ipython-input-53-a002beb664b7>:8: FutureWarning: The default value of regex will change from True to False in a future version.
  df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '').str.lower()
<ipython-input-53-a002beb664b7>:20: FutureWarning: The default value of regex will change from True to False in a future version.
  df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '')
```

```
# Printing the columns of the dataframe
df.columns
```

```
Index(['id', 'isTopCritic', 'originalScore', 'reviewState', 'reviewText',
       'scoreSentiment', 'review_length'],
      dtype='object')
```

New Section

```
# Encoding the sentiment scores using label encoding
encoder = LabelEncoder()
df['scoreSentiment'] = encoder.fit_transform(df['scoreSentiment'])
```

1. Descriptive Statistics:

```
# Printing the number of reviews and the average review length
```


```
# Encoding the sentiment scores using label encoding
encoder = LabelEncoder()
df['scoreSentiment'] = encoder.fit_transform(df['scoreSentiment'])
```

1. Descriptive Statistics:


```
# Printing the number of reviews and the average review length
num_reviews = df['reviewText'].count()
print(f"Number of reviews: {num_reviews}")
average_review_length = df['reviewText'].str.len().mean()
print(f"Average review length: {average_review_length}")
```

```
Number of reviews: 1375738
Average review length: 92.48325989396237
```

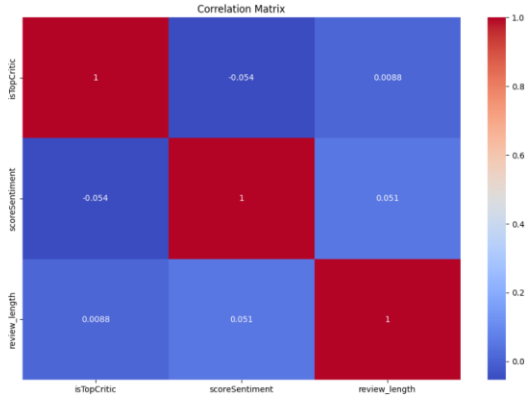
2.1. Visualizations:

```
# Importing necessary libraries for visualizations
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Calculating the correlation matrix
correlation_matrix = df.corr()

# Focusing on the correlations of the target variable
target_correlation = correlation_matrix['scoreSentiment']

# Plotting the correlation matrix
plt.figure(figsize=(12, 8))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm')
plt.title('Correlation Matrix')
plt.show()
```

```
<ipython-input-9-146516f1447d>:6: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  correlation_matrix = df.corr()
```


```
plt.show()
```

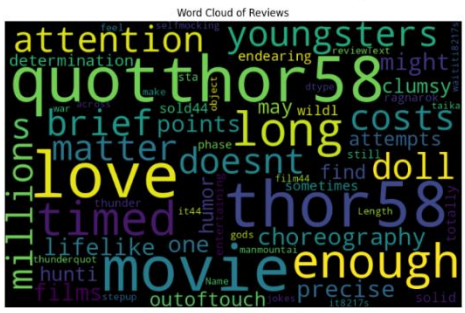
```
<ipython-input-9-146516f1447d>:6: FutureWarning: The default value of numeric_only in DataFrame.corr is deprecated. In a future version, it will default to False. Select only valid columns or specify the value of numeric_only to silence this warning.
  correlation_matrix = df.corr()
```


```
[ ]
```

```
# Plotting a word cloud
from wordcloud import WordCloud
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(str(df['reviewText']))
```

```
# Plotting a word cloud
from wordcloud import WordCloud
wordcloud = WordCloud(width=800, height=500, random_state=21, max_font_size=110).generate(str(df['reviewText']))
plt.figure(figsize=(10, 7))
plt.imshow(wordcloud, interpolation="bilinear")
plt.axis('off')
plt.title('Word Cloud of Reviews')
plt.show()
```



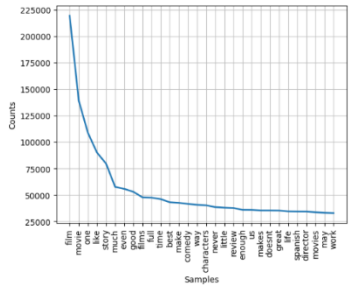
The size of each word indicates its frequency or importance with larger words appearing more frequently in the text data.

Words like "love", "movie", "enough", "youngsters", "clumsy", "timed", and "precise" are prominently displayed, suggesting they are commonly used in the reviews.

2.2. Word Frequency Distribution:

2.2. Word Frequency Distribution:

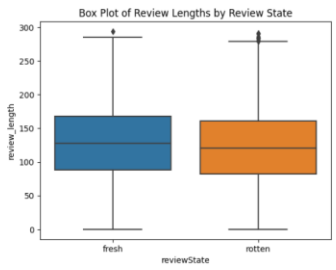
```
# Tokenizing the review text for the entire dataframe
tokens = word_tokenize(' '.join(df['reviewText']))

# Calculating word frequency for the entire dataframe
fdist = FreqDist(tokens)

# Plotting the frequency of the 30 most common words in the reviews
fdist.plot(30, cumulative=False)
plt.show()
```



It shows the frequency of different words in the movie reviews. The x-axis represents the words and the y-axis represents their counts. Words like "movie", "film", "love", "work", etc. appear to be quite frequent in the reviews

```
[ ] # Plotting the box plot of review lengths by review state
```

```
# Plotting the box plot of review lengths by review state
sns.boxplot(x='reviewState', y='review_length', data=df)
plt.title('Box Plot of Review Lengths by Review State')
plt.show()
```



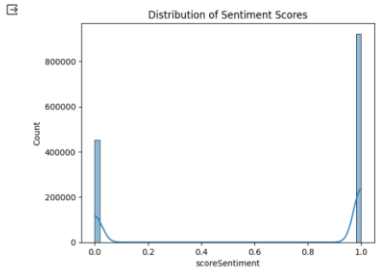
The box plot shows the distribution of review lengths for both 'fresh' and 'rotten' reviews.

Fresh Reviews: The box plot for "fresh" reviews shows that the median length (the line inside the box) is around 125. The box itself represents the interquartile range (IQR), which is the range within which the central 50% of the review lengths fall. For "fresh" reviews, the IQR extends from approximately 100 to 150. There is an outlier near 275, which means there is at least one "fresh" review significantly longer than the others.

Rotten Reviews: The box plot for "rotten" reviews also shows a median length around 125. However, the IQR is slightly broader, ranging from about 100 to 175. This suggests that the lengths of "rotten" reviews vary more than those of "fresh" reviews.

```
[ ] # Plotting the histogram of sentiment scores
    sns.histplot(df['scoreSentiment'], kde=True)
    plt.title('Distribution of Sentiment Scores')
    plt.show()
```

```
# Plotting the histogram of sentiment scores
sns.histplot(df['scoreSentiment'], kde=True)
plt.title('Distribution of Sentiment Scores')
plt.show()
```

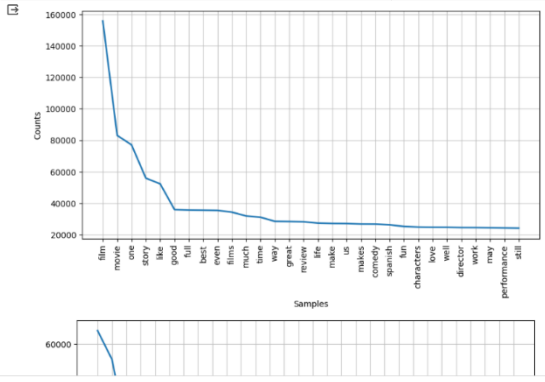


From the histogram, it appears that there are significantly more reviews with a sentiment score of 1 (positive) than reviews with a sentiment score of 0 (negative). This suggest that the dataset contains more positive reviews than negative ones.

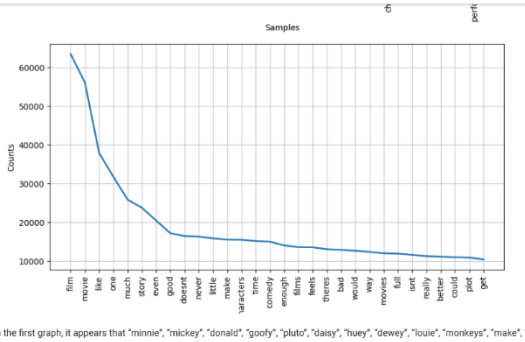
```
[ ] # Separating the reviews into positive and negative based on the sentiment score
    positive_reviews = df[df['scoreSentiment'] == 1]  # Assuming 1 is for positive sentiment
    negative_reviews = df[df['scoreSentiment'] == 0]  # Assuming 0 is for negative sentiment

    # Tokenizing the reviews for positive and negative reviews
    positive_tokens = word_tokenize(' '.join(positive_reviews['reviewText']))
    negative_tokens = word_tokenize(' '.join(negative_reviews['reviewText']))

    # Calculating word frequency for positive and negative reviews
    positive_fdist = FreqDist(positive_tokens)
    negative_fdist = FreqDist(negative_tokens)
```

+ Code + Text Connect

```
# Calculating word frequency for positive and negative reviews
positive_fdist = FreqDist(positive_tokens)
negative_fdist = FreqDist(negative_tokens)

# Plotting word frequency for positive and negative reviews
plt.figure(figsize=(10, 5))
positive_fdist.plot(30, cumulative=False)
plt.show()

plt.figure(figsize=(10, 5))
negative_fdist.plot(30, cumulative=False)
plt.show()
```



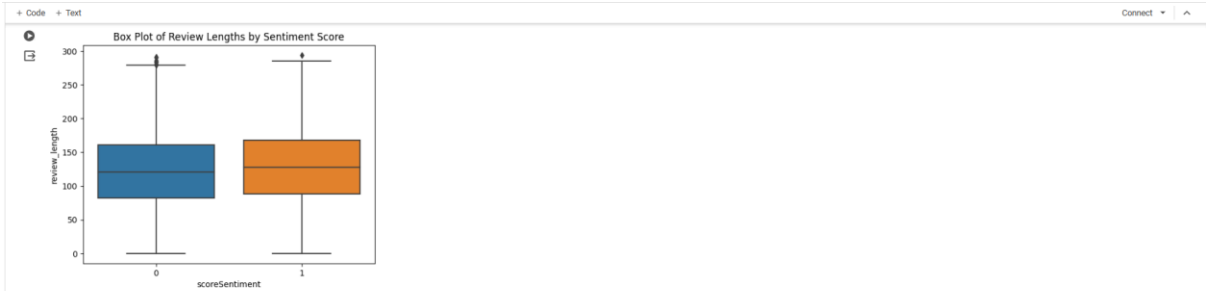
+ Code + Text Connect



From the first graph, it appears that "minnie", "mickey", "donald", "goofy", "pluto", "daisy", "huey", "dewey", "louie", "monkeys", "make", "chain", "daisy", "director", "firework", "performance", "skill" are some of the most frequently occurring words in the reviews.

From the second graph, it appears that "time", "more", "once", "second", "depress", "like", "change", "enough", "comes", "this", "would", "mean", "really", "could", "get" are some of the most frequently occurring words in the negative reviews.

```
[ ] # Plotting the box plot of review lengths by sentiment score
    sns.boxplot(x='scoreSentiment', y='review_length', data=df)
    plt.title('Box Plot of Review Lengths by Sentiment Score')
    plt.show()
```



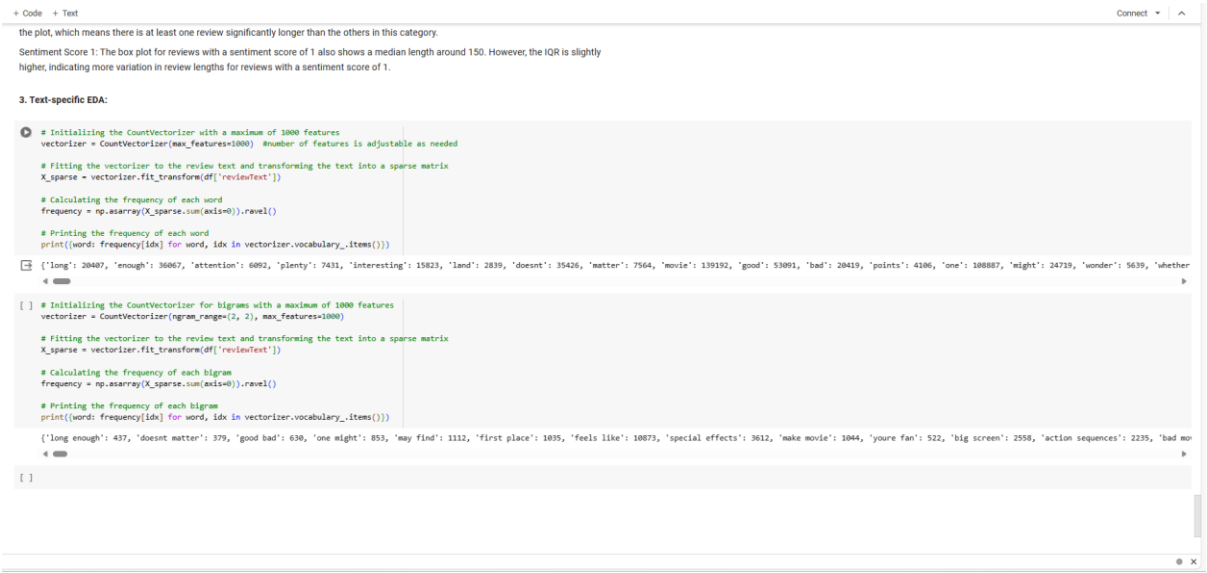
The box plot visualizes the distribution of review lengths for two sentiment scores: 0 and 1.

Sentiment Score 0: The box plot for reviews with a sentiment score of 0 shows that the median length (the line inside the box) is around 150. The box itself represents the interquartile range (IQR), which is the range within which the central 50% of the review lengths fall. For reviews with a sentiment score of 0, the IQR extends from approximately 100 to 200. There is an outlier indicated by a diamond shape near the top of the plot, which means there is at least one review significantly longer than the others in this category.

Sentiment Score 1: The box plot for reviews with a sentiment score of 1 also shows a median length around 150. However, the IQR is slightly higher, indicating more variation in review lengths for reviews with a sentiment score of 1.

3. Text-specific EDA:

```
[ ] # Initializing the CountVectorizer with a maximum of 1000 features
    vectorizer = CountVectorizer(max_features=1000)  #number of features is adjustable as needed

    # Fitting the vectorizer to the review text and transforming the text into a sparse matrix
    X_sparse = vectorizer.fit_transform(df['reviewText'])

    # Calculating the frequency of each word
```



the plot, which means there is at least one review significantly longer than the others in this category.

Sentiment Score 1: The box plot for reviews with a sentiment score of 1 also shows a median length around 150. However, the IQR is slightly higher, indicating more variation in review lengths for reviews with a sentiment score of 1.

3. Text-specific EDA:

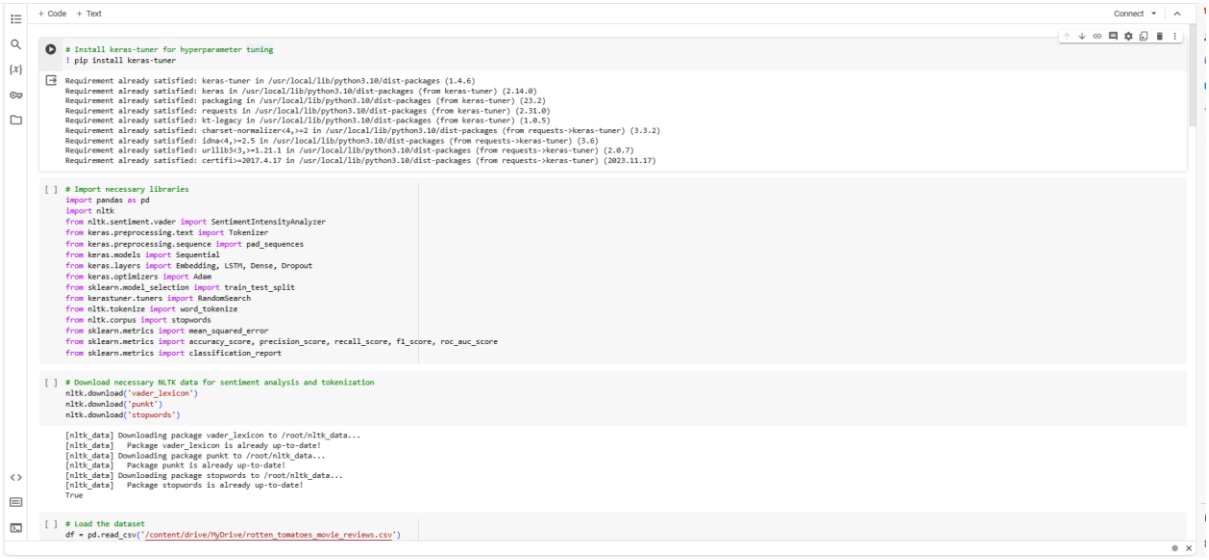
```
# Initializing the CountVectorizer with a maximum of 1000 features
vectorizer = CountVectorizer(max_features=1000)  #number of features is adjustable as needed

# Fitting the vectorizer to the review text and transforming the text into a sparse matrix
X_sparse = vectorizer.fit_transform(df['reviewText'])

# Calculating the frequency of each word
frequency = np.asarray(X_sparse.sum(axis=0)).ravel()

# Printing the frequency of each word
print({word: frequency[idx] for word, idx in vectorizer.vocabulary_.items()})
```

```
{'long': 20407, 'enough': 36067, 'attention': 6092, 'plenty': 7431, 'interesting': 15823, 'land': 2839, 'doesnt': 35426, 'matter': 7564, 'movie': 139192, 'good': 53091, 'bad': 20419, 'points': 4106, 'one': 108887, 'might': 24719, 'wonder': 5639, 'whether
```

```
[ ] # Initializing the CountVectorizer for bigrams with a maximum of 1000 features
    vectorizer = CountVectorizer(ngram_range=(2, 2), max_features=1000)

    # Fitting the vectorizer to the review text and transforming the text into a sparse matrix
    X_sparse = vectorizer.fit_transform(df['reviewText'])

    # Calculating the frequency of each bigram
    frequency = np.asarray(X_sparse.sum(axis=0)).ravel()

    # Printing the frequency of each bigram
    print({word: frequency[idx] for word, idx in vectorizer.vocabulary_.items()})
```

```
{'long enough': 437, 'doesnt matter': 379, 'good bad': 630, 'one might': 853, 'may find': 1112, 'first place': 1035, 'feels like': 10873, 'special effects': 3612, 'make movie': 1044, 'youre fan': 522, 'big screen': 2558, 'action sequences': 2235, 'bad mo
```

```
[ ]
```



```
# Install keras-tuner for hyperparameter tuning
! pip install keras-tuner
```

```
Requirement already satisfied: keras-tuner in /usr/local/lib/python3.10/dist-packages (1.4.6)
Requirement already satisfied: keras in /usr/local/lib/python3.10/dist-packages (from keras-tuner) (2.14.0)
Requirement already satisfied: packaging in /usr/local/lib/python3.10/dist-packages (from keras-tuner) (23.2)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-packages (from keras-tuner) (2.31.0)
Requirement already satisfied: kt-legacy in /usr/local/lib/python3.10/dist-packages (from keras-tuner) (1.0.5)
Requirement already satisfied: charset-normalizer<4,>=2 in /usr/local/lib/python3.10/dist-packages (from requests->keras-tuner) (3.3.2)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-packages (from requests->keras-tuner) (3.6)
Requirement already satisfied: urllib3<3,>=1.21.1 in /usr/local/lib/python3.10/dist-packages (from requests->keras-tuner) (2.0.7)
Requirement already satisfied: certifi>=2017.4.17 in /usr/local/lib/python3.10/dist-packages (from requests->keras-tuner) (2023.11.17)
```

```
[ ] # Import necessary libraries
    import pandas as pd
    import nltk
    from nltk.sentiment.vader import SentimentIntensityAnalyzer
    from keras.preprocessing.text import Tokenizer
    from keras.preprocessing.sequence import pad_sequences
    from keras.models import Sequential
    from keras.layers import Embedding, LSTM, Dense, Dropout
    from keras.optimizers import Adam
    from sklearn.model_selection import train_test_split
    from kerastuner.tuners import RandomSearch
    from nltk.tokenize import word_tokenize
    from nltk.corpus import stopwords
    from sklearn.metrics import mean_squared_error
    from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, roc_auc_score
    from sklearn.metrics import classification_report
```

```
[ ] # Download necessary NLTK data for sentiment analysis and tokenization
    nltk.download('vader_lexicon')
    nltk.download('punkt')
    nltk.download('stopwords')
```

```
[nltk_data] Downloading package vader_lexicon to /root/nltk_data...
[nltk_data]   Package vader_lexicon is already up-to-date!
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
True
```

```
[ ] # Load the dataset
    df = pd.read_csv('/content/drive/MyDrive/rotten_tomatoes_movie_reviews.csv')
```

```
# Load the dataset
df = pd.read_csv('/content/drive/MyDrive/rotten_tomatoes_movie_reviews.csv')

# Keep a copy of the original data
df_original = df.copy()
```

```
# Drop unnecessary columns
df.drop(columns=['criticName','creationDate','reviewId', 'publicationName', 'reviewUrl'], inplace=True)

# Drop rows with missing review text
df = df.dropna(subset=['reviewText'])

# Preprocess the review text: remove non-alphanumeric characters and convert to lowercase
df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '').str.lower()

# Calculate and store the length of each review
df['review_length'] = df['reviewText'].apply(len)

# Drop rows with missing review text again after cleaning
df = df.dropna(subset=['reviewText'])

# Convert review text to lowercase
df['reviewText'] = df['reviewText'].str.lower()

# Remove non-alphanumeric characters from review text
df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '')

# Tokenize the review text
df['reviewText'] = df['reviewText'].apply(word_tokenize)

# Define the list of stop words
stop_words = set(stopwords.words('english'))

# Remove stop words from the tokenized review text
df['reviewText'] = df['reviewText'].apply(lambda x: [word for word in x if word not in stop_words])

# Join the tokenized words back into a single string for each review
df['reviewText'] = df['reviewText'].apply(' '.join)
```

```
<ipython-input-31-c5213966a91c>:3: FutureWarning: The default value of regex will change from True to False in a future version.
  df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '').str.lower()
<ipython-input-31-c5213966a91c>:7: FutureWarning: The default value of regex will change from True to False in a future version.
  df['reviewText'] = df['reviewText'].str.replace('[^a-zA-Z0-9 ]', '')
```

```
df.columns
```

```
Index(['id', 'isTopCritic', 'originalScore', 'reviewState', 'reviewText',
       'scoreSentiment', 'review_length'],
      dtype='object')
```

+ Code + Text

converting the text data into sequences of tokens and then padding these sequences to ensure they all have the same length.

+ Code + Text

```
# Prepare the tokenizer
tokenizer = Tokenizer()
tokenizer.fit_on_texts(df['reviewText'])

# Convert the text to sequences of integers
sequences = tokenizer.texts_to_sequences(df['reviewText'])

# Pad the sequences so they all have the same length
data = pad_sequences(sequences)
```

```
# Calculate the sentiment score for each review using the Sentiment Intensity Analyzer from NLTK
sia = SentimentIntensityAnalyzer()
df['sentiment_score'] = df['reviewText'].apply(lambda text: sia.polarity_scores(text)['compound'])

# Prepare the labels: convert the sentiment scores to binary labels (1 for positive sentiment and 0 for negative sentiment
labels = df['sentiment_score']
labels = labels.apply(lambda x: 1 if x > 0 else 0)
```

```
# Split the data into training and testing sets
X_train, X_test, y_train, y_test = train_test_split(data, labels, test_size=0.2, random_state=1)
```

```
# Define the model
model = Sequential()
model.add(Embedding(input_dim=len(tokenizer.word_index)+1, output_dim=100))
model.add(LSTM(64, return_sequences=True))
model.add(LSTM(32))
model.add(Dense(1))

# Compile the model
model.compile(optimizer=Adam(), loss='mean_squared_error')

# Train the model
model.fit(X_train, y_train, epochs=10, validation_data=(X_test, y_test), batch_size=1024)
```

```
Epoch 1/10
1075/1075 [==============================] - 828s 766ms/step - loss: 0.0713 - val_loss: 0.0264
```

```
# Train the model
model.fit(X_train, y_train, epochs=10, validation_data=(X_test, y_test), batch_size=1024)
```

```
Epoch 1/10
1075/1075 [==============================] - 828s 766ms/step - loss: 0.0713 - val_loss: 0.0264
Epoch 2/10
1075/1075 [==============================] - 802s 746ms/step - loss: 0.0178 - val_loss: 0.0181
Epoch 3/10
1075/1075 [==============================] - 789s 734ms/step - loss: 0.0105 - val_loss: 0.0166
Epoch 4/10
1075/1075 [==============================] - 803s 747ms/step - loss: 0.0074 - val_loss: 0.0171
Epoch 5/10
1075/1075 [==============================] - 799s 744ms/step - loss: 0.0055 - val_loss: 0.0164
Epoch 6/10
1075/1075 [==============================] - 789s 734ms/step - loss: 0.0043 - val_loss: 0.0174
Epoch 7/10
1075/1075 [==============================] - 788s 733ms/step - loss: 0.0035 - val_loss: 0.0169
Epoch 8/10
1075/1075 [==============================] - 799s 743ms/step - loss: 0.0029 - val_loss: 0.0172
Epoch 9/10
1075/1075 [==============================] - 799s 743ms/step - loss: 0.0025 - val_loss: 0.0178
Epoch 10/10
1075/1075 [==============================] - 794s 739ms/step - loss: 0.0022 - val_loss: 0.0185
<keras.src.callbacks.History at 0x7b100a416bc0>
```

```
# Predict the sentiment scores for the test set
y_pred = model.predict(X_test)

# Print the Mean Squared Error
print(f"Mean Squared Error: {mean_squared_error(y_test, y_pred)}")
```

```
8599/8599 [==============================] - 87s 10ms/step
Mean Squared Error: 0.018480025814702144
```

```
# Predict the probabilities for the test set
y_pred_probs = model.predict(X_test)
```

```
8599/8599 [==============================] - 86s 10ms/step
```

```
# Convert probabilities into class labels
y_pred = [1 if p > 0.5 else 0 for p in y_pred_probs]
```

```
# Calculate metrics
accuracy = accuracy_score(y_test, y_pred)
precision = precision_score(y_test, y_pred)
recall = recall_score(y_test, y_pred)
```

```
[ ] # Convert probabilities into class labels
    y_pred = [1 if p > 0.5 else 0 for p in y_pred_probs]

    # Calculate metrics
    accuracy = accuracy_score(y_test, y_pred)
    precision = precision_score(y_test, y_pred)
    recall = recall_score(y_test, y_pred)
    f1 = f1_score(y_test, y_pred)
    roc_auc = roc_auc_score(y_test, y_pred_probs)

    print(f"Accuracy: {accuracy}")
    print(f"Precision: {precision}")
    print(f"Recall: {recall}")
    print(f"F1 Score: {f1}")
    print(f"ROC AUC: {roc_auc}")

    Accuracy: 0.9774957477430328
    Precision: 0.984724642367411
    Recall: 0.9783391139323343
    F1 Score: 0.9815214926050016
    ROC AUC: 0.9925194260195273

[ ] # Save the model
    model.save('/content/drive/MyDrive/sentiment_analysis_model.h5')

[ ] # Add sentiment analysis results to original dataframe
    df_original['sentiment_score'] = df['sentiment_score']
    df_original['sentiment_label'] = labels

[ ] # Group by movie and calculate the mean sentiment score
    df_grouped = df_original.groupby('id')['sentiment_score'].mean()

    # Convert the GroupBy object to a DataFrame
    df_grouped = df_grouped.reset_index()

    # Create a new column for box office success or failure
    df_grouped['box_office'] = df_grouped['sentiment_score'].apply(lambda x: 'success' if x > 0 else 'failure')

[ ] #df_original.to_csv('sentiment_analysis_results.csv'

    df_grouped.to_csv('/content/drive/MyDrive/average_sentiment_analysis_results.csv')
```