Using Sentiment Analysis to Compare Critics and Audience Movie Reviews

Mark Pacetta* **Justin Tran*** **Alex Tran***

Joshua Nguyen* **Jason Vu*** **Batool Alabduljabbar***

Abstract

This study aims to explore the disparity between online critic and audience reviews of movies through the application of sentiment analysis. By utilising sentiment analysis techniques, specifically focusing on HuggingFace with the RoBERTa model, we will analyse and compare the sentiment expressed in critic and audience reviews. The findings will provide insights into the varying perspectives and opinions of critics and audiences, shedding light on the factors contributing to the disparity. This research contributes to the understanding of the dynamic relationship between critics and audiences in the evaluation and perception of movies.

1 Introduction

Our aim of this project is to use sentiment analysis on critics' and audiences' reviews to scrape out ideas, keywords, phrases and common themes to draw out what criteria audiences base their reviews on and what critics base their reviews on. Our hypothesis is that audiences' reviews will mainly be based on entertainment and how they resonate with the movie, whereas critics judge the movie based on how well the movie is directed and written based on the screenplay and the overarching theme/message/social/political influences of the movie. Furthermore, our project

*Equal contribution and in Appendix

research assumption is that critic reviews would be less impacted by emotion which is proven to be true; however, it seems that critics are also more emotional with the negative sentiment presented through their reviews.

This report aims to address feedback received for our project proposal (Assessment Task 1) and incorporate comments from mentors and peer reviews. It will focus on concerns raised regarding the research problem, aim, research questions, literature review, and methodology. Based on the feedback, revised research questions and objectives will be presented to enhance the study's clarity and relevance.

2 Review of Feedback

2.1 Research Problem and Relevance

Reviewers noted that the research problem was inadequately defined and explained, making it challenging to determine the research contribution's relevance. However, the project proposal focuses on analysing and comparing reviews by critics and users on online rating sites, specifically addressing disparate review scores. The study aims to investigate differences and potential reasons behind these contrasting opinions, understanding the dynamics between critics and consumers. The implications are relevant for filmmakers, audiences, and online platforms.

2.2 Research Aim

The purpose of a research aim is to give the project a clear direction and purpose. Reviewers suggested that the primary purpose of the study remains unclear. However, the project proposal clearly states the aim as the analysis and comparison of reviews by critics and users on online rating sites. This sets the foundation for subsequent research questions and objectives, addressing data reliability, bias, and ethical considerations. It can be understood then that our research aim was not properly expressed in a manner that was easily understood by readers. In the revision, the aim will be more explicitly stated in order to avoid the same confusion.

2.3 Research Questions

Reviewers expressed concerns about the research questions, suggesting they focused more on data gathering and review comparison rather than proving a theory or methods. However, addressing questions related to data reliability, validity, bias, and ethical considerations is crucial for a rigorous analysis. The project proposal mentions using sentiment analysis and data visualisation as clear methodologies for analysis and interpretation.

2.4 Revised Research Questions and Objectives

Based on feedback, the following revised research questions and objectives are proposed:

- How do review scores differ between critics and users on online rating sites?

  Investigate disparities in review scores between critics and users.

Analyse factors contributing to these differences.

- What are the underlying reasons for contrasting opinions between critics and users?

  Identify potential reasons for disparities in opinion.

  Explore the impact of factors like personal preferences, biases, and motivations.

- How reliable and valid are the reviews provided by critics and users?

  Assess the reliability and validity of critic reviews.

  Evaluate the reliability and validity of user reviews.

- What are the ethical considerations associated with using reviews from online rating sites?

  Examine the ethical implications of using critic and user reviews.

  Discuss the potential consequences of biassed or manipulated reviews.

- How can sentiment analysis be applied to analyse and compare reviews from critics and users?

  Develop a sentiment analysis methodology tailored to study objectives.

  Apply sentiment analysis techniques to compare sentiment in critic and user reviews.

2.5 Literature Review

Research gaps were highlighted as being poorly explored and defined within our initial project proposal. The importance of identifying research gaps is to ensure that our own research is thorough and does not suffer from the shortcomings of other papers within the same domain. Due to the high availability and ease of access to information within this domain, there have already been a large number of gaps taken into consideration by these papers and therefore it is extremely difficult to find new gaps within the scope of this project. We also acknowledge the lack of proper citations in the literature review. Proper references will be included in the revised version to trace the sources of information and understand specific methodologies used.

The suggestion to reorder the literature review sections is appreciated, and the revised version will follow a more conventional approach for improved flow and organisation. It will also include a more extensive discussion of related work, methodologies, and approaches used in sentiment analysis techniques, providing a thorough understanding of existing techniques and their applicability.

2.6 Methodology

Feedback indicated that the methodology lacks a theoretical basis and requires citations and arguments supporting the chosen methods. The revised version will include appropriate citations and arguments to support the rationale behind the methodology, ensuring a stronger theoretical foundation. Visual aids such as images and charts have been incorporated to enhance the clarity of the project plan and contribution.

One reviewer stated that the lack of an explicit mention of the specific tools does not clearly communicate the specifics of the method. This also leads to another issue where specific reasoning can not be attributed to the planned techniques or tools. This has been addressed by deciding on a specific set of tools and techniques in order to achieve the research objectives set in this project. With a more concrete plan, it will be possible to effectively utilise the available resources and therefore provide the reasoning behind the selected methodology.

On the topic of web scraping, there was a concern about ethics due to the nature of the collection methods. In the case of this project, the approach was to use an API that specifically collects user-generated data from review websites that are publicly available and accessible. The review data posted here should be with the intention and understanding that their opinions will be accessed and viewed by many people on the internet. No other data will be collected using scraping and should therefore be able to be classified as ethical.

2.7 Conclusion

This report has addressed feedback for the project proposal, acknowledging the need for greater clarity in certain areas. The proposal outlines a relevant research problem, aim, research questions, and objectives. By addressing disparities between critic and user reviews, considering data reliability and validity, handling bias, and discussing potential reasons for differing opinions, the proposed project can contribute to understanding online review platforms' dynamics and their impact on film, product, or service perception.

By incorporating feedback, particularly in the literature review and methodology, we believe the revised project proposal will provide a solid foundation for our research. We appreciate the valuable suggestions provided and look forward to refining our work based on this feedback.

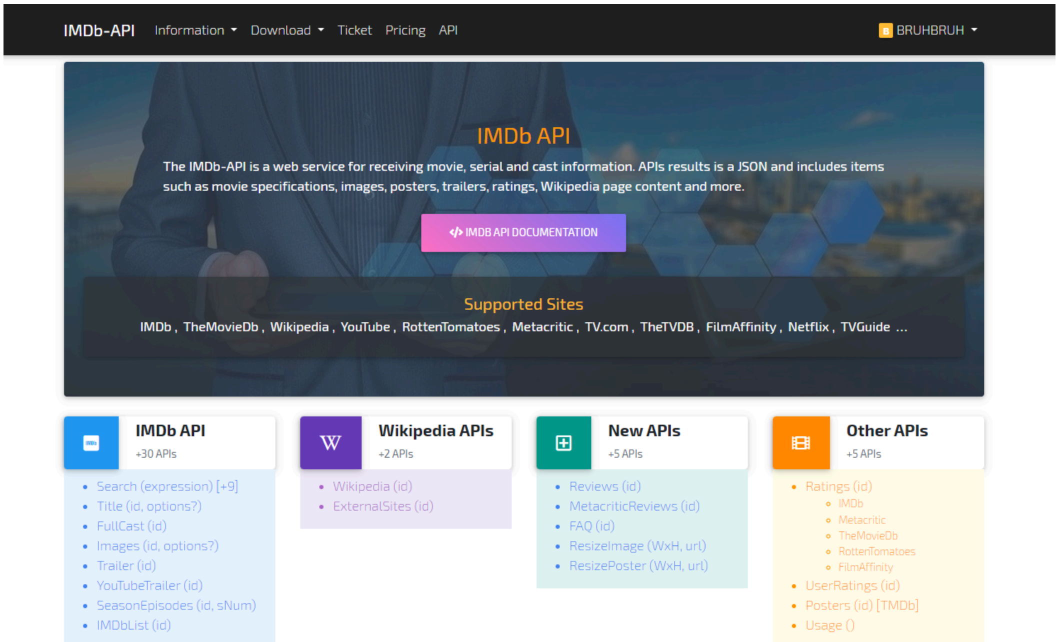
3 Design of Research

3.1 Web scraping

The first required step is to retrieve data, in this case, information such as movies and reviews are needed.

Where will this data be retrieved from?

A free Movie API was found that allowed for the retrieval of data such as movies and reviews from the IMDB database and after reading the documentation, it seems the following API calls will be required [1].



https://imdb-api.com/en/API/IMDbList/k_pjzm74qs/ls004285275	Retrieves the list of movies from a user-defined list that is denoted by its list code.
https://imdb-api.com/en/API/Ratings/k_6kz6050s/${movie.ID}	Retrieves all associated reviews from popular websites such as Rotten Tomatoes and Metacritic.
https://imdb-api.com/en/API/MetacriticReviews/k_9ihcmeyc/${movie.ID}	Fetches critic reviews from the site Metacritic.
https://imdb-api.com/en/API/Reviews/k_pjzm74qs/${movie.ID}	Fetches audience reviews from IMDB

Data that was determined to be needed including; Movie Title, Movie Year, Audience Score, Critic Score, Review Content and Review Rating.

Thus the plan for data retrieval includes first calling data, filtering out that data so that only movies with high disparities in audience and critic scores are gathered, organising that data in a readable format and only collecting properties that matter, such as the data mentioned above, and then converting the JSON format into CSV format so that the other group members are able to use the collected data. As well as this, since developing the code to write into a JSON file will take up more time than it saves, I will instead be printing out to the console and then copying and pasting the object into the file instead since it is more applicable for this situation.

1. Using the free IMDB API read up on documentation and determine how to utilise the API. Using a simple HTML, CSS and JS template, develop a script that runs in the browser to first call API → which will then convert the response into a readable JSON file using async and await.

```
// get list of top movies that critics loved more than audiences
const response = await fetch('https://imdb-api.com/en/API/IMDbList/k_6kz6050s/ls087388977');
const data = await response.json();
```

2. Retrieve Movies from the JSON object and filter out the data that is needed by mapping each object to a new object with only the selected properties, filtering out the unnecessary information. Note IMDB rating is multiplied by 10 since the IMDB rating is on a 10-point system e.g. 5/10, while Rotten Tomatoes rating is based on a 100-point system e.g. 60/100.

```
// Extract movie objects from the API response
  const movies = data.items.map(movie => ({
    ID: movie.id,
    Movie_Title: movie.title,
    Movie_Year: movie.year,
    audienceScore: (movie.imDbRating*10) // Use IMDb rating as audience score
  }));
```

3. For each corresponding movie, retrieve the Rotten Tomatoes score using the 'reviews query' and base that score as the critic score, updating the corresponding movie object with a critic score. Fetch the reviews using another API call → set the criticScore property of the movie object to the Rotten Tomatoes score retrieved

```
for (var movie of movies) {
    try {
      const response = await fetch(`https://imdb-
api.com/en/API/MetacriticReviews/k_9ihcmeyc/${movie.ID}`); //fetch metacritic reviews as critic reviews
      const data = await response.json();
      const reviewsObj = data.items.map(review => ({
        reviewContent: review.content,
        reviewRating: review.rate
      }));;
```

4. For each corresponding movie, retrieve the reviews using the Metacritic query from the IMDB API. Map the response JSON to a list of review objects that only store the review content and the review rating.

```
for (const movie of movies) {
    try {
      //fetch critic ratings for each movie
      const response = await fetch(`https://imdb-api.com/en/API/Ratings/k_6kz6050s/${movie.ID}`);
      const data = await response.json();
      movie.criticScore = data.rottenTomatoes //use rotten tomatoes as critic score

    } catch (error) {
      console.error(`Error fetching reviews for ${movie.Movie_Title}:`, error);
    }
  }
```

5. For each review, filter out the unnecessary data.
6. For each review, copy over the corresponding movie data so that each JSON object is a different review. Thus each JSON object is a different review but can have the same movie. The for loop terminates at 10 to limit each movie to 10 reviews.

```
for (i = 0; i <= 10; i++) { //limit to 10 reviews
       var objAdded = {...movie}//copy movie object
       objAdded.reviewContent = reviewsObj[i].reviewContent
       objAdded.reviewRating = reviewsObj[i].reviewRating
       //add both review values to movie
       //add new obj to list of objects
       newOBJ.push(objAdded)

     }
```

7. Convert the JSON file into a CSV file using an online tool; download the file and send it to the other group members for ease of use in other applications.

3.2 Sentiment Analysis

The sentiment analysis is the main essence of our project. It is the what and how the movie reviews are analysed and visualised. We need to label each movie review gathered from the web scraping dataset and append sentiment scores and emotions so that we can further analyse critics and audience reviews and draw conclusions on the discrepancy between the two groups. For background information, sentiment analysis represents the model's confidence or likelihood that a text is negative, neutral or positive through text classification. For example, if the predicted probabilities are [0.2, 0.5, 0.3] for neg, neu, and pos respectively, it means the model assigns a 20% probability to negative sentiment, 50% probability to neutral sentiment, and a 30% probability to positive sentiment for a given text input. Thus this text is labelled as neutral since it has the highest sentiment score.

The sentiment analysis process will now be discussed and we will explain step by step on how we will perform sentiment analysis. The full code and steps are in our Kaggle Notebook.

Link to the "Movie Reviews Sentiment Analysis" Kaggle Notebook [2]:

https://www.kaggle.com/code/hickman2049/movie-reviews-sentiment-analysis

1. Data preprocessing is required as the first step to clean the dataset web scraped. We had a look at the CSV file and there were some empty cells and unnecessary columns. So we dropped those using both Excel and Python.

```
df['ID'] = range(1, len(df) + 1)
df.dropna(subset=['ID', 'reviewRating'], inplace=True)
df.dropna(subset=['ID', 'criticScore'], inplace=True)
df['ID'] = range(1, len(df) + 1)
```

2. Import Necessary Modules such as the libraries needed for the sentiment analysis environment. Including numPy, pandas, matplotlib, NLTK, and seaborn. This has been done automatically by Kaggle when you create a new notebook.

Import Necessary Modules

```
In [1]:
# This Python 3 environment comes with many helpful analytics libraries installed
# It is defined by the kaggle/python Docker image: https://github.com/kaggle/docker-python
# For example, here's several helpful packages to load

import numpy as np # linear algebra
import pandas as pd # data processing, CSV file I/O (e.g. pd.read_csv)
import matplotlib.pyplot as plt # for the graphs
import seaborn as sns

plt.style.use('ggplot')

import nltk
# Input data files are available in the read-only "../input/" directory
# For example, running this (by clicking run or pressing Shift+Enter) will list all files under the in
put directory

import os
for dirname, _, filenames in os.walk('/kaggle/input'):
    for filename in filenames:
        print(os.path.join(dirname, filename))

# You can write up to 20GB to the current directory (/kaggle/working/) that gets preserved as output w
hen you create a version using "Save & Run All"
# You can also write temporary files to /kaggle/temp/, but they won't be saved outside of the current
session
```

3. Import the web scraped datasets into the Kaggle notebook and read in the data using pandas, converting the CSV dataset into a data frame. Here we called the data frame of the audience review dataset as df.

Read In Data

```
In [2]:
# Read in data in a data frame
df = pd.read_csv('../input/updated-movie-reviews-dataset/New_Audience.csv')
```

4. Now we need to import a model so that it can perform sentiment analysis. The chosen model is Facebook AI's RoBERTa pre-trained model was proposed in *RoBERTa: A Robustly Optimized BERT Pretraining Approach by Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov* [3]. It is based on Google's BERT model released in 2018. We chose this because it is a transformer-based model which not only accounts for the words but also the context related to other words as opposed to the NLTK VADER model which only accounts for individual words. This RoBERTa-base model was trained on ~58M Twitter tweets and finetuned for sentiment analysis with the TweetEval benchmark [4].

```
In [22]:
from transformers import AutoTokenizer
from transformers import AutoModelForSequenceClassification
from scipy.special import softmax
```

```
In [23]:
MODEL = f"cardiffnlp/twitter-roberta-base-sentiment" # Model from HuggingFace
tokenizer = AutoTokenizer.from_pretrained(MODEL)
model = AutoModelForSequenceClassification.from_pretrained(MODEL)
```

5. We made a function called polarity_scores_roberta which performs sentiment analysis for an input text using the RoBERTa model. We then made a loop that ran the polarity score on the entire dataset. Some reviews are too big for the RoBERTa model so it will result in a runtime error since the max token length of an input text is 512. We will skip those.

```
In [26]:
def polarity_scores_roberta(example):
    encoded_text = tokenizer(example, return_tensors='pt')
    output = model(**encoded_text)
    scores = output[0][0].detach().numpy()
    scores = softmax(scores)
    scores_dict = {
        'neg_sentiment' : scores[0],
        'neu_sentiment' : scores[1],
        'pos_sentiment' : scores[2],
}
    return scores_dict
```

```
In [27]:
#Run the polarity score on the entire dataset, but this time for the RoBERTa model
result = {}
broke_audience_ids = []  # Use a list to store the IDs of the records that broke

for i, row in tqdm(df.iterrows(), total=len(df)):
    try: # Some reviews are too big for the RoBERTa model so it will result in runtime error. We will
skip those
        text = row['reviewContent']
        myid = row['ID']
        roberta_result = polarity_scores_roberta(text)
        result[myid] = roberta_result
    except RuntimeError:
        broke_audience_ids.append(str(myid))  # Add the ID to the set

print(f'Broke for ids: {", ".join(broke_audience_ids)}')
```

6. Once the sentiment analysis is finished, we can save the polarity scores of neg, neu and pos into the original data frame by merging the result dictionary. We can then use head() to verify that our sentiment scores have been appended to the data frame.

```
In [28]:
audience_sentiment_df = pd.DataFrame(result).T
audience_sentiment_df = audience_sentiment_df.reset_index().rename(columns={'index': 'ID'})
audience_sentiment_df = df.merge(audience_sentiment_df, how='left', on='ID')
```

```
In [29]:
audience_sentiment_df.head()
```

Out[29]:

Movie_Title	Movie_Year	audienceScore	criticScore	reviewContent	reviewRating	neg_sentiment	neu_sentiment	pos_sentiment
National Treasure	2004	6.0	4.0	Jon Turtletaub's best film must endure the cri...	8	0.007512	0.051417	0.941072
National Treasure	2004	6.0	4.0	Very fun and entertaining movie. Some good act...	7	0.004976	0.022720	0.972304
National Treasure	2004	6.0	4.0	Director Jon Turtelaub and the usually enterta...	7	0.031195	0.467255	0.501550
National Treasure	2004	6.0	4.0	"National Treasure" is a thriller that was obv...	7	0.014135	0.128498	0.857366
National Treasure	2004	6.0	4.0	I enjoy my action/adventure movies so finally ...	8	0.001574	0.008486	0.989941

7. Now that we have finished the sentiment analysis for the audience, we can repeat steps 2-6 for the critic's dataset.
8. Now we can explore other methods of sentiment analysis beyond text classification by negative, neutral and positive scores. We will use a text classification model that can identify the emotions of input text. We will use the RoBERTa emotion model which is a fine-tuned RoBERTa model, trained on datasets with texts and associated emotions. The emotions are as follows: anger, disgust, fear, joy, neutral, sadness, and surprise. We can import the model from HuggingFace using a pipeline [5]. Then adjust the configurations, mainly to set the max_length to 512 otherwise there will be a runtime error later on.

```
In [36]:
from transformers import pipeline
```

```
In [37]:
classifier = pipeline(
    "text-classification",
    model="j-hartmann/emotion-english-distilroberta-base",
    tokenizer="j-hartmann/emotion-english-distilroberta-base",
    return_all_scores=True,
    padding=True,
    truncation=True,
    max_length=512  # Set the max_length parameter here
)
```

9. Similarly to the sentiment analysis, we can run the classifier from the pipeline for all reviews in the dataset using a loop

In [39]:
```python
audience_emotion_result = {}
broke_audience_ids = []  # List to store the IDs of the records that broke

# Define the list of emotions
emotions = ['anger', 'disgust', 'fear', 'joy', 'neutral', 'sadness', 'surprise']

for i, row in tqdm(df.iterrows(), total=len(df)):
    try:
        text = row['reviewContent']
        myid = row['ID']
        audience_roberta_emotion_result = classifier(text)[0]
        audience_emotion_result[myid] = audience_roberta_emotion_result
    except RuntimeError:
        broke_audience_ids.append(str(myid))  # Add the ID to the list

print(f'Broke for IDs: {", ".join(broke_audience_ids)}')
```

```
100% 941/941 [04:23<00:00, 2.91it/s]
```

10. Then we can append it to the data frame. Since we only want the scores and nothing else, we have to use numPy to filter out everything in the dictionary except for the scores.

In [40]:
```python
# Create a DataFrame from critics_emotion_result dictionary
audience_emotion_df = pd.DataFrame.from_dict(audience_emotion_result, orient='index', columns=emotio
ns)

# Extract only the score values from each cell
for emotion in emotions:
    audience_emotion_df[emotion] = audience_emotion_df[emotion].apply(lambda x: x['score'] if isinst
ance(x, dict) else None)

# Reset the index and rename the column
audience_emotion_df = audience_emotion_df.reset_index().rename(columns={'index': 'ID'})

# Merge with original dfc DataFrame
audience_emotion_df = audience_sentiment_df.merge(audience_emotion_df, how='left', on='ID')
```

In [41]:
```python
audience_emotion_df.head()
```

Out[41]:

g	neg_sentiment	neu_sentiment	pos_sentiment	anger	disgust	fear	joy	neutral	sadness	surprise
	0.007512	0.051417	0.941072	0.013110	0.012842	0.002603	0.705469	0.247277	0.007364	0.011335
	0.004976	0.022720	0.972304	0.001503	0.001213	0.000569	0.898903	0.028122	0.050292	0.019398
	0.031195	0.467255	0.501550	0.006374	0.007659	0.002185	0.384786	0.579179	0.006859	0.012957
	0.014135	0.128498	0.857366	0.005710	0.015070	0.001669	0.478549	0.485359	0.005966	0.007678
	0.001574	0.008486	0.989941	0.000975	0.000627	0.000250	0.989637	0.002494	0.001716	0.004301

11. We have now finished with the sentiment analysis. We will export the data frames in a CSV for audiences and critics with the sentiment and emotion scores for the next step in data visualisation.

```
In [52]:
# Exporting the final dataframe of audience after sentiment analysis to csv file
audience_emotion_df.to_csv('Sentiment_Analysis_Audience.csv', index=False)

In [53]:
# Exporting the final dataframe of critic after sentiment analysis to csv file
critic_emotion_df.to_csv('Sentiment_Analysis_Critic.csv', index=False)
```

12. Final step is data post-processing, where we clean up the CSV file and drop any empty cells. We also only want the scores with the highest emotion and its label so we achieved that through Excel:

```
=IF(AND(J2>=H2,J2>=I2), "pos", IF(AND(I2>=H2,I2>=J2), "neu", "neg"))

=IF(AND(K2>=L2,K2>=M2,K2>=N2,K2>=O2,K2>=P2,K2>=Q2), "anger",
    IF(AND(L2>=K2,L2>=M2,L2>=N2,L2>=O2,L2>=P2,L2>=Q2), "disgust",
        IF(AND(M2>=K2,M2>=L2,M2>=N2,M2>=O2,M2>=P2,M2>=Q2), "fear",
            IF(AND(N2>=K2,N2>=L2,N2>=M2,N2>=O2,N2>=P2,N2>=Q2), "joy",
                IF(AND(O2>=K2,O2>=L2,O2>=M2,O2>=N2,O2>=P2,O2>=Q2), "neutral",
                    IF(AND(P2>=K2,P2>=L2,P2>=M2,P2>=N2,P2>=O2,P2>=Q2), "sadness",
                        IF(AND(Q2>=K2,Q2>=L2,Q2>=M2,Q2>=N2,Q2>=O2,Q2>=P2), "surprise",
                            "unknown"
                        )
                    )
                )
            )
        )
    )
)
```

3.3 Data Visualisation

Once we have performed sentiment analysis on the critic and audience reviews using the RoBERTA model, we plan to represent the data in an intuitive and visually appealing manner. To showcase the disparity between the two groups, we will create bar graphs comparing the overall sentiment scores. Additionally, we will utilise pie charts to illustrate the distribution of sentiments, highlighting the proportions of positive, negative, and neutral reviews from both critics and audiences. These graphical representations will allow us to effectively communicate the differences in sentiments and provide a comprehensive visual understanding of the online movie review landscape. Example graphs we may consider are a double bar graph of average ratings, a pie graph for overall sentiment, accuracy line graph.

3.4 Analysis of Results

Once we have represented the sentiment analysis data through bar graphs, pie graphs and line graphs, we will delve into an in-depth analysis to gain deeper insights. We will examine the specific themes and topics prevalent in both positive and negative reviews. By identifying common patterns and recurring sentiments within each group, we aim to uncover the underlying reasons behind the disparity between critic and audience reviews. This detailed analysis will provide a nuanced understanding of the factors influencing the differing perspectives, enriching our overall findings and contributing to a comprehensive exploration of online movie reviews.

4 Results Analysis

4.1 Movie Rating

Average Score

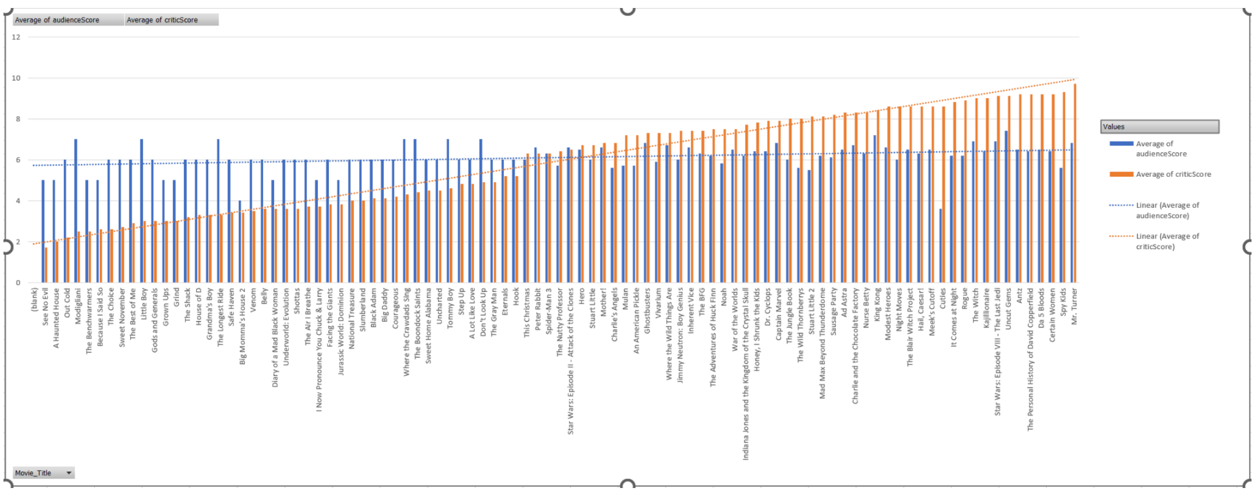


Figure i) Comparison of average critic score and average audience score. (Mark)

4.1.1 Analysis

The critics' reviews displayed a greater range of scores, as evidenced by the lowest score of 1.7 for "See No Evil" and the highest score of 9.7 for "Mr. Turner." In contrast, the audience scores exhibited a narrower range, with the lowest rating of 3.6 for "Cuties" and the highest rating of 7.4 for "Uncut Gems." Consequently, the difference in range between the two groups was notable, with the audience reviews spanning 3.6 points and the critics' reviews spanning 8 points.

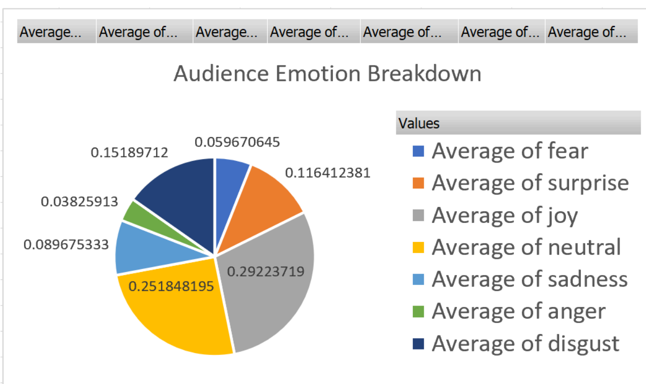
This discrepancy in range further emphasises the contrasting approaches taken by critics and audiences in evaluating films. Critics, due to their professional expertise and critical analysis, may assign scores that vary significantly based on their evaluation of various cinematic elements such as direction, acting, writing, cinematography, and overall artistic merit. This results in a wider spectrum of scores, as seen in the range of 8 points observed among the critics.

On the other hand, audiences, who are primarily seeking entertainment and personal enjoyment, tend to assign scores based on their overall satisfaction with the film. Their ratings may be influenced by factors such as relatability, genre preferences, emotional engagement, and the desire for escapism. As a result, the range of scores from the audience tends to be more compressed, spanning a narrower range of 3.6 points in this particular case.

Overall, the range disparity between critics and audiences underscores the different perspectives they bring to the reviewing process. Critics approach films from an analytical and artistic standpoint, leading to more diverse and polarising scores. Audiences, on the other hand, consider a broader range of factors, resulting in a more balanced distribution of ratings.

4.2 Sentiment Analysis

Average Emotion



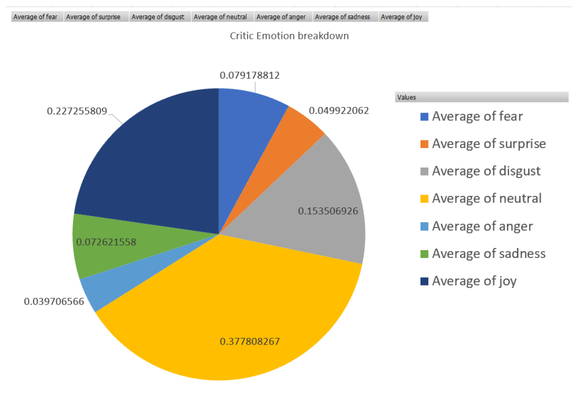


Fig ii and iii: The breakdown of the emotions displayed in reviews through sentiment analysis specific to audience and critic (Mark)

4.2.1 Analysis

The critics' reviews have the three top categories of neutral (37.78%), fear (22.72%) and disgust (15.35%). This contrasts to the audience reviews whose top three categories are joy (29.22%), disgust (15.18%) and neutral (25.18%). This links to the analysis of figure i) as the critics' reviews of fear and disgust have a large impact (53.24%) on the overall negative sentiment value, whereas with the audience reviews the negative sentiment emotions only add up to 33.92% of the overall emotions gathered from the reviews.

The main emotion for audiences, which is joy making up 29.22 per cent of all reviews, can possibly be explained by the fact that audiences tend to resonate with nostalgia since joy is linked to nostalgia. Since many movies are created around the main idea of nostalgia, many movies such as The Super Mario Bros, will tend to resonate more with audiences but are resented by critics since critics themselves don't find nostalgia a factor in their rating and instead will view the movie without nostalgia, defeating the whole point/idea of the movie. For example, for the movie Ghostbusters, in which the film tried to reinvent the movie in a new light for a new generation, we are able to see how changing the film's core as a whole can lead to differing opinions between audiences and critics. As we know audiences respond well to familiarity or nostalgia while for critics, they value ideas, concepts, and movies that reinvent ideas, change ideas, and flip norms. This can be evidently seen within the reviews for Ghostbusters 2016

A review from an audience member who gave the film a score of 3 states
"I admit from the start that I'm a massive fan of the original GHOSTBUSTERS, although the sequel isn't so hot. Both films look like absolute classics compared to this ill-advised reboot of the series, recasting with an all-female ghostbusting team. Although the film looks similar to the original movies, it feels absolutely NOTHING like them. Gone is the character interplay and genuinely funny and droll performances that could be enjoyed by kids and adults alike; it's replaced by the kind of modern-day sarcasm and idiocy that I always end up hating. The new cast members are all unlikeable, and the likes of Kristen Wiig and Melissa McCarthy seem to be competing as to who can be the most irritating of the lot. Yes, the new-fangled CGI effects are very colourful and pretty, but they don't have any of the character and finesse of the old effects, which you knew were hard to do, thus all the more impressive. The plot is as predictable as they come, and the only person who raised the occasional smile was Chris Hemsworth. I'm afraid the new GHOSTBUSTERS should never have been made."

It is established that from the beginning that this audience member is a fan of the original Ghostbusters which may have prompted them to watch the movie in the first place, however, they state that "*Although the film looks similar to the original movies, it feels absolutely NOTHING like them. Gone is the character interplay and genuinely funny and droll performances"* indicating that while this is a reboot of an old movie and should feel familiar it does not, "*modern-day sarcasm and idiocy that I always end up hating"* which is due to the film's direction and how it strives to update ghostbusters for a new audiences but ends up, causing to film to fall far from its origins and thus it is evidently shown that challenging familiarity and nostalgia can have negative effects on audiences which sometimes, they resonate strongly with.

On another note, critics' reviews evidently were harsher with their review rating metrics as presented by the analysis of the sentiment value results. It should be noted that although the critics tended to have a higher range of negative sentiment emotions within their reviews they also had more neutral sentiment emotions with 37.78% compared to the 25.18% of the audience reviews. Our project research assumption is that critic reviews would be less impacted by emotion which is proven to be true however it seems that critics are also more emotional with the negative sentiment presented through their reviews.

The critics' sentiment analysis also presents the large percentage of the fear emotion represented in their reviews. This links to the hypothesis that critics do not rate horror movies highly. As horror movies are stereotypically scary this could be a resulting linkage to why horror movies are presented as low ratings by critics. Many critics may not be big fans of horror due to the grotesque and over the top gore presented within them. On top of this, many critics prefer the positive aspects of score, cinematography and production design. Horror movies tend to have low budgets and therefore score less on these categories. Some examples of this are "See No Evil" which has a budget of $8 million USD and "A Haunted House" which has a budget of $2.5 million USD. The average film budget is considered to be $ 50 million - $ 100 million. The horror-themed movies selected do not even reach a fifth of the average movie budget and therefore must lack in some aspect of cinematography.

4.3 Film Theme

Average Score

Drama

- Mr. Turner - 9.7
- Certain Women - 9.2
- Da 5 Bloods - 9.2
- The Personal History of David Copperfield - 9.2

Childrens films

- Antz - 9.2
- Spy Kids - 9.3

Horror tended to score badly

- See No Evil - 1.7
- A Haunted House - 2

Comedy Tended to score badly

- Benchwarmers - 2.5
- Out Cold - 2.2
- Because I Said So - 2.6

4.3.1 Analysis

Through a small snapshot of the 100 movies used within the project, we were able to deduce that Real life drama and children's movies tended to score well with the critic reviews whereas comedy and horror movies tended to get lower scores by critic reviews. This may be due to the fact that the enjoyment of critics and audiences is inherently different. For critics, they primarily watch movies not for the enjoyment, but rather to critique the work, to explore the movie as a piece of work, analysing all its intricate details. This means that movies such as comedy or horror which aren't groundbreaking and rather are quite formulaic do poorly with audiences since they're mostly the same, while for audiences, its mainly the opposite, as they watch movies primarily for enjoyment and do not mind if the content of the movie just entails 'mindless action'. This notion can be further exemplified when specific reviews are explored.

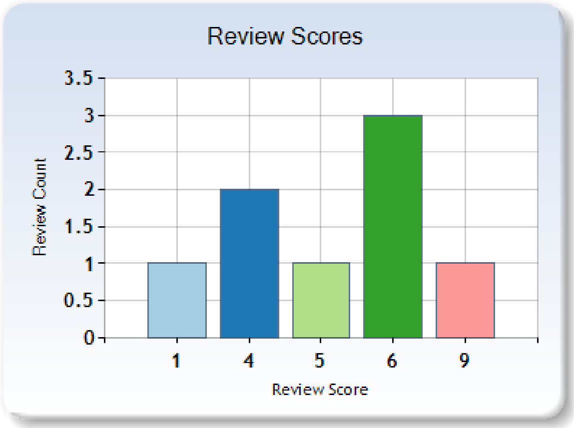
For the movie A haunted house; The review below gave the movie a 9/10

"Let's get this straight right now: I appreciate good movies. Was this a good movie? No, in fact it was pretty stupid. Will people hate it? Absolutely. Did I enjoy it? A whole lot. I don't know what it was about this movie, but at times I could not stop myself from laughing. In my head I knew it was stupid, but I just didn't care. It's immature, raunchy, and ridiculous, but I loved it. The only Marlon Wayans movies I like, along with this one, is the first Scary Movie, and White Chicks. All three of them were pretty stupid. But that's the point. Sometimes you just have to see those movies that are obviously not good, but will entertain you. I love good, well made movies as much as the next guy, but I also enjoy just kicking back and watching a movie that will keep me smiling for 90-120 minutes. So if you're bored, have $7-$10, and an hour and a half to burn, watch this movie."

As described by the audience member while the audience states " Was this a good movie? No", even calling the movie "pretty stupid", they then go on to emphasise the fact that the audience member still enjoyed the movie stating " I could not stop myself from laughing". This supports the fact that while audience members may not hold a movie in high regard in terms of quality, as long as the movie entertains the audience member; makes them laugh, then that is all that really matters to them, which can be reflected within the reviewer's score of 9. Most importantly the audience member states this directly within the review " Sometimes you just have to see those movies that are obviously not good, but will entertain you".

This is completely contrasted by the critic reviews which from our data set all 10 reviews gave the movie "A haunted house" scores of 2 while audiences scores ranged from 1-9, with 6 being the most frequent. As seen below within the graph. This is also supported within the review content in which the critic displays his emotions of displeasure which severely contrasts the review content of the audience review.

"No-holds-barred comedy is one thing, hurtful thoughtlessness is something else entirely. An ostensible comedy shouldn't have so many moments that feel so ugly."



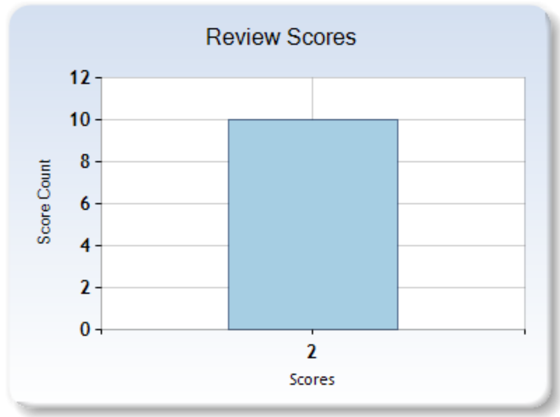


Figure iv Review Score frequency of the movie A Haunted House

It is important to note that while this snapshot is a great example of how audiences resonate more with entertainment than anything else, we cannot disregard this data as an outlier, as since we only choose 10 reviews each for both audiences and reviews, we were only able to gather a sample of their opinion. Thus it is important to note that in any case there can be specific outliers to the proposed hypothesis which must be taken into account.

Therefore it is essential to approach any hypothesis or generalisation with an understanding that there can always be exceptions or outliers that challenge the anticipated patterns. These deviations can result from various factors, including the individual nature of film perception and the diverse range of expectations and preferences within both critics and audiences. Taking into account these outliers allows for a more comprehensive and nuanced interpretation of the data, fostering a deeper understanding of the complex dynamics involved in film reception and evaluation.

4.4 Result Analysis Summary

To summarise, the differences between audiences and critics can be visualised through the analysis of 3 main concepts; average movie score, sentiment analysis (emotion and sentiment), movie theme; individual review content. For the average movie score, it can be noted that audiences had a range that centred around the middle on 1-10 point scale, whilst critics had reviews that were either very low or very high. This demonstrates the first difference between audiences and critics being that reviews of the audience are more balanced and the reviews of critics are much more polarising. In terms of sentiment analysis, it can be said that audiences express mainly emotions of joy (29.22%), disgust (15.18%) and neutral (25.18%) while for critics' reviews have the three top categories of neutral (37.78%), fear (22.72%) and disgust (15.35%). This was explained, to be due to the reason that audiences resonate with nostalgia more and thus are more likely to watch movies based around the idea of nostalgia and express said nostalgia with joy. Lastly, for movie themes, it can be seen that critics tended to rate movies such as horror and comedy lower while children's movies and real-life drama tended to get higher scores. This phenomenon was explained to be the fact that, for audiences, they prioritise entertainment above most other factors in a movie. For cheap thrills and laughs, namely comedy and horror, audiences tended to be more forgiving; even if a movie was poor. This was contrary for critics since they view the movie not as entertainment but rather something to be critiqued. Therefore, through the analysis of the differing review styles of audiences and critics, certain conclusions can be drawn regarding the respective perspectives and values. In this case, it can be seen that audiences view films for entertainment and enjoyment whilst critics will regard the film as a work of art. This simple difference in perspective creates the gap between a critic and an average viewer.

5 Verify and Validate Research Outcomes

5.1 Accuracy of the RoBERTa Model

In order to ensure that the validity of the results used for analysis is accurate, we need to make sure that the sentiment analysis model is proven to be accurate and effective. We conducted a comparison experiment between our chosen sentiment analysis model from RoBERTa from HuggingFace and the standard VADER model that came with the NLTK Python library. We would need to perform sentiment analysis using two datasets, the audience's movie reviews and the critics' movie reviews. Then taking the sentiment scores of negative, neutral and positive within the CSV files, we can plot it in a line plot graph with a regression line to see the trend as seen in figure ii, iii, iv and v. After conducting our sentiment analysis using HuggingFace with the RoBERTa model and comparing it to NLTK with the VADER model, we have arrived at some interesting findings.

```
fig, axs = plt.subplots(1, 3, figsize=(15, 3))

sns.lineplot(data=audience, x='reviewRating', y='pos_sentiment', ax=axs[0])
sns.regplot(data=audience, x='reviewRating', y='pos_sentiment', ax=axs[0], scatter=False)

sns.lineplot(data=audience, x='reviewRating', y='neu_sentiment', ax=axs[1])
sns.regplot(data=audience, x='reviewRating', y='neu_sentiment', ax=axs[1], scatter=False)

sns.lineplot(data=audience, x='reviewRating', y='neg_sentiment', ax=axs[2])
sns.regplot(data=audience, x='reviewRating', y='neg_sentiment', ax=axs[2], scatter=False)

axs[0].set_title('Positive')
axs[1].set_title('Neutral')
axs[2].set_title('Negative')

for ax in axs:
    ax.set(yticks=[0, 0.2, 0.4, 0.6, 0.8, 1])
    ax.set(xticks=range(1, 11, 1))

plt.tight_layout()
plt.show()
```

Figure i) The code for plotting the comparison graphs [6]

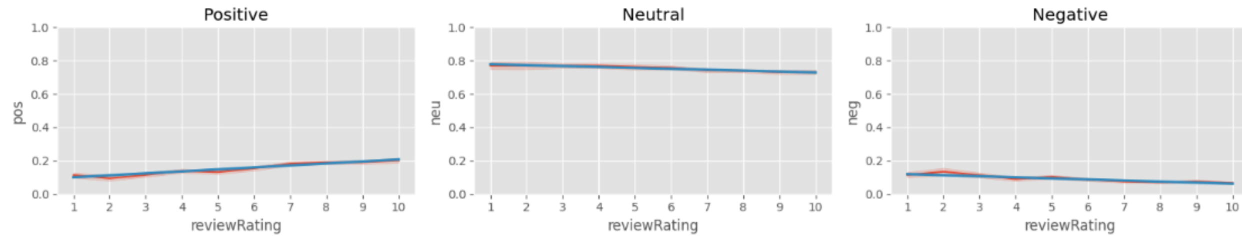


Figure ii) The accuracy of the VADER sentiment analysis model for Audience movie reviews

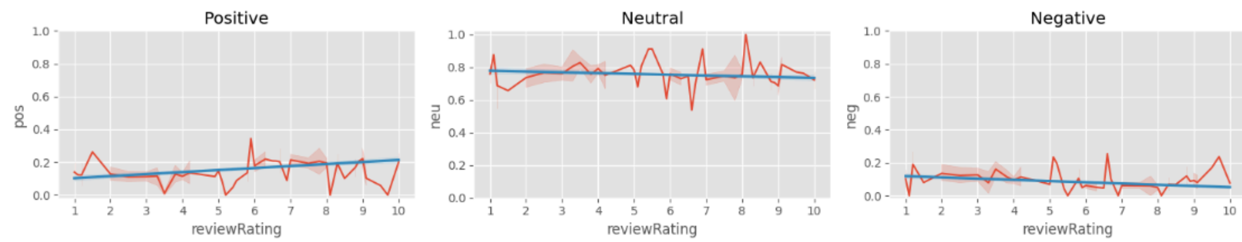


Figure iii) The accuracy of the VADER sentiment analysis model for Critic movie reviews

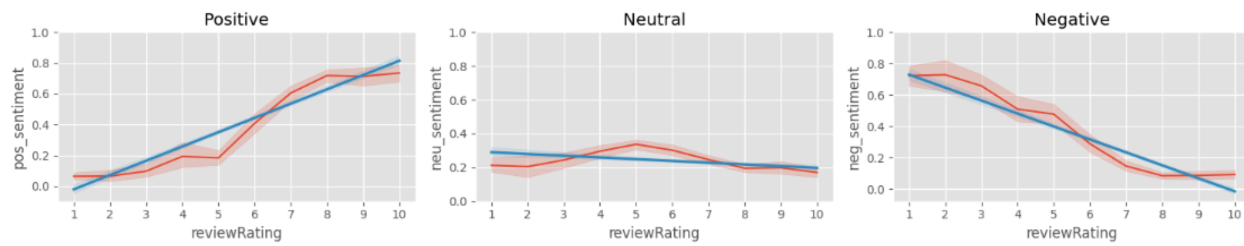


Figure iv) The accuracy of the RoBERTa sentiment analysis model for Audience movie reviews

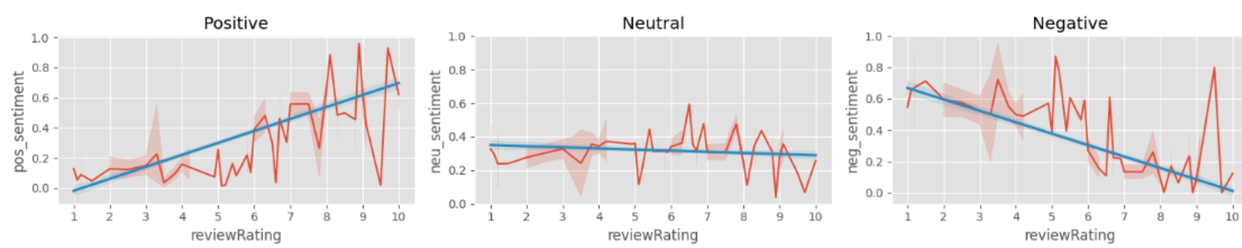


Figure v) The accuracy of the RoBERTa sentiment analysis model for Critic movie reviews

As seen in the figures ii, iii, iv and v, the use of HuggingFace with the RoBERTa model proved to be a significant improvement over NLTK with the VADER model in our sentiment analysis tasks. The sentiment scores indicate the confidence and likelihood that the movie review is either positive, neutral or negative. The aim of this comparison is to see which model accurately labels the movie reviews based on the review ratings. So the lower the review rating (close to 1), the lower the positive sentiment and the higher negative sentiment. Likewise, for the higher review ratings (close to 10), there should be a higher positive sentiment and a lower negative sentiment.

As seen in the VADER model in Figures ii and iii, although there is a slight upwards trend for the positive sentiment, the likelihood score is not confident in labelling the review as positive or negative but most labels it as neutral. Comparing this to the RoBERTa model in figure iv and v, it is evident that the sentiment scores are much higher in confidence, giving reviews with high ratings a higher positive sentiment and giving reviews with lower ratings a higher negative sentiment.

We can conclude that the RoBERTa model, being a state-of-the-art pre-trained with large datasets and fine-tuned for sentiment analysis, provided advanced features and capabilities that outperformed the VADER model. It demonstrated superior accuracy in capturing nuanced sentiment and detecting emotions within movie reviews. This validates the reliability of the results of analysis, ensuring the results are credible and accurate.

5.2 Verification of the Data Source

The data source; IMDB, which allowed us to gather our results was quite valid by which it gathered real data from the IMDB, as certain snapshots of the data were cross-checked with the official IMDB website to ensure the API was fetching data straight from the website. However, while IMDB is an extremely reliable and valid source, IMDB has no real way of verifying user reviews, this means that users could possibly create multiple accounts and influence the reviews and/or ratings of a certain movie, this means that while the external information such as movie title, movie rating, movie critic rating may be valid, the user reviews may not be.

5.3 Verification of Results Analysis

The results obtained discussed the hypothesis through the use of the aim and objectives being fluently carried out. The main segment of the research paper analyses the different methods in analysing the outcomes of the project methodology through the use of the average scores of critics and audience, the average emotion of the sentiment analysis and the film themes presented. These results presented a fluent and comprehensive analysis of the project proposal. However, it did present some forthcomings regarding the research questions. They lacked overall sophistication in areas such as ethical issues as this is hard to quantify through the use of graphs and results. This, however, was included in our research aims and therefore could have been discussed at greater length throughout the overall project.

5.4 Discussion on Validation of Research Outcomes

The RoBERTa model used for sentiment analysis was thoroughly validated to ensure the accuracy of the results. A comparison experiment was conducted between RoBERTa and the VADER model from the NLTK Python library. Sentiment analysis was performed on the audience and critics' movie reviews, and the sentiment scores were plotted in line graphs with regression lines. The findings showed that the RoBERTa model outperformed the VADER model, demonstrating higher confidence in accurately labelling movie reviews based on ratings.

The use of RoBERTa provided advanced features and capabilities, capturing nuanced sentiment and detecting emotions more effectively. The data source, IMDb, was verified by crosschecking snapshots with the official IMDb website, although it should be noted that user reviews may be influenced by potential biases. The results analysis effectively addressed the research questions, aims, and objectives, providing a comprehensive analysis of average scores, sentiment analysis, and film themes. However, there were limitations in addressing ethical issues, which were acknowledged in the research aims and could have been discussed in greater detail. Overall, the verification and validation processes ensured the reliability and credibility of the research outcomes.

6 Discussion and Conclusions

6.1 Research Study Discussion

The results discussion has been made in the Results Analysis, now we will discuss other findings we encountered during this research.

6.1.1 The Model Used

One of the key strengths of HuggingFace is its extensive library of pre-trained models, including RoBERTa, which allowed us to leverage the expertise of the NLP community. These models are trained on vast amounts of data and have learned intricate patterns, making them highly effective in analysing natural language text. This wide range of pre-trained models gave us the flexibility to experiment and select the most suitable one for our research.

The RoBERTa model's advanced architecture and fine-tuning capabilities played a crucial role in its superior performance. It exhibited a deeper understanding of language semantics and contextual information, enabling more accurate sentiment analysis. The fine-tuning process allowed us to adapt the model to our specific domain, resulting in better alignment with movie reviews and improved accuracy.

Furthermore, HuggingFace provided comprehensive documentation and a supportive community, making it easier for us to integrate and implement the RoBERTA model in our research. The availability of code examples, tutorials, and forums enhanced our understanding and facilitated a smoother workflow. We found HuggingFace to be user-friendly and highly accessible, even for researchers with limited prior experience in NLP.

6.1.2 Limitations

During our research project, we encountered several limitations that impacted our analysis and findings. Firstly, we faced a constraint on the length of reviews, as those exceeding 512 tokens were either ignored by the model or caused runtime errors. This restriction limited our ability to analyse longer and more detailed reviews, potentially missing out on valuable insights contained within them. Another significant limitation was the varying technical skills within our group. Some members struggled with coding and had to find alternative methods for generating graphs and data visualisations, bypassing the use of programming languages like Python or R. This skill disparity hindered our efficiency and may have impacted the quality and accuracy of our visual representations.

Time constraints also posed a challenge, as we were unable to thoroughly analyse all aspects of the sentiment analysis data. This resulted in some unused data, potentially missing out on valuable trends or patterns. Additionally, given that we were dealing with the latest technologies and models, finding comprehensive guides or resources proved challenging, and we heavily relied on reading documentation. This reliance on documentation may have introduced a higher risk of misinterpretation or incomplete understanding of the methodologies and techniques used.Furthermore, due to the nature of human language, our sentiment analysis model may not accurately capture elements such as sarcasm or irony. These nuances can significantly impact the interpretation of reviews and introduce potential biases or inaccuracies in our analysis.

In addition to the aforementioned limitations, we also encountered constraints related to the number of reviews we could web scrape and the limitations imposed by API usage. Due to the IP limits and restrictions imposed by websites, our ability to collect a large volume of reviews was restricted. This limited sample size may have impacted the representativeness and generalisability of our findings. Furthermore, API limits the diversity and breadth of our dataset. These limitations underscore the importance of considering data availability and accessibility when conducting research, as they can influence the comprehensiveness and robustness of our analyses.

Lastly, as we analysed online movie rating reviews, we encountered the issue of unreliable data. Some individuals may provide random or inconsistent ratings that do not align with their actual review. This discrepancy can distort our findings and introduce noise into our analysis, affecting the overall reliability and validity of our results. Despite these limitations, we made conscientious efforts to address and mitigate their impact throughout our research project. Awareness of these constraints allows for a more nuanced interpretation of our findings and highlights areas for improvement in future studies.

6.2 Future Directions

Based on the study we have just conducted, it can be observed that there are a wide range of potential applications and future directions that sentiment analysis could take, namely:

1) Incorporating user-specific content and recommendations: One potential direction for sentiment analysis is to develop new models that take into account the differing tastes, preferences, opinions and biases. By integrating these models, sentiment analysis can provide a more unique and personalised experience to any integrated system that can provide more accurate predictions than the current recommendation models. If we use the context of movies in this instance, we can see that sentiment analysis could analyse our opinions and provide more personalised and accurate recommendations for the type of shows and movies we watch.
2) Dynamic Sentiment Analysis: Currently sentiment is constantly evolving especially due to evolving opinions or trends, one noticeable example is the environmental movement, in which opinions are ever-growing and changing, thus dynamic sentiment analysis could be a potential system to capture and track the trends of sentiment within a given field, providing real-time data if integrated into a form of media such as social media. This would potentially be valuable in tracking public opinion or just regular social media trends.
3) Multimodal sentiment analysis, currently all forms of sentiment analysis that were studied, only work through text input, meaning that sentiment analysis cannot be applied to other forms of modality such as images (facial expressions), audio (vocal tone) and video. While text can provide many forms of sentiment, most forms of sentiment are often expressed with facial expression and vocal tone, thus by integrating multiple forms of modality, sentiment analysis could become an even more powerful tool that could analyse the opinions in many different fields.
4) It is evident that sentiment analysis currently struggles with classifying sarcasm and irony, as this may lead to potential misinterpretation and could create a possible skew in the results. Future directions could involve developing models that have the capabilities to handle sarcasm and irony to better analyse and capture the sentiment.
5) We could expand our analysis beyond movie reviews and apply the HuggingFace and RoBERTa combination to other domains, such as product reviews, social media sentiment analysis, or customer feedback. This would allow us to evaluate its effectiveness and generalisability across different contexts.
6) Additionally, we can delve deeper into the analysis of specific aspects of sentiment, such as identifying sarcasm, and irony, or detecting sentiment shifts over time. This would require fine-tuning the RoBERTa model on annotated datasets specific to these aspects, enabling more nuanced sentiment analysis.
7) Moreover, incorporating additional features such as contextual information, user demographics, or review metadata could enhance the accuracy and depth of our sentiment analysis. By considering these factors, we could gain a better understanding of the factors influencing the disparity between audience and critic movie reviews.

6.3 Conclusion

In conclusion, our analysis of critic and audience reviews has unveiled intriguing insights into their contrasting perspectives. Audiences primarily rate movies based on enjoyment, leading to a narrower range of scores, while critics, with their analytical approach, assign a wider range of scores, focusing on cinematic elements. Additionally, the prevalence of joy as the main emotion in audience reviews suggests a strong connection to nostalgia, which may contribute to the differing evaluations between critics and audiences. Our examination of specific genres highlights that real-life drama and children's movies tend to receive higher scores from critics, emphasising their appreciation for storytelling and impactful narratives. Conversely, comedy and horror movies, perceived as formulaic, receive lower scores. These disparities stem from the distinct goals of critics and audiences, with critics prioritising artistic merit and audiences seeking entertainment. Overall, our findings underscore the inherent disparity between critic and audience perspectives. Understanding these differences provides valuable insights for filmmakers, critics, and audiences alike, enhancing our understanding of the complexities surrounding movie evaluations.

7 Related Work

"Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open Source Technologies" (B. Narendra, 2016) [7] investigates the suitability and performance of machine learning algorithms versus open source technologies for conducting sentiment analysis on movie reviews. The study compares the Naïve Bayes Classifier machine learning algorithm with the MapReduce framework in Apache Hadoop. We did a similar comparison with the VADER model and the RoBERTa model for sentiment analysis of movie reviews.

Swathi et. al. [8] conducted a comparative sentiment analysis between movie reviews that were posted to The New York Times and those posted on Twitter. The reason for this analysis was to “gain insight on research analysis on social media against traditional media sources”. They wanted their results to improve the understanding of future perceptions of reviews. Our research also gained insight as we wanted to find the discrepancy between the reviews set by the audience against those of critics.

"Sentiment Analysis in Social Media and Its Application," by Zulfadzli & Khalid, Haliyana. (2019) [9] explores the overall topic of sentiment analysis, how it is performed and its application. The paper introduces sentiment analysis by highlighting the growing importance of sentiment analysis in social media and then proceeds to outline its objectives which are to identify the types of social media platforms used, the languages analysed and the applications of sentiment analysis. Our research also aims to explore the topics of sentiment analysis within the social paradigm and to draw conclusions on the sentiment of two different groups.

8 Appendix

<table>
<tr><th>Group Member</th><th>Contribution</th></tr>
<tr><td>Mark Pacetta</td><td>4. Results- Creation of Fig i, ii, iii and analysis
Verification of results analysis - 5.3.

Validation:
The research outcomes obtained were a reflection of the research questions which we presented within our overall project. The main outcomes which we discussed were the impact of money on critic reviews, the themes of movies affecting critic reviews, the sentiment value comparison between critics and audience and the overall rating comparison between critics and audiences.

In order to validate the data which we have received and to make sure the sentiment analysis was accurate we used figure iv) which presents the correlation between sentiment and review rating score. This shows that although the sentiment ratings are not perfectly accurate they definitely follow the trend of the movie ratings and have statistical correlation. It is acceptable as in many cases the sentiment analysis fails to note aspects of irony and sarcasm.

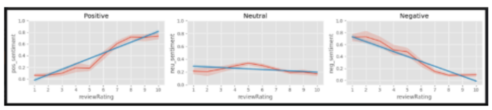


Fig v) The testing of the expected sentiment against the scores

Verification:

The results obtained discussed the hypothesis through the use of the aim and objectives being fluently carried out. The main segment of the research paper analyses the different methods in analysing the outcomes of the project methodology through the use of the average scores of critics and audience, the average emotion of the sentiment analysis and the film themes presented. These results presented a fluent and comprehensive analysis of the project proposal however did present some forthcomings surrounding the research questions. They lacked overall sophistication in areas such as ethical issues as this is hard to quantify through the use of graphs and results. This however was included in our research aims and therefore could have been discussed in greater length throughout the overall project.

Discussion:
Main Points:
- Was the model used the most effective
The RoBERTA model which we used was the most in-depth and analytical tool and was crucial in the verification and validation of the results as it provided accurate results.
- Future directions of the project

Conclusion:

In response to our hypothesis and project aims we have analysed and proven the existing links between critic and audience reviews through the use of sentiment analysis. The pinnacle points for discussion surround themes, sentiment and specific emotions allowed for in depth analysis. The report has concluded that there is an overall link between critic and audience reviews with critics basing their reviews off cinematography and acting whereas the audience will base their reviews on whether they enjoyed the film. On the critic perspective for film reviews the genre of films also played a large role in scores given by the critics.

Notes brought to discussion for validation, verification, discussion and conclusion</td></tr>
<tr><td>Batool Alabduljabbar</td><td>Review Project Proposal
Contributed to the analysis 4.3.1 - 5.4</td></tr>
</table>

<table>
<tr>
<td></td>
<td>
Individual work Verify and Validate Research Outcomes:

5.4 Discussion on Validation of Research Outcomes

The RoBERTa model used for sentiment analysis was thoroughly validated to ensure the accuracy of the results. A comparison experiment was conducted between RoBERTa and the VADER model from the NLTK Python library. Sentiment analysis was performed on audience and critics movie reviews, and the sentiment scores were plotted in line graphs with regression lines. The findings showed that the RoBERTa model outperformed the VADER model, demonstrating higher confidence in accurately labelling movie reviews based on ratings. The use of RoBERTa provided advanced features and capabilities, capturing nuanced sentiment and detecting emotions more effectively. The data source, IMDb, was verified by crosschecking snapshots with the official IMDb website, although it should be noted that user reviews may be influenced by potential biases. The results analysis effectively addressed the research questions, aims, and objectives, providing a comprehensive analysis of average scores, sentiment analysis, and film themes. However, there were limitations in addressing ethical issues, which were acknowledged in the research aims and could have been discussed in greater detail. Overall, the verification and validation processes ensured the reliability and credibility of the research outcomes.
</td>
</tr>
<tr>
<td>Jason Vu</td>
<td>
Performed sentiment analysis on the movie reviews CSV files that had been web scraped using a fine-tuned RoBERTa model from HuggingFace via Kaggle. We also used NLTK VADER's sentiment analysis model but did a comparison of accuracy with the RoBERTa model and concluded that the RoBERTa model had higher confidence in the sentiment labelling

Contributed to
2 Review of Feedback
3.2 Sentiment Analysis
5.1 Accuracy of the RoBERTa Model
6.1.2 Limitations
6.2 Future Directions
7 Related Work

Individual work Verify and Validate Research Outcomes:

5.1 Accuracy of the RoBERTa Model

In order to ensure that the validity of the results used for analysis is accurate, we need to make sure that the sentiment analysis model is proven to be accurate and effective. We conducted a comparison experiment between our chosen sentiment analysis model from RoBERTa from HuggingFace and the standard VADER model that came with the NLTK Python library. We would need to perform sentiment analysis using two dataset, the audience movie reviews and the critics movie review. Then taking the sentiment scores of negative, neutral and positive within the CSV files, we can plot it in a line plot graph with a regression line to see the trend as seen in figure ii, iii, iv and v. After conducting our sentiment analysis using Huggingface with the RoBERTa model and comparing it to NLTK with the VADER model, we have arrived at some interesting findings. In this section, we will discuss and summarise our research study, highlighting the strengths of HuggingFace and RoBERTA while identifying potential future directions for our study.

The code used for validating and comparing the RoBERTa model:
</td>
</tr>
</table>

	https://www.kaggle.com/code/hickman2049/accuracy-of-roberta-vs-vader-movie-reviews
Joshua Nguyen	Assisted Jason with performing sentiment analysis using the RoBerta Model through Huggingface from Kaggle's resources. Mainly dealt with critic reviews. Additionally used NLTK VADER's sentiment analysis model, and concluded that the RoBERTA model was more effective in the sentiment labelling. Contributed to 3.2 Sentiment Analysis 3.3 Data Visualisation 3.4 Analysis of Results 5.3 Verification of Results Analysis 6.1.1 The Model Used 6.2 Future Directions 6.3 Conclusion Individual work on the 'Verify and Validate Research Outcomes', specifically 5.3 **5.3 Verification of Results Analysis** The results obtained discussed the hypothesis through the use of the aim and objectives being fluently carried out. The main segment of the research paper analyses the different methods in analysing the outcomes of the project methodology through the use of the average scores of critics and audience, the average emotion of the sentiment analysis and the film themes presented. These results presented a fluent and comprehensive analysis of the project proposal however did present some forthcomings surrounding the research questions. They lacked overall sophistication in areas such as ethical issues as this is hard to quantify through the use of graphs and results. This however was included in our research aims and therefore could have been discussed in greater length throughout the overall project. **6.3 Conclusion** In conclusion, our analysis of critic and audience reviews has unveiled intriguing insights into their contrasting perspectives. Audiences primarily rate movies based on enjoyment, leading to a narrower range of scores, while critics, with their analytical approach, assign a wider range of scores, focusing on cinematic elements. Additionally, the prevalence of joy as the main emotion in audience reviews suggests a strong connection to nostalgia, which may contribute to the differing evaluations between critics and audiences. Our examination of specific genres highlights that real-life drama and children's movies tend to receive higher scores from critics, emphasising their appreciation for storytelling and impactful narratives. Conversely, comedy and horror movies, perceived as formulaic, receive lower scores. These disparities stem from the distinct goals of critics and audiences, with critics prioritising artistic merit and audiences seeking entertainment. Overall, our findings underscore the inherent disparity between critic and audience perspectives. Understanding these differences provides valuable insights for filmmakers, critics, and audiences alike, enhancing our understanding of the complexities surrounding movie evaluations.
Alex Tran	Performed web scraping by running js scripts within the browser to access critical movie & movie review data through a public API and filtered said data into JSON and CSV format.

	Contributed to 3.1 Web Scraping 4.2 Sentiment analysis 4.3 Film theme 4.4 Result analysis conclusion 6.2 Future Directions Analysed reviews in depth to back up points Individual work Verify and Validate Research Outcomes **5.2 Verification of the Data Source** The data source; IMDB, which allowed us to gather our results was quite valid by which it gathered real data from the IMDB, as certain snapshots of the data were crosschecked with the official IMDB website to ensure the api was fetching data straight from the website. However while IMDB is a extremely reliable and valid source, IMDB has no real way of verifying user reviews, this means that user could possibly create multiple accounts and influence the reviews and/or ratings of a certain movie, this means that while the external information such as, movie title, movie rating, movie critic rating may be valid, the user reviews may not be.
Justin Tran	Performed general document maintenance Contributed to 2 Review of Feedback Individual work: Verify and Validate Research Outcomes: The data generated from the web scraping was stored into a CSV file and then manually cross-referenced with the reviews on the respective website that they were taken from. This ensures that the API utilised was returning the correct and expected output.

References

[1] IMDb-API. (n.d.). *API for IMDb, Wikipedia and more - IMDb-API.* IMDb API. https://imdb-api.com/

[2] Vu, J., & Nguyen, J. (2023). *Movie Reviews Sentiment Analysis*. Kaggle. https://www.kaggle.com/code/hickman2049/movie-reviews-sentiment-analysis

[3] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, July 26). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv.org. https://arxiv.org/abs/1907.11692

[4] cardiffnlp. (2023). *Twitter-roBERTa-base for Sentiment Analysis*. HuggingFace. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment

[5] j-hartmann. (2023). *Emotion English DistilRoBERTa-base*. HuggingFace. https://huggingface.co/j-hartmann/emotion-english-distilroberta-base

[6] Vu, J. (2023). *Accuracy of RoBERTa vs VADER Movie Reviews*. Kaggle. https://www.kaggle.com/code/hickman2049/accuracy-of-roberta-vs-vader-movie-reviews

[7] B. Narendra, K. Uday Sai, G. Rajesh, K. Hemanth, M. V. Chaitanya Teja, K. Deva Kumar. (2016). *Sentiment Analysis on Movie Reviews: A Comparative Study of Machine Learning Algorithms and Open*. http://doi.org/10.5815/ijisa.2016.08.08

[8] Swathi, H., Aravinth, S., Nivethitha, V., Saranya, T., Nivethanandhini, R. (2019). *Sentiment Analysis of Movie Review using Data Analytics Techniques.* https://www.irejournals.com/formatedpaper/1701029.pdf

[9] Drus, Zulfadzli & Khalid, Haliyana. (2019). *Sentiment Analysis in Social Media and Its Application: Systematic Literature Review*. Procedia Computer Science. 161. 707-714. 10.1016/j.procs.2019.11.174. https://www.sciencedirect.com/science/article/pii/S187705091931885X