INSTITUTO
UNIVERSITÁRIO
DE LISBOA

Impact of Cinematic Aspects on Ratings: A Sentiment Analysis of Movie Reviews

Catarina Filipa Lourenço de Sousa

Master in Information Systems Management

Supervisor:
PhD Eugénio Alves Ribeiro, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

September, 2025

TECNOLOGIAS
E ARQUITETURA

Department of Information Science and Technology

Impact of Cinematic Aspects on Ratings: A Sentiment Analysis of Movie Reviews

Catarina Filipa Lourenço de Sousa

Master in Information Systems Management

Supervisor:
PhD Eugénio Alves Ribeiro, Assistant Professor,
Iscte – Instituto Universitário de Lisboa

September, 2025

Acknowledgements

I would like to express my gratitude to my supervisor, Professor Eugénio Ribeiro, for his availability and support throughout the development of this dissertation. His expertise and constructive feedback were fundamental in shaping this work.

Special thanks go to my family and friends, whose patience and encouragement gave me the resilience to face this demanding process. Their trust in my abilities and their constant support, both in moments of doubt and of progress, were vital in allowing me to complete this dissertation. I could not have reached this stage without their presence and motivation throughout the journey.

Resumo

As críticas de filmes online constituem uma fonte valiosa de feedback do público, mas a sua riqueza avaliativa ultrapassa frequentemente a análise de sentimentos global. A Análise de Sentimentos Baseada em Aspetos (ABSA) permite associar sentimentos a dimensões cinematográficas específicas. Esta dissertação investiga quais os aspetos que mais influenciam as classificações dos espetadores, através de uma metodologia sistemática e multifásica. Foi utilizado um *dataset* de 1.000 críticas do IMDb, segmentadas e pré-processadas antes de serem analisadas por três abordagens ABSA. Duas abordagens seguiram uma estrutura em *pipeline*, combinando a deteção de aspetos por palavras-chave com a classificação de sentimentos pelo modelo léxico VADER ou pelo transformador DistilBERT, enquanto a terceira adotou um desenho *end-to-end*, recorrendo a modelos de linguagem de grande escala (GPT-4o mini e GPT-4.1-mini) para extrair diretamente aspetos e sentimentos. O desempenho foi avaliado com base num *gold standard* manual de aspetos e nas etiquetas de polaridade do *dataset*, utilizando métricas de *precision*, *recall, F1 score,* e *accuracy*, cujos resultados comparativos orientaram a seleção do método mais fiável para as análises subsequentes. Com este método, realizaram-se análises ao nível da crítica, de aspetos dominantes e de combinações de aspetos. Constatou-se que Enredo e Elenco são os principais determinantes das classificações, enquanto Realização e Ambiente Audiovisual têm contributos secundários, mas relevantes. Revelou-se ainda que as classificações emergem da interação entre múltiplos aspetos, sobretudo quando Enredo e Elenco se alinham. Assim, o estudo oferece contributos académicos para a ABSA e implicações práticas para a indústria cinematográfica e plataformas de *streaming*.

Palavras-Chave: Análise de Sentimentos; Análise de Sentimentos Baseada em Aspetos; Aspetos Cinematográficos; Comentários Online; Filmes.

Abstract

Online movie reviews constitute a valuable source of audience feedback, yet their evaluative richness often exceeds the scope of overall sentiment analysis. Aspect-Based Sentiment Analysis (ABSA) provides a finer-grained approach by linking sentiments to specific cinematic dimensions. This dissertation investigates which aspects of films most strongly influence audience ratings, addressing the research question through a systematic, multi-phase methodology. A dataset of 1,000 IMDb reviews was used, with texts segmented and preprocessed before being analysed through three ABSA approaches. Two followed a pipeline structure, combining keyword-based aspect detection with sentiment classification via either the lexicon-based VADER model or the transformer-based DistilBERT, while the third adopted an end-to-end design using large language models (GPT-4o mini and GPT-4.1 mini) to extract aspects and sentiments directly. Performance was evaluated against a manually annotated gold standard for aspects and the dataset's polarity labels for sentiment, using precision, recall, F1 score, and accuracy, with the comparative results guiding the selection of the most reliable method for subsequent analyses. With the selected method, review-level, dominant-aspect, and combination-aspect analyses were conducted. The findings demonstrate that Plot and Cast are most influential aspects in shaping ratings, while Directing and Ambience act as secondary but meaningful contributors. Moreover, results reveal that ratings emerge not from isolated dimensions but from the interaction of multiple aspects, particularly when Plot and Cast align. The study combines methodological rigor with applied analysis, providing academic insights into ABSA and offering practical implications for the film industry and streaming platforms.

Keywords: Sentiment Analysis; Aspect-Based Sentiment Analysis (ABSA); Cinematic Aspects; Online Reviews; Movies.

Index

Acknowledgements ..... **i**
Resumo ..... **ii**
Abstract ..... **iii**
Index ..... **iv**
Tables index ..... **vi**
Figures index ..... **vii**
List of abbreviations ..... **viii**
Chapter 1 – Introduction **1**
1.1. Motivation and topic relevance 2
1.2. Questions and research goals 3
1.3. Methodologic approach 3
1.4. Structure and organisation of dissertation 4
Chapter 2 – Literature review **5**
2.1. An overview of cinema 5
2.2. Online reviews 6
2.2.1. Importance of online reviews 6
2.2.2. Online movie reviews 7
2.3. Sentiment analysis 8
2.3.1. Definition and importance of sentiment analysis 8
2.3.2. Application of sentiment analysis in movie reviews 8
2.4. Aspect-based sentiment analysis 10
2.4.1. Definition of aspect-based sentiment analysis 10
2.4.2. Aspect extraction (AE) 11
2.4.3. Opinion extraction (OE) 11
2.4.4. Aspect sentiment classification (ASC) 12
2.5. Aspect-based sentiment analysis in movie reviews 12
Chapter 3 – Research methodology **15**
3.1. Research design 15
3.2. Data collection and preparation 15
3.2.1. Dataset selection 15
3.2.2. Gold standard construction 16
3.2.3. Text preprocessing 18
3.3. ABSA methods 19

3.3.1. Pipeline-based approach 19
3.3.2. End-to-end approach 22
3.4. Method evaluation and selection 23
Chapter 4 – Results presentation and analysis 25
4.1. Evaluation of ABSA methods 25
4.2. Review-level analyses 29
4.2.1. Aspect prevalence, sentiment, and co-occurrence 29
4.2.2. Aspect–rating relationships 32
4.3. Dominant aspect analyses 35
4.3.1. Prevalence and rating distribution of dominant aspects 36
4.3.2. Agreement between the dominant aspect approaches 37
4.3.3. Explanatory power of dominant aspects 39
4.4. Combination analyses 41
Chapter 5 – Conclusions and future research 43
5.1. Main conclusions 43
5.2. Contributions to the scientific and business community 44
5.3. Research limitations 44
5.4. Future research proposals 45
References 46
Appendices 51
Appendix A 51
Appendix B 52

Tables index

Table 1 - Example of the dataset structure with one sample review 16

Table 2 - Example of manual aspect annotation for a single review 17

Table 3 - Cohen's Kappa values for inter-annotator agreement by cinematic aspect 17

Table 4 - Interpretation of Cohen's Kappa values according to Landis and Koch (1977) 18

Table 5 - Example of the dataset after sentence segmentation and text preprocessing 19

Table 6 - Aspects, definitions, and keywords (adapted from Kit and Joseph, 2023) 20

Table 7 - Evaluation of aspect extraction by cinematic aspect 25

Table 8 - Overall evaluation of aspect extraction (macro and micro averaging) 26

Table 9 - Performance of sentiment classification (Neutral mapped to Negative) 27

Table 10 - Performance of sentiment classification (Neutral mapped to Positive) 27

Table 11 - Correlations between aspect sentiments and review ratings (Pearson and Spearman) 33

Table 12 - Impact of aspect sentiments on ratings (OLS regression results) 34

Table 13 - Concordance between frequency- and score-based dominant aspect identification (per review) 38

Table 14 - Correlations between dominant-aspect sentiment and review ratings (Pearson and Spearman) 39

Table 15 - Impact of dominant-aspect sentiment on review ratings (OLS regression results) 39

Figures index

Figure 1 - Example of a rating on IMDb 2

Figure 2 - Example of the four key sentiment elements in ABSA 11

Figure 3 - Research design of the study 15

Figure 4 - Review-level analyses: reviews mentioning each aspect (left) and total aspect mentions (right). 30

Figure 5 - Review-level analyses: average sentiment by aspect 30

Figure 6 - Review-level analyses: sentiment distribution by aspect (100%) 31

Figure 7 - Review-level analyses: aspect co-occurrence by sentiment 31

Figure 8 - Review-level analyses: aspect co-occurrence with positive–negative contradictions 32

Figure 9 - Review-level analyses: average sentiment by rating and aspect 33

Figure 10 - Dominant aspect analyses: prevalence of dominant aspects 36

Figure 11 - Dominant aspect analyses: distribution of ratings by dominant aspect 37

Figure 12 - Dominant aspect analyses: contingency table, frequency vs. score 38

Figure 13 - Aspect combinations across rating intervals (counts and row-normalised proportions; shading reflects proportions) 42

List of abbreviations

ABSA - Aspect-Based Sentiment Analysis

API - Application Programming Interface

BERT - Bidirectional Encoder Representations from Transformers

BiLSTM - Bi-directional Long Short-Term Memory

CGI - Computer-Generated Imagery

CNN - Convolutional Neural Networks

CRF - Conditional Random Fields

FN - False Negatives

FP - False Positives

GCN - Graph Convolutional Networks

LDA - Latent Dirichlet Allocation

LLM – Large Language Model

LSTM - Long Short-Term Memory

MLP - Multilayer Perceptron

NLP - Natural Language Processing

NLTK - Natural Language Toolkit

OLS - Ordinary Least Squares

RNN - Recurrent Neural Networks

RNTN - Recursive Neural Tensor Networks

SVM - Support Vector Machines

TF-IDF - Term Frequency-Inverse Document Frequency

TN - True Negatives

TP - True Positives

VADER - Valence Aware Dictionary and sEntiment Reasoner

Chapter 1 – Introduction

Movies have been a fundamental form of art and entertainment for several decades, playing an important role in shaping society's culture and leisure (Curran & Hesmondhalgh, 2019). From the emergence of traditional cinemas to the digital era, the way in which audiences engage with cinematic content has evolved significantly. Within this transformation, online movie reviews have become increasingly prominent, mediating the relationship between audiences and films and shaping collective perceptions of cinematic works.

In the present day, viewers have easier access to movie reviews on digital platforms, enabling them to share and consult public opinions conveniently and immediately. This accessibility is particularly important because movies are considered *experience goods*, meaning their quality can only be evaluated after consumption. As a result, many viewers rely on third-party opinions before deciding whether to watch a particular film (Kim et al., 2013). Beyond influencing individual decision-making, online reviews also play a crucial role in the popularity and success of films. On the one hand, the volume of reviews, whether positive or negative, significantly impacts box office revenues, particularly during the initial weeks of release (Liu, 2006). On the other hand, the sentiments expressed in these reviews are a key factor in predicting a movie's success, influencing both audience opinions and external publicity (Dellarocas et al., 2007).

Given the relevance and influence of online movie reviews in the film industry, understanding which specific aspects of films contribute to their overall rating is a crucial task. Elements such as the cast, plot, direction, and soundtrack can play a significant role in influencing viewers' perception of a movie, directly affecting their final rating.

While traditional sentiment analysis typically focuses on the general sentiments expressed — positive or negative — towards an entire movie (Devi et al., 2020), aspect-based sentiment analysis (ABSA) adopts a more targeted approach by isolating sentiments associated with specific components, such as acting or plot. This method provides deeper insights into which elements resonate most with audiences, offering a more nuanced understanding of viewer preferences (Onalaja et al., 2021).

In this context, the present study aims to apply ABSA to uncover patterns in audience feedback, providing actionable insights that can guide decision-making processes within the film industry.

1.1. Motivation and topic relevance

With the growing influence of the internet and social media, online reviews have become an essential tool in consumer decision-making, and movies are no exception (Tsao, 2014). Increasingly, individuals turn to platforms like IMDb[1] to evaluate the opinions and ratings of other viewers before deciding whether to watch a particular film. However, the overall movie rating, typically represented by a score or classification (Figure 1), often appears vague and subjective, as it does not explicitly indicate which cinematic aspects — such as cast, plot, direction, and soundtrack — contribute most to the rating.

Figure 1 - Example of a rating on IMDb

In light of this, this research aims to address the limited understanding of the cinematic components that viewers prioritize when evaluating a film. Although numerous studies have focused on sentiment analysis in online movie reviews (Turney, 2002; Pang et al., 2002; Maas et al., 2011; Ali et al., 2019; Devi et al., 2020), few explicitly explore the relationship between cinematic aspects and overall ratings, leaving a gap that this study seeks to fill.

By addressing this issue, the research not only aims to identify the elements that carry the most weight in ratings but also seeks to provide valuable insights for filmmakers and streaming platforms. Understanding audience preferences enables the creation of content that aligns more closely with their expectations, ultimately enhancing audience satisfaction and retention.

The relevance of this topic becomes even more evident as technological advancements open new possibilities for the film industry, driving profound changes in movie production, distribution, and exhibition (Chen, 2023). This evolving scenario highlights the need for a critical analysis of audience preferences and expectations.

[1] IMDb is the world's most popular and reliable source for movie, TV, and celebrity content, designed to help fans explore the cinematic universe and decide what to watch (https://www.imdb.com).

1.2. Questions and research goals

Given the identified problem, the following research question is proposed: *What cinematic aspects identified in movie reviews have the greatest impact on their respective ratings?*

To address this question, the research aims to understand how different aspects of a film influence viewer ratings by employing ABSA techniques.

Therefore, the objectives of this study are as follows:

- Identify the main cinematic aspects mentioned in movie reviews;
- Analyse the sentiments expressed regarding the identified aspects;
- Determine the impact of each aspect on the overall movie rating;
- Provide valuable insights for the film industry.

1.3. Methodologic approach

To understand which cinematic components have the greatest impact on movie ratings, a quantitative research design was applied to online movie reviews. This process was divided into four main phases.

The first phase involved data collection and preparation, including dataset selection and text preprocessing. The second phase applied three different approaches to extract aspects and classify sentiments: two pipeline structures, combining keyword-based aspect detection with either the lexicon-based VADER model or the transformer-based DistilBERT, and one end-to-end design, where large language models (GPT-4o mini and GPT-4.1-mini) were prompted to extract aspects and their associated sentiments directly from the reviews. The third phase evaluated the performance of these approaches using precision, recall, F1 score, and accuracy, with the comparative results guiding the choice of the most reliable method for subsequent analysis. Finally, the fourth phase comprised descriptive, explanatory, and inferential analyses of the results, examining the dataset with the selected method to identify the aspects that most strongly influenced overall rating.

This structured design ensured that the study advanced from raw data preparation to systematic method evaluation and, ultimately, to analytical findings that directly addressed the research question.

1.4. Structure and organisation of dissertation

The remainder of this document is organised in four chapters that intend to reflect the different phases until its conclusion. Chapter 2 is dedicated to the literature review, where the concepts, theories, and relevant studies on the topic are discussed. Chapter 3 describes the adopted methodology, detailing the processes of data collection and processing, as well as the analysis methods used. Chapter 4 presents the analysis of the obtained results, following the methodology deemed appropriate. Finally, Chapter 5 outlines the study's conclusions, including recommendations, identified limitations, and suggestions for future work.

Chapter 2 – Literature review

This chapter reviews the literature relevant to this study, starting with an overview of cinema and film criticism. It then examines the importance of online reviews as a form of user-generated content, focusing specifically on movie reviews. The discussion progresses to sentiment analysis as a method for interpreting opinions in text, culminating in a detailed exploration of Aspect-Based Sentiment Analysis (ABSA), the central focus of this dissertation, and its crucial application in analysing online movie reviews.

2.1. An overview of cinema

Cinema emerged in the late 19th century as a groundbreaking form of entertainment. Early inventors, such as Thomas Edison and the Lumière brothers, developed key devices like the Kinetoscope and the Cinématographe, which allowed moving images to be projected for audiences. These innovations played a crucial role in transforming moving pictures into a widely recognized medium for storytelling and entertainment (Vaniuha et al., 2024).

Initially, films were slow to gain serious critical attention. Although art criticism has long existed alongside visual art and literature, film was often dismissed as a fleeting novelty. By the time the medium was acknowledged for its artistic value, the public had already embraced it, which complicated the process of critical evaluation (Battaglia, 2010).

The late 20th and early 21st centuries witnessed a digital revolution that reshaped the filmmaking landscape. Technological advancements, such as digital cameras, computer-generated imagery (CGI), and online distribution platforms, made filmmaking more accessible and affordable. These innovations facilitated the rise of new tools, like motion capture, which continue to push the boundaries of visual storytelling. Consequently, the production of films has evolved, and so too has the way audiences engage with and experience cinema (Babbar, 2024).

In recent years, the decline of print journalism and the rise of social media have significantly transformed film criticism. Traditional reviews are increasingly overshadowed by aggregator websites that consolidate professional opinions into a single numerical score. Additionally, the rise of amateur bloggers has led to a saturation of opinions, diminishing the influence of established critics. Hollywood, recognizing the value of bloggers and influencers, has increasingly shifted its focus to these groups, finding it easier to engage them than traditional critics. As a result, professional critics are often pressured to align their views with either studio

preferences or the prevailing online sentiment, complicating the future of film criticism (Battaglia, 2010).

2.2. Online reviews

2.2.1. Importance of online reviews

Online reviews have become a crucial component of e-commerce (Wu et al., 2020). With advancements in the internet and information technology consumers are now empowered to share their product evaluations online, making these reviews significantly influence both consumer decisions and business performance in online marketplaces (Thakur, 2018).

From the perspective of the consumer, online reviews serve as a crucial source of information and trust. Positive reviews evoke emotional trust and increase confidence, reassuring potential buyers about their choices, while negative reviews encourage deeper thinking and comparison, often leading consumers to seek additional information or avoid purchasing to minimize risk (Chen et al., 2022). Furthermore, the perceived expertise and credibility of the review source significantly influence how consumers interpret and act on the information, with expert reviews holding greater sway over purchasing decisions (Filieri et al., 2018).

For businesses, online reviews play an equally significant role. Positive reviews act as endorsements, enhancing a company's credibility and attracting new customers. A high volume of favourable reviews can also improve a business's visibility in search engine results, as algorithms prioritize well-reviewed entities. Additionally, reviews provide invaluable feedback, offering businesses a direct line to customer sentiment and highlighting areas for improvement (Patil & Rane, 2023). Companies that actively engage with their reviews — by responding to feedback and addressing concerns — demonstrate a commitment to customer satisfaction, which can strengthen their reputation and foster loyalty. However, it is crucial that responses avoid self-promotion, as this can damage customer relations and reduce repurchase intentions (Li et al., 2020).

In this sense, online reviews have become a cornerstone of modern commerce, bridging the gap between consumers and businesses. Platforms such as Amazon, Airbnb, Yelp, and TripAdvisor have emerged as influential intermediaries, magnifying the role of online reviews across various industries (Pocchiari et al., 2024). These reviews empower consumers to make well-informed decisions and provide businesses with essential insights to build trust, refine their offerings, and adapt to the evolving demands of their customers. Thus, the strategic

management of online reviews is indispensable for businesses striving to succeed and remain competitive in today's digital economy.

2.2.2. Online movie reviews

Before the internet, movie discussions occurred primarily face-to-face. However, with the advent of Web 2.0, platforms such as Rotten Tomatoes and IMDb emerged as spaces where users could share opinions online. Social media further expanded these interactions, enabling individuals to not only share reviews but also engage directly with other viewers and related content (Oh et al., 2017). As a result, online movie reviews have become an integral part of the film industry, with consumers increasingly relying on these platforms to evaluate films before deciding whether to watch them (Kim et al., 2013).

As these platforms grew in influence, their impact on box office performance evolved. The effect of online user reviews on movie sales is now shaped not only by the ratings but also by the volume of reviews. While an individual review might have a limited impact on consumer decisions, the high frequency of reviews generates an "awareness effect", acting as a strong signal of word-of-mouth that significantly contributes to driving box office revenues (Duan et al., 2008). In addition to volume, the depth and credibility of online movie reviews also significantly impact box office performance. Casual moviegoers often write superficial reviews, which may lack depth and authenticity, whereas more engaged viewers tend to provide in-depth analyses. Authentic, credible reviews with detailed content can positively influence box office sales. In contrast, fake or manipulated reviews, which often resemble superficial content, tend to have a detrimental effect. However, even less credible reviews may still contribute to boosting sales (Kim et al., 2023).

Further expanding on this, Gupta et al. (2024) highlight the critical roles of both consumer engagement and external factors — such as media coverage — in shaping movie ratings on platforms like IMDb. External influences, including news articles for newly released films and the accumulation of awards for older movies, are crucial in driving positive ratings. Personal engagement, particularly the number of votes and likes on movie trailers (e.g., on YouTube), also plays an important role in boosting ratings. While interactive engagement, such as the number of user reviews and trailer comments, does contribute to ratings, its impact is generally less pronounced compared to personal engagement. These findings underline the importance of both external influences and personal engagement in determining how movies are evaluated online.

The combination of volume, credibility, and engagement creates a complex web of factors that influence the online reputation of movies, highlighting the evolving role of online reviews in shaping consumer perceptions and box office outcomes.

2.3. Sentiment analysis

2.3.1. Definition and importance of sentiment analysis

Sentiment Analysis, also known as Opinion Mining, is a subfield of Natural Language Processing (NLP) that focuses on identifying and extracting subjective information from texts (Wankhade et al., 2022). This process facilitates the determination of the polarity of emotions, such as happiness, sadness, hate, anger, or affection, as well as the opinions expressed in texts, reviews, posts, and other online content (Baid et al., 2017).

The origins of sentiment analysis can be traced back to early 20th-century research on public opinion. Initial studies, which primarily focused on post-World War II public opinion, such as views on communism in war-torn countries, laid the foundation for the field. However, sentiment analysis remained a relatively dormant area of study until the mid-2000s, when it gained prominence due to the increasing demand for and availability of online product reviews (Mäntylä et al., 2018). In recent years, its applications have expanded into diverse domains, including business, social media, finance, politics, and education. This growth reflects the increasing importance of sentiment analysis in understanding user opinions, monitoring brands and topics perception, and gathering valuable feedback to support strategic decision-making (Sharma et al, 2024).

2.3.2. Application of sentiment analysis in movie reviews

Extensive research on sentiment analysis in movie reviews has led to the development of a wide range of techniques. These methodologies have proven effective in evaluating film reception, identifying patterns in audience preferences, and supporting strategic decisions in the film industry.

Early studies, such as the one by Turney (2002), investigated sentiment classification using a lexicon-based approach, specifically focusing on semantic orientation to determine the sentiment of text. His objective was to assess whether unsupervised methods could effectively classify reviews as recommended or not recommended based on the co-occurrence of sentiment-laden words. To achieve this, he applied the method to a dataset of 410 reviews from Epinions, using a predefined set of positive and negative terms. The results demonstrated that while the semantic orientation approach achieved an accuracy of approximately 66% for movie

reviews, it performed better in other domains such as automobiles (84%) and banks (80%). This lower accuracy for movie reviews could be attributed to the complexities involved in analysing movies. Unlike other domains, movie reviews encompass both concrete elements, such as actors and plot events, as well as more abstract, subjective aspects, like the overall artistic style and tone of the film. For instance, while the phrase “more evil” may suggest a negative sentiment, an evil character does not necessarily imply that the film itself is bad. Nonetheless, Turney's study contributed significantly to the development of lexicon-based sentiment analysis and laid the groundwork for subsequent advancements in the field.

As a next step in advancing sentiment analysis, Pang et al. (2002), investigated sentiment classification using machine learning techniques, including Naive Bayes, maximum entropy, and Support Vector Machines (SVM). Their objective was to determine whether machine learning models could effectively classify text based on sentiment (e.g., positive or negative opinions) within movie reviews. For that, they used a dataset of 752 negative and 1301 positive movie reviews from 144 distinct reviewers, sourced from IMDb. The results demonstrated that machine learning methods, particularly SVM, were highly effective in sentiment classification, outperforming traditional rule-based approaches. This study provided a foundation for sentiment analysis across various domains, showcasing the potential of machine learning for analysing subjective content.

Building on this foundation, recent research has advanced the use of deep learning models. For instance, Ali et al. (2019) applied models such as Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM), Convolutional Neural Networks (CNN), and a hybrid CNN-LSTM model to the IMDb dataset, which consists of 50K movie reviews files (25K positive reviews and 25K negative reviews). Preprocessing included the use of Word2Vec for word embeddings. The findings revealed that the hybrid CNN-LSTM model achieved the highest accuracy (89.2%), surpassing individual models (CNN: 87.7%, MLP: 86.74%, LSTM: 86.64%). These results demonstrate the enhanced performance of deep learning models over traditional approaches like SVM, Naive Bayes, and Recursive Neural Tensor Networks (RNTN).

More recently, Danyal et al. (2024) applied advanced language models, such as XLNet and BERT, to the IMDb dataset (50K reviews) and the Rotten Tomatoes dataset. XLNet achieved the highest accuracy, with 93.48% on IMDB and 87.78% on Rotten Tomatoes. BERT also performed well, achieving 86.27% accuracy on IMDB and 83.38% on Rotten Tomatoes. Both XLNet and BERT consistently outperformed traditional machine learning methods, such as

SVM, Logistic Regression and Naive Bayes. These findings underscore the superior capabilities of XLNet and BERT in capturing nuanced sentiment, offering significant potential for improving personalised movie recommendations and targeted marketing strategies.

In parallel, prompt-based methods have recently been applied to sentiment analysis of movie reviews, offering an alternative paradigm to traditional supervised fine-tuning. Stilwell (2024), in his Master's thesis, applied prompt templates to the IMDb dataset and compared human-engineered prompts with trainable prompt-learning approaches. The first experiment (prompt engineering) evaluated a set of 12 manually designed prompts across several models (BERT, RoBERTa, DistilBERT, ELECTRA, GPT, and LLaMA 2) without fine-tuning, showing that performance varied depending on the formulation but could already achieve competitive results. The second experiment fine-tuned these models on datasets constructed from the completed prompts, a process referred to as prompt learning. This approach consistently improved performance across all models, with LLaMA 2 achieving very high accuracy (up to 98.5%), thereby demonstrating the effectiveness of prompt learning over prompt engineering in movie review sentiment analysis.

2.4. Aspect-based sentiment analysis

2.4.1. Definition of aspect-based sentiment analysis

Aspect-Based Sentiment Analysis (ABSA) is a subdomain of Sentiment Analysis that focuses on analysing the sentiment associated with specific aspects identified in a text (Pontiki et al., 2014). ABSA serves as a crucial tool for extracting and summarising opinions in online reviews (Hu & Liu, 2004).

Typically, ABSA involves four key elements: the aspect term, the aspect category, the opinion term, and the sentiment polarity. The aspect term refers to the explicit target of the opinion in the text. The aspect category represents a specific characteristic of an entity, predefined for a given domain. The opinion term is the expression used by the opinion holder to communicate their sentiment toward the target. Finally, sentiment polarity describes the orientation of the sentiment (positive, negative, or neutral) regarding an aspect term or category (Zhang et al., 2023). For instance, in the context of online movie reviews, the sentence "The acting was incredible" illustrates the following elements: "acting" (aspect term), "cast" (aspect category), "incredible" (opinion term), and "positive" (sentiment polarity). In this example, "acting" and "incredible" are explicitly mentioned, while "cast" and "positive" belong to predefined categories and sentiments (Figure 2).

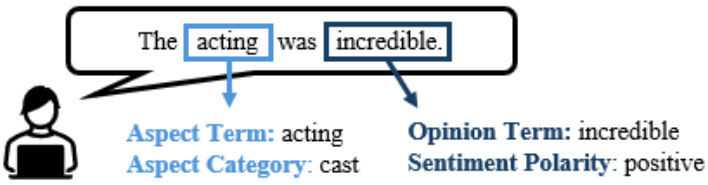


Figure 2 - Example of the four key sentiment elements in ABSA

In this sense, the ABSA solution involves several distinct tasks. According to Li et al. (2022), the fundamental tasks include Aspect Extraction (AE), Opinion Extraction (OE), and Aspect Sentiment Classification (ASC).

2.4.2. Aspect extraction (AE)

Aspect Extraction involves identifying the aspects of an entity about which opinions are expressed (Luo et al., 2019). These aspects can be either explicit, when directly mentioned in the text, or implicit, when inferred from the context (Nazir et al., 2022).

For instance, in the sentence "The storyline was very well constructed and full of surprises," the aspect "plot" is explicit because the term "storyline" is explicitly mentioned. Conversely, in the sentence "I didn't expect that ending!" the aspect "plot" is implicit, as the comment refers to the story's conclusion without directly mentioning the plot. Identifying implicit aspects remains a challenging domain in ABSA literature due to its complexity and ambiguous nature (Maitama et al., 2020).

There are three main approaches to aspect extraction: supervised, semi-supervised, and unsupervised. Each approach has its advantages and limitations, depending on the availability of annotated data and the application context. Supervised approaches require large volumes of annotated data and utilise techniques such as Conditional Random Fields (CRF) (Yin et al., 2016), Recurrent Neural Networks (RNN) (Liu et al., 2015), Convolutional Neural Networks (CNN) (Xu et al., 2018), and domain-specific fine-tuned BERT models (Xu et al., 2019). While these methods achieve high precision, they demand significant effort in data annotation. Semi-supervised approaches combine annotated and unannotated data, leveraging techniques like progressive self-training (Wang et al., 2021) to reduce the reliance on extensive annotated datasets. Unsupervised approaches do not depend on annotated data (Zhang et al., 2023).

2.4.3. Opinion extraction (OE)

Opinion Extraction focuses on identifying expressions of opinion related to specific aspects. This task is often framed as a Token Classification problem, where models can simultaneously

extract aspect and opinion terms or, alternatively, identify opinions associated with pre-identified aspects in the text (Zhang et al., 2023).

To illustrate, in the sentence "The movie's music was incredible, but the ending didn't surprise me," the model would extract the opinion terms "incredible" (associated with the aspect term "music") and "didn't surprise me" (associated with the implicit aspect "plot").

A range of techniques is employed for opinion extraction. Dependency tree-based models and attention mechanisms are commonly used to map relationships between aspects and opinions. LSTM networks (Fan et al., 2019) are frequently applied to incorporate aspect information and transfer sentiment analysis knowledge. More advanced methods, such as BiLSTM (Mensah et al., 2021) with positional embeddings and Graph Convolutional Networks (GCN) (Kipf & Welling, 2017), have also been introduced to capture syntactic and structural relationships in the text (Zhang et al., 2023).

2.4.4. Aspect sentiment classification (ASC)

Aspect Sentiment Classification (ASC) determines the sentiment polarity (positive, negative, or neutral) associated with specific aspects in a text (Zhou et al., 2019). For example, in the comment "The performance of the lead actors was excellent, but the story was predictable", the aspect categories are Cast and Plot with positive and negative sentiments, respectively.

Recent advances in deep learning have significantly improved performance in this area. Satyarthi and Sharma (2023) report that combining architectures such as LSTM, GCN, and BERT has proven to be highly effective for ASC tasks.

2.5. Aspect-based sentiment analysis in movie reviews

Aspect-Based Sentiment Analysis (ABSA) applied to online movie reviews enables the identification of the cinematic elements most valued by audiences. This task focuses on extracting specific aspects of a film, such as the plot, direction, cast, and soundtrack, and analysing the sentiment polarity associated with each. Some studies in the literature have employed various methods to achieve this.

Thet et al. (2010) utilised a lexical approach to analyse movie reviews in discussion forums on IMDb. Their method calculates the sentiment of each clause based on the grammatical structure of sentences, leveraging predefined sentiment scores from SentiWordNet (Esuli & Sebastiani, 2006). The study also emphasizes the importance of creating domain-specific

keyword lists to identify different film aspects, facilitating the categorization of opinions. In terms of performance, the authors reported clause-level sentiment classification accuracies of 75% for overall movie sentiment, 86% for director, 83% for cast, 80% for story, 90% for scene, and 81% for music, demonstrating the effectiveness of their approach across multiple cinematic dimensions.

Mir and Mahmood (2020) proposed the Movie Aspects Identification Model (MAIM), which employs a hybrid BiLSTM-CRF technique to identify specific movie aspects and named entities, such as names of people and movie titles. The model detects both frequent and infrequent aspects, assigning sentiment values to each. Notably, this model introduces an annotation process for classifying aspects, entities, and sentiment words, alongside an aspect pruning method to eliminate irrelevant data. Using the IMDb dataset, MAIM reached strong results, with a precision of 89.9%, recall of 88.9%, and an F1-score of 89.4%. These values indicate a clear improvement over the baseline CRF and LSTM-CRF models, confirming the advantage of combining BiLSTM with CRF for ABSA.

Wang et al. (2020) developed a filtering mechanism to identify words associated with various movie aspects using a modified lexicon. These words were then used to calculate sentiment intensity via the VADER model, which is particularly effective in analysing informal texts due to its ability to handle slang, emojis, and abbreviations commonly found in online reviews, such as those on IMDb. Their model proved to be both feasible and effective, achieving accuracies of 81% for overall sentiment, 83% for actors, 80% for director, 84% for plot, and 77% for music. These results highlight the capacity of lexicon-based filtering combined with VADER to capture emotions tied to distinct cinematic components.

Onalaja et al. (2021) applied both supervised models (e.g., Logistic Regression, Naive Bayes, SVM) and deep learning models (e.g., Recurrent Neural Networks) to classify sentiments linked to movie aspects. Using the IMDb dataset, the researchers employed the spaCy tool for entity extraction, along with techniques such as Latent Dirichlet Allocation (LDA) for topic identification and TF-IDF and CountVectorizer for text vectorization. Despite challenges in constructing precise lexicons for aspect-related terms, the study showed that including driving factors such as aspect and film genre increased accuracy by 3–4% on average, with the best configuration achieving 68% compared to 63% without them. This improvement resulted in more effective sentiment predictions.

Recent studies, such as those by Kit and Joseph (2023) and Horsa and Tune (2023), have also applied machine learning methods. Kit and Joseph, working with the IMDb dataset, first

extracted aspects using a keyword-based list, which provided the labels necessary for supervised classification. With these annotations, Decision Trees achieved 98% accuracy for aspect prediction, while Logistic Regression reached 92%. For sentiment analysis, Logistic Regression outperformed other models with 93% accuracy, compared to 91% with Multinomial Naïve Bayes. These results suggest that Decision Trees are particularly effective for aspect prediction, whereas Logistic Regression is more suitable for sentiment classification.

On the other hand, Horsa and Tune addressed aspect-based sentiment analysis in an underexplored language, Afaan Oromoo, by collecting 2,800 YouTube movie reviews through the YouTube Data API. Seven predefined aspects of the reviews were manually annotated by three human annotators, who also assigned positive or negative sentiment to each aspect occurrence. Annotation disagreements were resolved using Cohen's Kappa, ensuring reliability of the gold standard dataset. With this annotated data, the authors trained machine learning models—Random Forest, Logistic Regression, SVM, and Multinomial Naïve Bayes—using Bag of Words (BoW) and TF-IDF for text representation. Their experiments showed competitive results, with Random Forest and Multinomial Naïve Bayes both achieving 88% accuracy, SVM reaching 88% with BoW and 87% with TF-IDF, and Logistic Regression obtaining 87% in both configurations. These findings demonstrate that even in low-resource language settings, traditional machine learning methods paired with simple text representations can achieve robust performance.

Chapter 3 – Research methodology

Research methodology refers to the systematic plan that guides the processes of collecting, analysing, and interpreting data to address the research questions or test hypotheses. It provides the rationale for the selection of specific methods and procedures, ensuring that the study is conducted in a structured, valid, and replicable manner (Creswell & Creswell, 2017).

3.1. Research design

As outlined in Chapter 1, this study addresses the research question: *What cinematic aspects identified in movie reviews have the greatest impact on their respective ratings?* To answer this question, a quantitative research design was adopted since it allows systematic measurement of data and the use of statistical techniques to identify relationships between variables. In this context, it enables the comparison between aspect-level sentiments and overall ratings, supporting objective analysis and facilitating replicability.

Within this framework, aspect-based sentiment analysis (ABSA) was applied to online movie reviews. This method makes it possible to decompose overall opinions into specific aspects and assign sentiments to each, which is essential for understanding which dimensions of a film most strongly influence audience evaluations. The overall process was structured into four sequential phases, illustrated in Figure 3.

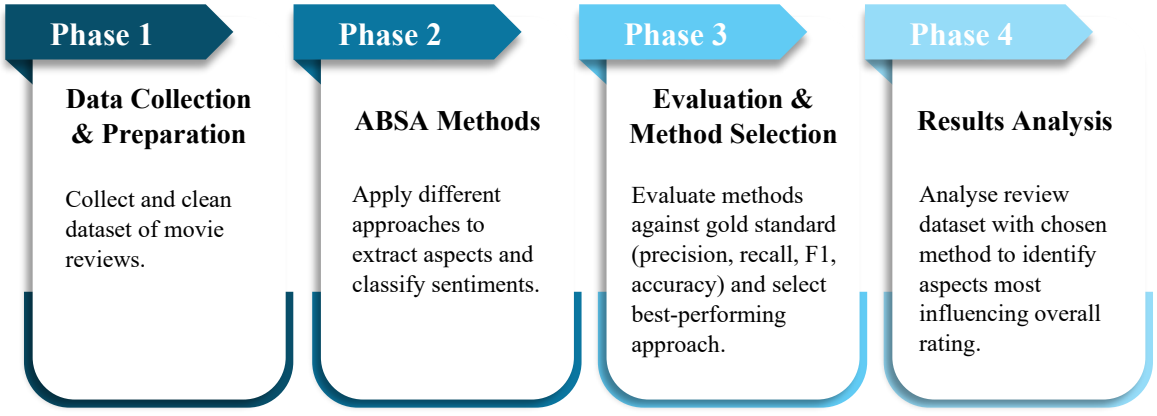


Figure 3 - Research design of the study

3.2. Data collection and preparation

3.2.1. Dataset selection

The dataset employed in this study (Benlahbib, 2019) consists of 1,000 reviews covering 10 films, with 100 reviews per film. The films included are: *2012; A Beautiful Mind; Amadeus; Avatar; Clash of the Titans; Les Miserables; Star Wars Episode I - The Phantom Menace; The*

Expendables I; The Godfather; The Matrix Revolution. The reviews were randomly extracted and selected to ensure representativeness based on IMDb users' weighted average ratings.

Each record in the dataset contains four fields: (1) the *review*, which is the full user-generated text; (2) the *polarity*, a manually annotated sentiment classification where 1 indicates a positive review and 0 indicates a negative review; (3) the *rating*, a numerical score assigned by the reviewer ranging from 1 to 10; and (4) the *movie*, referring to the title of the reviewed film, as shown in Table 1. In total, the dataset comprises 756 positive and 244 negative reviews.

The selection of this dataset was guided by the requirement that it should contain the titles of the reviewed films, the numerical ratings, and the sentiment polarity annotations. The presence of movie titles is particularly important because it enables the later identification of aspects related to actors and directors. For this reason, the widely used IMDb dataset of 50,000 reviews (Maas et al., 2011) was not selected, since it does not include film titles and therefore would not allow this type of analysis. The numerical ratings are also essential for determining which aspects most influence the overall evaluation of a film, while the sentiment polarity annotations enable comparisons between manually annotated sentiment labels and the results produced by automated sentiment analysis models. In addition, since no publicly available dataset with pre-annotated aspects for movie reviews was found, it was necessary to select a corpus that at least fulfilled the other essential requirements of this research — film titles, numerical ratings, and sentiment polarity annotations. Aspect identification was therefore implemented as part of the methodology.

Table 1 - Example of the dataset structure with one sample review

Review	polarity	rating	movie
I like John Cusack. He usually makes some pretty good movies. This movie is a dog. I know movies stretch the imagination, but this one wants you to remove you head and not even think. The physics are just WAAAAAAY to hard to believe. Earthquakes and volcanic eruptions are nowhere near as big as they are in this movie. And its just a stupid plot all together.	0	3	2012

3.2.2. Gold standard construction

In order to evaluate the performance of the aspect identification methods described in Section 3.3 (*ABSA Methods*), a gold standard was constructed based on a stratified random sample of 100 reviews, with 10 reviews taken from each of the 10 films in the dataset to ensure balance. The presence of each aspect in a review was manually annotated using a binary scheme

(1 = aspect identified; 0 = aspect not identified), independently of whether the aspect had an associated sentiment.

Table 2 provides an example of the manual aspect annotation format used in the evaluation.

Table 2 - Example of manual aspect annotation for a single review

review_id	review	movie	cast	directing	plot	ambience
305	This movie was a complete shocker. The acting was so bad I could not stay with it. When acting is terrible it is tough to stay "in the movie". They really should have cast some more accomplished actors. The scenery was so terrible. It was just too fake and plastic looking for me to settle in and enjoy it. The aspect that really got me was the predictable and boring story-line. It was so predictable that I considered switching it off numerous times. The only reason I watched the entire movie is because of the hype that it generated at the box office. Boring!	Avatar	1	0	1	1

To ensure reliability in the manual annotation process, two independent annotators evaluated the presence of each aspect in the selected reviews. Inter-annotator agreement was then assessed using Cohen's Kappa coefficient (Cohen, 1960), a statistical measure that quantifies the degree of agreement between annotators while accounting for the possibility of agreement occurring by chance. Cohen's Kappa (κ) is defined as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{1}$$

where p_o represents the observed proportion of agreement between annotators, and p_e represents the proportion of agreement expected by chance. The results for each aspect are presented in Table 3.

Table 3 - Cohen's Kappa values for inter-annotator agreement by cinematic aspect

Aspects	Cohen's Kappa (κ)
Plot	0.689
Cast	0.746
Directing	0.783
Ambience	0.753

For interpretation, the scale proposed by Landis and Koch (1977) was adopted (Table 4).

Table 4 - Interpretation of Cohen's Kappa values according to Landis and Koch (1977)

Kappa value	Strength of agreement
< 0	Poor
0.00–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

According to this interpretation scale, all aspects achieved substantial agreement (0.61–0.80). Directing recorded the highest agreement ($\kappa = 0.783$), followed closely by Ambience ($\kappa = 0.753$) and Cast ($\kappa = 0.746$). Plot achieved the lowest agreement ($\kappa = 0.689$), although it still falls within the substantial agreement range. The lower agreement for Plot may be attributed to its broader and more subjective nature, which can lead to greater differences in annotators' judgments compared to more concrete aspects such as Directing.

In total, 34 reviews contained at least one disagreement between annotators. In these cases, a third annotator was consulted to adjudicate the correct label. The final adjudicated dataset was used as the gold standard for the evaluation of the automatic aspect identification methods presented in Section 3.3 (*ABSA methods*).

3.2.3. Text preprocessing

The preprocessing stage began with the segmentation of each review into individual sentences using the *sent_tokenize* function from the NLTK library (Bird et al., 2009). This function is based on the Punkt tokenizer, a pre-trained unsupervised model that exploits punctuation and statistical distributions of character sequences to identify sentence boundaries with high accuracy (Kiss & Strunk, 2006). Structuring the dataset at the sentence level was particularly important for aspect-based sentiment analysis, since a single review often contains evaluations of multiple aspects of a movie, typically articulated in separate sentences. To ensure traceability, each sentence was stored together with its corresponding review identifier, sentiment polarity, rating, and movie title, thereby maintaining the link with the original dataset.

After segmentation, all sentences were converted to lowercase to ensure consistency during subsequent processing steps. No additional normalization steps, such as lemmatization or

stopword removal, were applied, as they were not necessary for the ABSA methods employed in this study.

An illustrative excerpt of the dataset after sentence segmentation and preprocessing is presented in Table 5. It should be noted that polarity and rating values correspond to the full review, not to the individual sentence.

Table 5 - Example of the dataset after sentence segmentation and text preprocessing

review_id	sentence	preprocessed_sentence	polarity	rating	movie
1	I have to say this movie is very tense.	i have to say this movie is very tense.	1	9	2012

3.3. ABSA methods

To perform ABSA on the collected dataset, three different methods were implemented, deliberately selected to complement established approaches in the literature and to ensure a representative coverage of sentiment analysis techniques. These methods differ mainly in how aspects are identified within the reviews and in the sentiment analysis models applied to those aspects.

The first two approaches follow a pipeline structure: aspects are identified through keyword-based rules applied to the preprocessed, sentence-level dataset, and sentiments are then classified using either (i) the lexicon-based VADER model or (ii) the transformer-based DistilBERT. The third approach adopts an end-to-end structure, employing large language models (GPT-4o mini and GPT-4.1 mini) prompted to extract aspects and their associated sentiments directly from the raw reviews.

The implementation of these three methods made it possible to identify the approach that provides the most consistent and reliable basis for subsequent analysis. This process contributes to ensuring that the aspects detected and their associated sentiments more faithfully approximate the evaluations articulated by reviewers, thereby enhancing the validity and robustness of the response to the research question.

3.3.1. Pipeline-based approach

3.3.1.1. Aspect identification

In the pipeline approaches, the first step was aspect detection, which relied on a predefined set of keywords covering four cinematic dimensions — Plot, Cast, Directing and Ambience —

together with a residual category, General, used whenever no aspect-specific evidence was found in a review. The keywords list was adapted from Kit and Joseph (2023), who compiled it based on previous research, surveys, questionnaires, and interviews with movie enthusiasts and content creators. Table 6 presents the aspects, their definitions, and the corresponding keywords.

Table 6 - Aspects, definitions, and keywords (adapted from Kit and Joseph, 2023)

Aspect	Definition	Keywords
Plot	Represents the story of the movie	plot, story, storyline, ending, storytelling, drama, writing, twist, script, end, movie
Cast	Represents the actors and their performances.	acting, role, character, act, actress, actor, villain, protagonist, antagonist, performance, performed, play, played, playing, casting, cast, crew, artist, portray
Directing	Represents the flow of the movie and the method by which it was directed.	direct, directing, direction, filming, cinematography, filmmaker, cinematic, director
Ambience	Represents the immersive elements of the movie such as visuals and sounds	visual, effect, animation, cgi, graphics, scenery, stunt, design, audio, sound, music, track
General	Represent the reviews not mentioning any of the above aspects	-

To improve coverage, the names of actors and directors associated with the films in the dataset were retrieved using the OMDb API[2] and incorporated into the keyword lists for the Cast and Directing aspects. This ensured that references to specific individuals (e.g., *Al Pacino*, *Keanu Reeves* for Cast; *Francis Ford Coppola*, *James Cameron* for Directing) were accurately captured. The complete lists of names are provided in Appendix A.

The detection process was implemented using substring matching, which allowed morphological variants (e.g., *direct* in *directing*) to be captured without requiring additional normalization. Sentences with no aspect-specific keywords remained unassigned and the residual category General was only applied when no aspect was detected in any sentence of the review.

[2] https://www.omdbapi.com/

3.3.1.2. Sentiment analysis

Based on the previous aspect annotations, sentiment analysis was conducted at the sentence level using VADER and DistilBERT. When multiple aspects were identified in the same sentence, the same sentiment label was assigned to all of them.

VADER (Hutto & Gilbert, 2014) is a rule-based model designed for short, informal texts, combining a sentiment lexicon with heuristics to capture polarity and intensity. For each sentence, VADER produces a compound score ranging from –1 (most negative) to +1 (most positive). Following standard thresholds, scores ≥ 0.05 were classified as positive, scores ≤ –0.05 as negative, and those in between as neutral.

For DistilBERT, the pre-trained *distilbert-base-uncased-finetuned-sst-2-english* model from the Hugging Face platform[3] was employed. This variant is fine-tuned on the Stanford Sentiment Treebank (SST-2) dataset for binary sentiment classification, producing probabilities between 0 and 1 for the *positive* and *negative* classes. DistilBERT (Sanh et al., 2019) applies knowledge distillation to compress BERT (Devlin et al., 2019), reducing the number of parameters by 40% and achieving 60% faster inference, while retaining approximately 97% of BERT's performance on NLP benchmarks. By design, it remains a context-aware model capable of capturing dependencies between words in a sentence, in contrast to lexicon-based methods such as VADER, which rely on predefined vocabularies and heuristic rules.

Although the model is intrinsically binary, a neutral class was introduced in order to maintain consistency with other approaches considered in this study, which classify sentences into three categories: positive, negative, or neutral. This extension is also justified by the fact that many reviews contain a balanced mixture of positive and negative statements, where the overall polarity cannot be clearly assigned to one side. To operationalise this adjustment, the probability score returned by DistilBERT was assigned a positive sign if the predicted label was *positive* and a negative sign if it was *negative*. These signed values were then aggregated by review and by aspect and the same thresholding scheme applied with VADER was adopted: aggregated scores ≥ 0.05 were classified as positive, those ≤ –0.05 as negative, and values between –0.05 and 0.05 were labelled as neutral. In this way, the neutral category reflects cases where positive and negative statements offset each other, resulting in no clear evaluative orientation at the review level.

[3] https://huggingface.co/

For both models, the sentence-level results were then aggregated at the review level in three complementary ways. First, the frequency of each aspect was determined by counting how many times it was mentioned across the sentences of a review. Second, the average sentiment per aspect was calculated as the mean of the compound scores of all sentences in which that aspect appeared, thereby capturing the overall attitude expressed towards each cinematic dimension. Lastly, the overall sentiment of the review was computed as the mean of all sentence-level compound scores, ensuring that both positive and negative statements distributed across different sentences contributed proportionally to the final evaluation.

3.3.2. End-to-end approach

The end-to-end approach employed a prompt-based approach with large language models (LLMs). In this case, GPT-4o mini and GPT-4.1 mini were used to extract both aspects and associated sentiments directly from the raw review text, without additional preprocessing. Unlike the pipelines approaches described earlier, this method performs end-to-end aspect-based sentiment analysis, with the model simultaneously responsible for aspect identification, sentiment classification, and structured output formatting.

A detailed prompt was designed to enforce a structured output format that would allow systematic analysis of the results. The instruction specified: (i) the aspect categories and their scope; (ii) disambiguation rules (e.g., restricting Ambience to visual and sound elements, preventing generic adjectives such as *epic* or *fun* from triggering aspect matches); (iii) named entities to be recognised as Cast or Directing mentions; (iv) rules ensuring that each aspect was counted at most once per sentence; (v) independence of aspect detection from sentiment expression; and (vi) the sentiment labelling scheme. Aspect sentiment values were constrained to the interval $[-1,1]$, while overall sentiment was assigned using thresholds consistent with the previous methods (≥ 0.05 = positive; ≤ -0.05 = negative; otherwise neutral).

The output schema enforced by the prompt was a table containing, for each review, the counts and sentiments of all aspects (Plot, Cast, Directing, Ambience, General), together with overall sentiment and overall label. This ensured that results from the LLM-based approach could be integrated seamlessly into the analysis pipeline and compared directly with those obtained from the keyword-based methods.

The complete prompt and model execution parameters are provided in Appendix B to ensure transparency and reproducibility.

3.4. Method evaluation and selection

The evaluation of the methods described in Section 3.3. was conducted using four performance metrics: precision, recall, F1 score, and accuracy. These metrics are defined as:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

Precision measures the proportion of retrieved instances that are relevant, while recall measures the proportion of relevant instances that are correctly retrieved (Manning et al., 2008). The F1 score, defined as the harmonic mean of precision and recall, provides a balanced measure that accounts for both false positives and false negatives. Accuracy, defined as the proportion of correctly classified instances over the total number of instances, is also reported for completeness, although it is known to be less informative in cases of class imbalance (Powers, 2011).

Aspect extraction was evaluated separately for each cinematic dimension by comparing the performance of the keyword-based approach, GPT-4o mini, and GPT-4.1 mini against the manually annotated gold standard introduced in Section 3.2.2.

For sentiment classification, the evaluation was conducted by comparing the outputs of VADER, DistilBERT, GPT-4o mini, and GPT-4.1 mini against the polarity labels provided in the dataset, in which each review was annotated as positive (1) or negative (0). Unlike the dataset labels, which are limited to binary annotations (positive or negative), the sentiment analysis methods under evaluation produced three categories at the review level: positive, negative, and neutral. To maintain consistency in the number of reviews evaluated across methods, as the number of neutral predictions varied between them, the neutral class was not discarded. Instead, two evaluation strategies were applied: (1) mapping neutral predictions to negative, and (2) mapping neutral predictions to positive. This procedure ensured comparability while also providing insights into how neutral instances influenced the alignment between the automatic methods and the gold standard binary labels.

During this process, it was identified that in a small number of cases (12 reviews, 1.2% of the dataset), the GPT-4o mini model did not return an overall sentiment. In these instances, the output of GPT-4.1 mini was adopted as a fallback, ensuring completeness and consistency of the dataset prior to evaluation.

Since the ABSA methods under evaluation operate as integrated pipelines, their components cannot be combined across approaches. The selection of the most suitable method for subsequent analyses was therefore based on the overall performance of each complete pipeline, jointly considering both aspect extraction and sentiment classification. The comparative results of this evaluation are presented in Section 4.1.

Chapter 4 – Results presentation and analysis

This chapter begins by presenting the comparative evaluation of the ABSA methods introduced in the previous chapter, covering both aspect extraction and sentiment classification. The selection of the most suitable method for subsequent analyses was based on these results. The analysis is then structured progressively, moving from review-level examinations to dominant aspect evaluations and finally to aspect combinations. This layered approach ensures that the research question *Which cinematic aspects identified in movie reviews have the greatest impact on their respective ratings?* is addressed comprehensively, both descriptively and inferentially.

4.1. Evaluation of ABSA methods

This section reports the comparative evaluation results of ABSA methods, beginning with aspect extraction and sentiment classification, and concluding with the method selection.

4.1.1. Aspect extraction performance

Aspect extraction performance was evaluated for each cinematic dimension. The results are presented in Table 7, while Table 8 reports the macro-averaged and micro-averaged scores that summarise overall performance across the three methods.[4]

Table 7 - Evaluation of aspect extraction by cinematic aspect

Methods	Aspect	Precision	Recall	F1	Accuracy
Keyword-based	Plot	0.830	0.963	0.891	0.810
	Cast	0.831	**0.987**	**0.902**	0.840
	Directing	**0.865**	0.914	0.889	**0.920**
	Ambience	0.857	0.787	0.821	0.790
GPT-4o mini	Plot	0.802	0.951	0.870	0.770
	Cast	0.703	0.693	0.698	0.550
	Directing	0.444	0.457	0.451	0.610
	Ambience	0.648	0.574	0.609	0.550
GPT-4.1 mini	Plot	0.811	0.951	0.875	0.780
	Cast	0.705	0.733	0.719	0.570
	Directing	0.378	0.400	0.389	0.560
	Ambience	0.661	0.672	0.667	0.590

[4] Macro-averaging gives equal weight to each aspect by averaging metrics across aspects, whereas micro-averaging pools all instances, giving more weight to frequent aspects.

Table 8 - Overall evaluation of aspect extraction (macro and micro averaging)

Methods	Averaging	Precision	Recall	F1	Accuracy
Keyword-based	Macro	**0.846**	**0.913**	**0.876**	**0.840**
	Micro	**0.841**	**0.921**	**0.879**	**0.840**
GPT-4o mini	Macro	0.649	0.669	0.657	0.620
	Micro	0.692	0.714	0.703	0.620
GPT-4.1 mini	Macro	0.639	0.689	0.662	0.625
	Micro	0.688	0.742	0.714	0.625

The evaluation results presented in Table 7 and Table 8 reveal marked performance differences between the three aspect identification methods. The keyword-based approach emerged as the most effective, consistently surpassing the GPT-based alternatives across all metrics, both at the aspect level and in the aggregated macro- and micro-averages. This superiority can be attributed to the domain-specific lexicon employed, which captures explicit references to cinematic dimensions more reliably than general-purpose prompting.

At the aspect level (Table 7), the keyword-based method achieved the highest F1-scores for Plot (0.891), Cast (0.902), and Directing (0.889), supported by particularly strong recall for Plot (0.963) and Cast (0.987). This indicates that nearly all relevant mentions were retrieved, while precision ensured that false positives remained limited. The only aspect with comparatively lower results was Ambience, which proved challenging for all methods. This difficulty reflects the inherent subjectivity and ambiguity of references to visual and sound effects, which are often conveyed indirectly. Nonetheless, even in this category, the keyword-based method attained an F1-score of 0.821, outperforming the GPT-based approaches.

In contrast, GPT-4o mini and GPT-4.1 mini delivered weaker and less balanced results, particularly for Directing (F1 below 0.46), reflecting both missed detections and misclassifications, which likely stem from the indirect way this aspect is expressed in reviews, often through comments on pacing, narrative flow, or scene composition. Although GPT-4.1 mini obtained slight improvements in Ambience (F1 = 0.667), these were not enough to narrow the performance gap.

Aggregated results (Table 8) reinforce these findings. The keyword-based method reached macro- and micro-averaged F1-scores of 0.876 and 0.879, far ahead of GPT-4o mini (0.657 and 0.703) and GPT-4.1 mini (0.662 and 0.714).

In summary, the keyword-based method demonstrated greater robustness and reliability in identifying aspects in movie reviews. Its consistently high recall ensured that nearly all relevant mentions were captured, while its high precision minimised false positives, resulting in superior F1-scores and accuracy. In contrast, the GPT-based methods underperformed, particularly in aspects that are less explicitly expressed or more context-dependent, such as Directing and Ambience. These findings highlight that, within the present experimental setup, rule-based keyword detection remains more effective than prompt-based large language models for ABSA in the film domain, unless the latter are fine-tuned or otherwise adapted to the task.

4.1.2. Sentiment classification performance

Sentiment classification performance was evaluated by comparing the outputs of VADER, DistilBERT, GPT-4o mini, and GPT-4.1 mini against the dataset labels. To account for the presence of the neutral class, results are reported under two complementary evaluation strategies. Table 9 presents the outcomes when neutral predictions are mapped to the negative class, whereas Table 10 shows the results when neutral predictions are mapped to the positive class. This dual reporting allows for a more balanced interpretation of performance across methods.

Table 9 - Performance of sentiment classification (Neutral mapped to Negative)

Methods	Precision	Recall	F1	Accuracy
VADER	0.858	0.796	0.826	0.746
DistilBERT	0.962	0.706	0.815	0.757
GPT-4o mini	**0.989**	0.819	0.896	0.856
GPT-4.1 mini	0.981	**0.866**	**0.920**	**0.886**

Table 10 - Performance of sentiment classification (Neutral mapped to Positive)

Methods	Precision	Recall	F1	Accuracy
VADER	0.817	0.906	0.859	0.776
DistilBERT	0.950	0.798	0.867	0.815
GPT-4o mini	**0.980**	0.854	0.913	0.877
GPT-4.1 mini	0.968	**0.909**	**0.937**	**0.908**

The evaluation results presented in Table 9 and Table 10 reveal consistent performance differences between the four sentiment classification methods, depending on how neutral cases are handled.

When neutral instances were mapped to the negative class (Table 9), the GPT-based approaches demonstrated the strongest balance. GPT-4.1 mini achieved both high precision (0.981) and high recall (0.866), resulting in the highest F1-score (0.920) and accuracy (0.886). This indicates that it successfully identified most true cases while maintaining a low rate of false positives. GPT-4o mini prioritised precision even further (0.989), producing highly reliable predictions, though its slightly lower recall (0.819) shows that it overlooked more positive cases compared to GPT-4.1 mini. DistilBERT also displayed very high precision (0.962) but with substantially lower recall (0.706), reflecting a conservative classification style that minimised false positives at the expense of missing a considerable proportion of true cases. By contrast, VADER reached lower levels of precision (0.858) and recall (0.796), resulting in weaker F1 (0.826) and accuracy (0.746), underscoring the limitations of its lexicon-based design.

When neutral cases were instead mapped to the positive class (Table 10), the overall ranking of models remained stable, though the precision–recall trade-offs shifted. GPT-4.1 mini once again showed the most balanced performance, with precision of 0.968 and recall of 0.909, yielding the highest F1-score (0.937) and accuracy (0.908). GPT-4o mini maintained its precision-oriented behaviour (0.980), though at the cost of lower recall (0.854), leading to an F1-score of 0.913. DistilBERT behaved consistently with the previous setting, showing strong precision (0.950) but weaker recall (0.798), which limited its F1 to 0.867. Notably, VADER achieved almost the highest recall of all methods (0.906), capturing most true cases, but its lower precision (0.817) revealed a tendency to misclassify neutral or negative reviews as positive.

The imbalance in the dataset, with a significantly larger proportion of positive reviews (756 positives vs. 244 negatives), partly explains VADER's performance. Its lexicon-based rules are biased towards positive polarity, which helps it capture most true positives but simultaneously increases the likelihood of false positives, thus lowering precision.

Overall, the GPT-based models clearly outperformed both VADER and DistilBERT under both evaluation schemes, with GPT-4.1 mini consistently achieving the best balance between precision and recall. Nevertheless, the results also highlight the complementary strengths of the other methods: DistilBERT may be preferable when minimising false positives is crucial, whereas VADER can be advantageous in contexts where maximising recall is more important than overall accuracy, particularly in datasets with a strong positive bias. These findings

underline the importance of considering multiple evaluation metrics, as each sheds light on a different dimension of model performance.

4.1.3. Comparative assessment and method selection

The comparative evaluation of methods highlighted different strengths in aspect extraction and sentiment classification. For aspect identification, the keyword-based approach achieved the most reliable results, with substantially higher precision, recall, and F1 scores than the GPT-based methods, both at the aspect level and in the aggregated averages.

In the case of sentiment classification, GPT-based approaches, especially GPT-4.1 mini, achieved the highest scores, showing superior balance between precision and recall. DistilBERT also performed well, particularly in terms of precision, while VADER remained the weakest alternative.

Considering both dimensions together, the Keyword-based + DistilBERT method was selected as the most suitable approach for this study. This choice balances the robustness of keyword-based aspect extraction with the stronger performance of DistilBERT in sentiment classification, resulting in a reliable and coherent pipeline for Aspect-Based Sentiment Analysis of movie reviews. The selection prioritises accurate identification of aspects while ensuring consistent sentiment detection, providing a solid basis for the subsequent analysis of which aspects most influence overall movie evaluation.

4.2. Review-level analyses

This first stage of analysis considers the role of all aspects as they are mentioned across reviews. The aim is to establish how frequently each aspect is addressed, how sentiments are distributed, and how these evaluative tendencies align with the overall ratings provided by users.

4.2.1. Aspect prevalence, sentiment, and co-occurrence

Figure 4 presents the distribution of aspect mentions across the dataset, distinguishing between the proportion of reviews in which each aspect is mentioned at least once and the total number of times the aspect is referred to across all reviews. The results reveal a clear hierarchy of aspect mentions: Plot is referred to in 95.8% of reviews, meaning that nearly all texts mention the storyline at least once, with a total of 4,586 mentions. Cast appears in 91.5% of reviews (3,777 mentions in total), confirming that actors' performances are also a central focus of audience discourse. By contrast, Ambience is mentioned in 59.3% of reviews (1,288 mentions) and Directing in only 41.4% (670 mentions). This distinction shows that while all four aspects

contribute to the evaluative framework, Plot and Cast dominate both in breadth (presence across most reviews) and depth (high frequency of references), whereas Ambience and Directing play more peripheral roles.

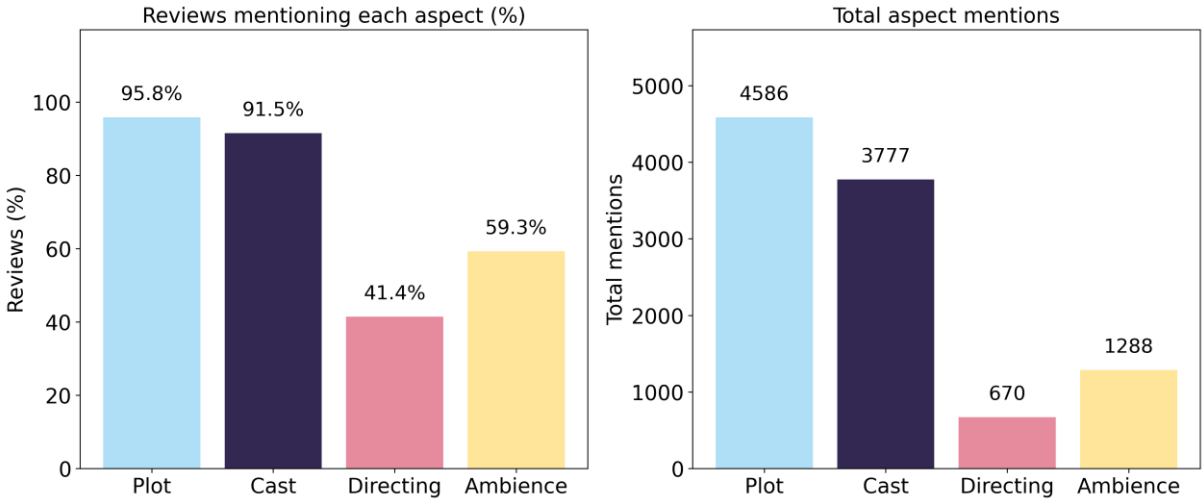


Figure 4 - Review-level analyses: reviews mentioning each aspect (left) and total aspect mentions (right).

Turning to sentiment, Figure 5 shows that all aspects are, on average, evaluated positively, which is consistent with the fact that the ten films in the dataset have relatively high ratings (around 7 on average). Nonetheless, the magnitude of sentiment varies across aspects: Directing achieves the highest mean sentiment score (0.35), followed by Ambience (0.29), Cast (0.21), and Plot (0.16). This pattern suggests two complementary mechanisms. On the one hand, aspects that are less frequently discussed, such as Directing, tend to be mentioned in particularly positive contexts, raising their average. On the other hand, Plot and Cast, being central and more scrutinised, attract a greater diversity of opinions, including criticism, which lowers their mean despite their predominance in frequency.

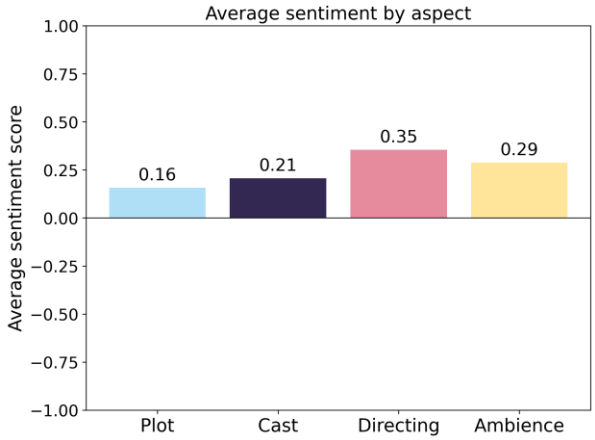


Figure 5 - Review-level analyses: average sentiment by aspect

The distribution of sentiments confirms this duality (Figure 6). Positive evaluations prevail across all aspects, especially for Directing (65.9%) and Ambience (61.9%), whereas Plot (54.6%) and Cast (56.5%) attract a larger share of negative opinions (36.6% and 35.0%). Neutral mentions remain residual, between 5.6% and 8.8%, showing a general tendency for reviewers to express clear stances.

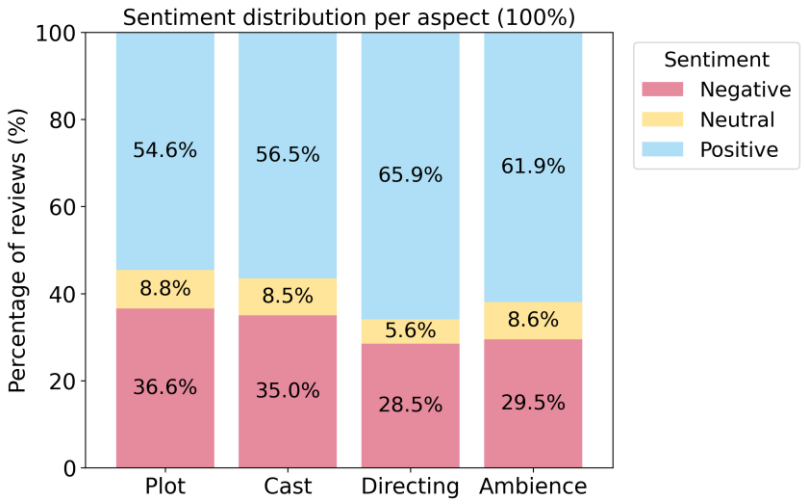


Figure 6 - Review-level analyses: sentiment distribution by aspect (100%)

Aspect co-occurrence, visualised in Figure 7, highlights how cinematic dimensions tend to be evaluated in combination rather than isolation. Each heatmap shows co-occurrences of aspects within the same sentiment category at the review level. Plot and Cast are most frequently associated, both positively (393 reviews) and negatively (225), indicating that storyline and acting are often judged jointly in user evaluations. Ambience also appears frequently alongside these two, particularly in positive contexts (245 with Plot and 252 with Cast), suggesting that the atmosphere of a film tends to reinforce prevailing impressions. By contrast, Directing co-occurs less often with other aspects, pointing to its more peripheral role in audience discourse.

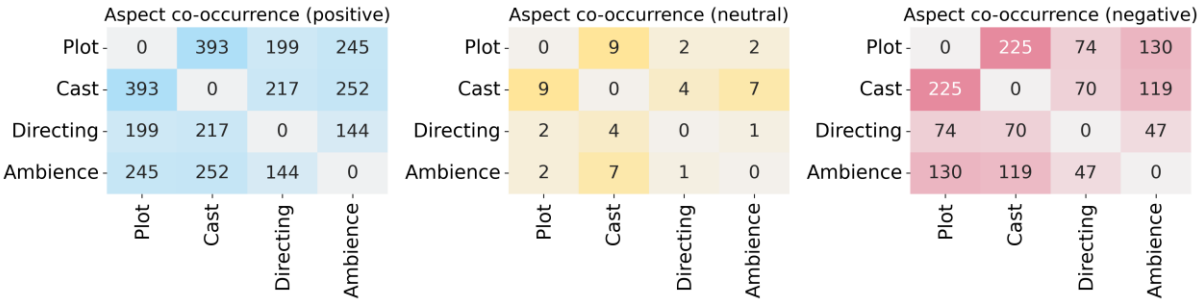


Figure 7 - Review-level analyses: aspect co-occurrence by sentiment

To complement this picture, Figure 8 presents co-occurrences where one aspect is evaluated positively and the other negatively, again at the review level. The most frequent contradictions involve Plot and Cast (123 reviews), followed by Plot–Ambience (102) and Cast–Ambience (101). These cases indicate that users often praise one dimension while criticising another, such as appreciating the storyline but disliking the acting or the atmosphere. Other contradictions are less common, for instance those involving Directing (e.g., 66 with Plot and 64 with Cast), but they nonetheless illustrate that disagreements can extend across all cinematic dimensions. Taken together, these patterns reveal that while Plot and Cast tend to dominate joint evaluations, they are also the aspects most subject to diverging opinions, highlighting the complex trade-offs that shape overall judgements.

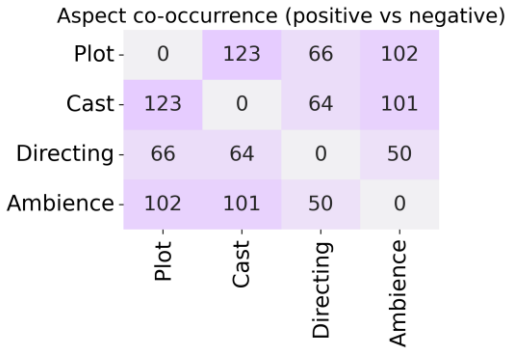


Figure 8 - Review-level analyses: aspect co-occurrence with positive–negative contradictions

In sum, the descriptive analysis confirms a dual structure: Plot and Cast dominate the discourse and attract more polarised evaluations, while Directing and Ambience, though secondary, are typically framed in positive terms when mentioned.

4.2.2. Aspect–rating relationships

As illustrated in Figure 9, the average sentiment score for each aspect increases steadily with higher ratings on the 1–10 scale. At the lowest ratings (1–3), all aspects are evaluated negatively, reflecting a broad dissatisfaction that cuts across dimensions. From rating 6 onwards, aspect sentiments rise sharply, becoming predominantly positive for high ratings (8–10). The curves show that Plot and Cast follow smoother and more consistent trajectories, while Directing and Ambience display steeper increases at intermediate values, particularly above rating 6, where they reach the highest average sentiment levels. This suggests that although less frequently mentioned, Directing and Ambience become decisive markers of particularly favourable reviews, amplifying the positivity of high ratings.

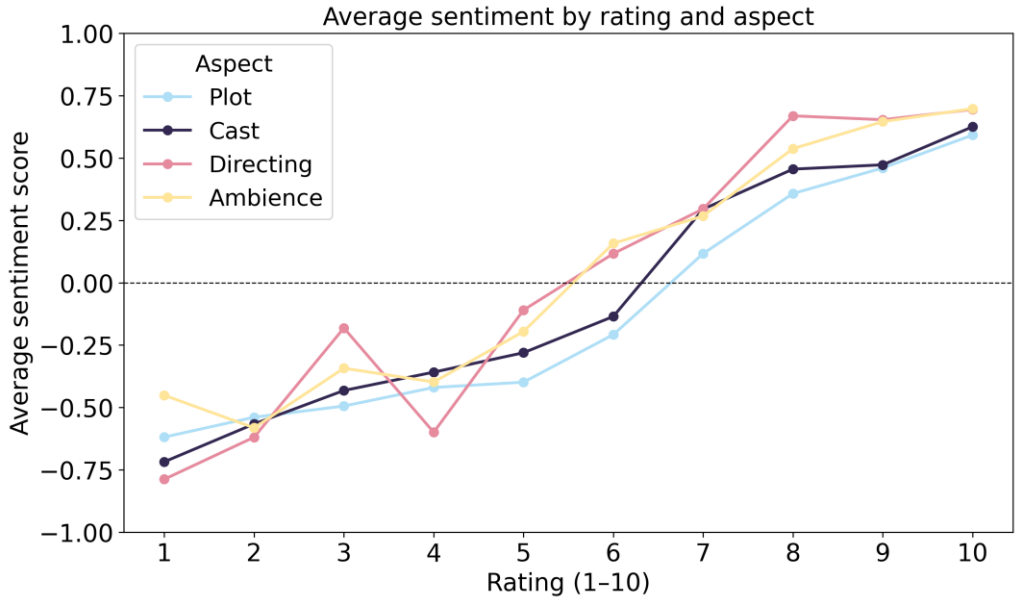


Figure 9 - Review-level analyses: average sentiment by rating and aspect

To quantify these associations, both Pearson's r and Spearman's ρ were computed. These coefficients range from –1 to +1, where positive values indicate that higher sentiment scores are associated with higher ratings. Negative values, in contrast, indicate that higher sentiment scores tend to be associated with lower ratings.

Table 11 shows that all coefficients are positive and significant at $p < .001$, confirming that more favourable evaluations of aspects are strongly aligned with higher ratings. Plot shows the highest correlation ($r = 0.63$; $\rho = 0.63$), followed closely by Cast ($r = 0.60$; $\rho = 0.59$). Directing ($r = 0.53$; $\rho = 0.51$) and Ambience ($r = 0.53$; $\rho = 0.53$) display slightly weaker but still substantial correlations. These results highlight Plot and Cast as the aspects most strongly associated with overall ratings, while Directing and Ambience emerge as complementary but meaningful contributors.

Table 11 - Correlations between aspect sentiments and review ratings (Pearson and Spearman)

Aspect	Pearson *r*	p-value	Spearman ρ	p-value
Plot	**0.631**	p < .001	**0.626**	p < .001
Cast	0.600	p < .001	0.587	p < .001
Directing	0.527	p < .001	0.506	p < .001
Ambience	0.533	p < .001	0.533	p < .001

Note. All correlations are significant at p < .001.

While correlations capture the strength of pairwise associations, regression analysis allows for estimating the unique contribution of each aspect sentiment when considered simultaneously. An Ordinary Least Squares (OLS) regression was therefore estimated. Coefficients (β) represent the expected change in ratings for a one-unit change in aspect sentiment. Positive coefficients indicate that higher (more positive) aspect sentiments are associated with higher ratings, whereas negative coefficients indicate that higher aspect sentiments are associated with lower ratings. The coefficient of determination (R^2, ranging from 0 to 1 in this context) reflects the proportion of variance in ratings accounted for by the model.

The results, presented in Table 12, show that the model explains a substantial proportion of the variance in ratings ($R^2 = 0.447$). Although this indicates that the four cinematic aspects capture an important share of rating variation, more than half remains unexplained, suggesting that additional determinants beyond the present framework are likely to influence rating formation. Plot sentiment exerts the strongest effect ($\beta = 1.62$, $p < .001$), followed by Cast ($\beta = 0.99$, $p < .001$), confirming their central role in rating formation. Directing ($\beta = 0.47$, $p < .001$) and Ambience ($\beta = 0.46$, $p < .001$) also have statistically significant effects, though with smaller magnitudes, suggesting they act as supporting dimensions that reinforce positive evaluations when present.

Table 12 - Impact of aspect sentiments on ratings (OLS regression results)

Predictor	Coefficient (β)	p-value
Plot	**1.622**	p < .001
Cast	0.993	p < .001
Directing	0.470	p < .001
Ambience	0.462	p < .001

Model fit: $R^2 = 0.447$; Adjusted $R^2 = 0.445$

Note. Dependent variable: review rating. All predictors are significant at $p < .001$.

In sum, the explanatory analysis demonstrates that aspect-level sentiments align closely with user ratings. Plot and Cast emerge as the strongest and most consistent drivers, while Directing and Ambience enhance the positivity of higher-rated reviews in a complementary way. These findings provide an essential bridge from descriptive patterns to inferential analyses, paving the way for the examination of dominant aspects in the following section.

4.3. Dominant aspect analyses

While the previous section considered all aspect mentions within reviews, this section adopts a simplification strategy by focusing on dominant aspects within each review. The aim is to test whether one or more aspects can be identified as the main driver of the overall evaluation, thereby offering a more parsimonious explanation of rating formation. This perspective provides a useful contrast to review-level analyses: if a small set of dominant aspects consistently emerges and aligns with ratings, it would suggest that film evaluations can be reduced to a limited set of primary evaluative dimensions. Conversely, if dominance proves weakly associated with ratings, this would reinforce the view that reviews are multidimensional and shaped by several aspects in combination.

Two complementary approaches were developed to operationalise dominant aspect identification. The frequency-based approach defines the dominant aspect as the one most frequently mentioned within a review, measured by the number of sentences in which the aspect appears. This method assumes that greater discursive prominence reflects evaluative salience: the more attention a reviewer devotes to an aspect, the more central it is to their overall judgement. In cases of ties, all tied aspects were recorded as dominant, each associated with its corresponding sentiment score.

The score-based approach, by contrast, identifies the dominant aspect as the one with the highest sentiment intensity. In this perspective, a single strongly polarised opinion, whether positive or negative, can prevail over multiple references to other aspects when those evaluations are expressed only in neutral or moderate terms. In the event of ties with equal sign, all tied aspects were considered dominant. For ties of equal magnitude but opposite polarity (e.g., +0.5 vs. –0.5), the overall sentiment of the review was used as a tiebreaker, with positive reviews privileging the positive aspect and negative reviews the negative one. If the review sentiment was neutral, both aspects were retained.

Applying both approaches provides two distinct but equally plausible perspectives: one grounded in discursive prominence, the other in evaluative weight. Convergence between them would reinforce the robustness of findings, confirming the centrality of certain aspects, while divergence highlights alternative pathways through which aspects may influence ratings. This dual perspective strengthens the analytical framework, enabling a more nuanced understanding of how different dimensions of a film may emerge as primary drivers of evaluation.

4.3.1. Prevalence and rating distribution of dominant aspects

Figure 10 reveals that the prevalence of dominant aspects varies considerably depending on the operationalisation adopted. Under the frequency-based definition, Plot emerges as the dominant aspect in most reviews (671 cases), followed by Cast (421). In contrast, Ambience (63) and particularly Directing (28) are seldom identified as dominant. This pattern suggests that when dominance is defined by the sheer number of mentions, reviewers devote greater discursive attention to storyline and acting, reinforcing the descriptive evidence presented in Section 4.2.1. By contrast, the score-based approach yields a more even distribution: although Plot remains important (239), both Ambience (309) and Cast (307) surpass it, and Directing rises to 230 cases. This indicates that even less frequently mentioned aspects can become dominant when associated with highly polarised sentiments.

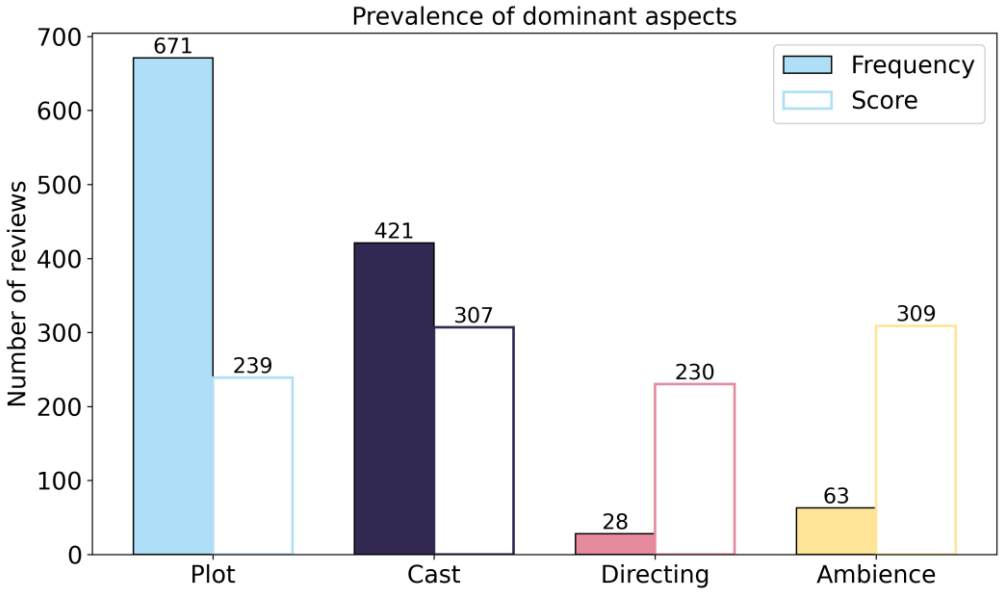


Figure 10 - Dominant aspect analyses: prevalence of dominant aspects

The rating distributions in Figure 11 provides further insights. In both approaches, reviews where Plot or Cast are dominant tend to show higher medians (around 7–8), reinforcing their central role. Ambience exhibits the greatest variability, being linked to both highly positive and more critical reviews. Directing appears more consistently associated with higher ratings, particularly under the frequency-based criterion, although this pattern may partly reflect the smaller number of cases. While the two methods yield broadly similar tendencies, the score-based approach produces occasionally produces broader distributions, indicating that single polarised statement can shift the dominant aspect towards more extreme evaluations.

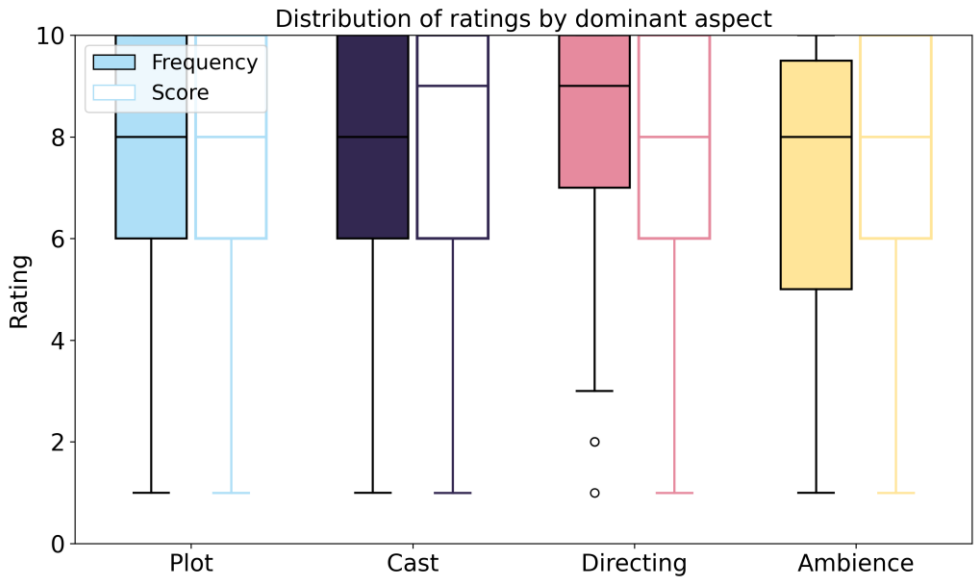


Figure 11 - Dominant aspect analyses: distribution of ratings by dominant aspect

4.3.2. Agreement between the dominant aspect approaches

As shown in Table 13, concordance between the two approaches is limited. Exact matches are observed in only 19.7% of reviews, while 26.3% overlap on at least one dominant aspect, and a clear majority (73.7%) are fully disjoint.

To further examine this divergence, a contingency table and Cohen's κ are reported. As both measures require a single dominant aspect per review, ties were resolved by applying a deterministic rule based on the hierarchy of predictive importance derived from the OLS regression in Section 4.2.2. This procedure, whereby Plot was selected first, followed by Cast, Directing and Ambience, ensured that the assignment of a single label was consistent and empirically grounded rather than arbitrary.

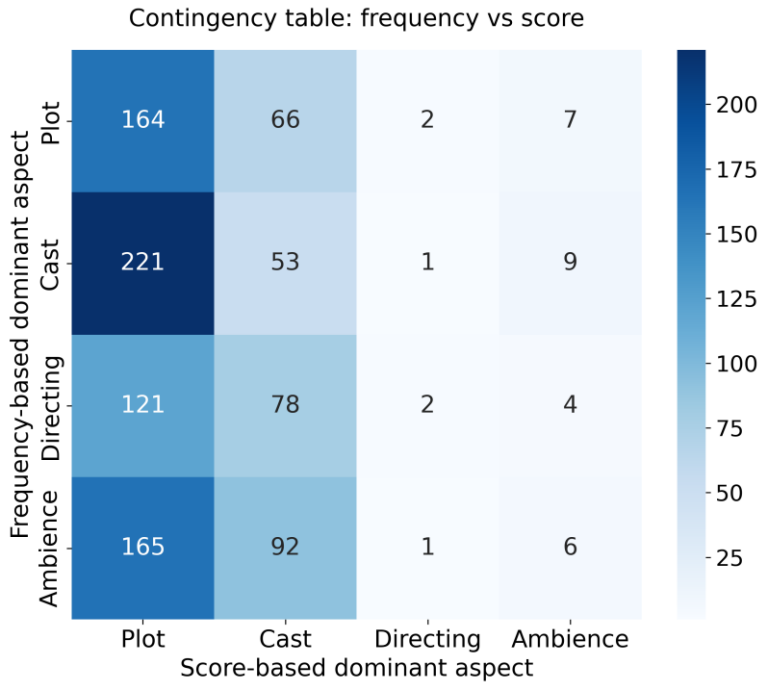
Under this adjustment, Figure 12 provides a cross-tabulation of dominant aspects identified by the two approaches. The diagonal cells capture agreement, most notably for Plot (164 reviews) and Cast (53 reviews). The off-diagonal cells, however, reveal substantial reallocations, such as 221 reviews identified as Cast-dominant under frequency but reassigned to Plot under score, along with substantial reallocations from Ambience (165 cases) and Directing (121 cases) towards Plot. These patterns confirm that the two operationalisations often attribute dominance differently, which explains the limited overall concordance reported in Table 13.

This visual evidence is consistent with the statistical results. Even under the single-label adjustment, Cohen's κ indicated virtually no agreement (κ = –0.037, overall agreement =

22.7%), underscoring that the two approaches frequently yield divergent dominant aspects. It should also be noted that κ underestimates concordance in multilabel contexts, as it cannot account for partial overlaps such as those observed in Table 13.

Table 13 - Concordance between frequency- and score-based dominant aspect identification (per review)

Measure	Value
Exact match	19.7% (195/992)
Overlap (≥1 aspect in common)	26.3% (261/992)
Disjoint	73.7% (731/992)



Agreement = 22.7%; Cohen's κ = –0.037

Figure 12 - Dominant aspect analyses: contingency table, frequency vs. score

Overall, these findings show that the identification of a dominant aspect is strongly dependent on the method used. While the frequency-based approach, which prioritises sentiment intensity, tends to concentrate dominance on Plot, the score-based approach distributes relevance more evenly, giving greater weight to Cast and Ambience. This divergence indicates that dominance is not an intrinsic property of the reviews but rather a result of methodological choice, meaning that interpretations should be made with caution and with awareness of each approach's biases.

4.3.3. Explanatory power of dominant aspects

Beyond concordance, the explanatory strength of dominant aspects was evaluated. Results in Table 14 and Table 15 indicate clear and statistically significant associations between dominant-aspect sentiment and review ratings under both dominance definitions.

When dominance is defined by frequency, Ambience shows the highest correlation (Pearson $r = 0.728$, $R^2 = 0.529$), followed by Plot ($r = 0.665$, $R^2 = 0.443$), Cast ($r = 0.594$, $R^2 = 0.353$), and Directing ($r = 0.581$, $R^2 = 0.338$). Although Ambience attains the strongest statistical association under this criterion, it appears as the dominant aspect in only 63 reviews, which increases the likelihood of sampling variability. Plot and Cast, by contrast, are dominant in a considerably larger number of reviews ($n = 671$ and $n = 421$), offering greater stability and robustness for interpretation.

When dominance is defined by sentiment intensity, the pattern remains broadly consistent. Plot ($r = 0.694$, $R^2 = 0.482$) and Cast ($r = 0.679$, $R^2 = 0.461$) continue to display the highest associations with review ratings, while Directing ($r = 0.540$, $R^2 = 0.291$) and Ambience ($r = 0.548$, $R^2 = 0.301$) show more moderate but still statistically significant relationships. These findings suggest that strongly positive or negative expressions about Plot or Cast tend to align closely with the overall evaluation of the film.

Table 14 - Correlations between dominant-aspect sentiment and review ratings (Pearson and Spearman)

Aspect	Frequency (Pearson *r)*	Frequency (Spearman ρ)	Score (Pearson *r*)	Score (Spearman ρ)
Plot	0.665, p < .001	0.658, p < .001	0.694, p < .001	0.669, p < .001
Cast	0.594, p < .001	0.584, p < .001	0.679, p < .001	0.612, p < .001
Directing	0.581, p < .01	0.431, p < .05	0.540, p < .001	0.537, p < .001
Ambience	0.728, p < .001	0.599, p < .001	0.548, p < .001	0.570, p < .001

Note. All correlations are statistically significant. Significance thresholds: p < .05; p < .01; p < .001.

Table 15 - Impact of dominant-aspect sentiment on review ratings (OLS regression results)

Predictor	Frequency β	Frequency R^2	Score β	Score R^2
Plot	**2.773**	0.443	2.336	0.482
Cast	2.390	0.353	2.260	0.461
Directing	1.861	0.338	1.611	0.291
Ambience	2.853	0.529	1.618	0.301

Note. All coefficients are statistically significant at p < .001, except Directing in the Frequency model, which is significant at p < .01.

Overall, the dominant-aspect analyses indicate that sentiment associated with individual cinematic aspects is meaningfully related to review ratings, regardless of whether dominance is determined by frequency of mention or by sentiment intensity. Plot and Cast consistently emerge as the most influential dimensions, showing strong statistical associations and relatively high explanatory power across both dominance criteria. Directing and Ambience also present significant relationships, although with lower R^2 values and, in the frequency-based approach, more limited sample sizes, which introduces additional uncertainty. These results support the premise that reviewers often structure their evaluations around a primary aspect that plays a central role in shaping their opinion.

Nevertheless, the explanatory power of the models—ranging from 0.29 to 0.53—indicates that no single aspect fully captures the complexity of audience evaluations. While dominant-aspect sentiment provides valuable information about the overall direction of a rating, it does not encompass all factors that contribute to the final assessment. This suggests that ratings are likely influenced by a broader evaluative process in which multiple aspects interact and jointly shape the reviewer's judgement.

Frequency-based and intensity-based dominance each reveal distinct yet complementary insights. Frequency highlights the aspects that consistently draw audience attention, indicating where viewers focus their commentary and which dimensions of a film matter most at scale. Intensity, in turn, identifies the aspects that trigger the strongest emotional reactions, which is particularly useful for tracking sentiment shifts or detecting enthusiasm and dissatisfaction early. Rather than relying on a single criterion, treating both indicators jointly provides a more grounded interpretation of audience behaviour. For example, a studio monitoring feedback on a newly released film could use frequency to identify that viewers discuss the plot extensively, while intensity could reveal that strong emotional responses are driven by the cast. Leveraging both signals allows decision-makers to prioritise marketing messages, allocate improvement efforts, or anticipate reception trends more effectively.

In summary, dominant-aspect sentiment is a meaningful and analytically valuable signal, but it represents only one layer of audience evaluation. A deeper understanding of rating formation likely requires considering how multiple aspects contribute jointly to the final judgement.

4.4. Combination analyses

This section examines whether overall ratings are shaped by the joint polarity of multiple aspects. Each review was represented by the set of detected aspects and their associated polarity (positive, neutral, or negative), thereby forming an evaluative configuration. Since ratings range from 1 to 10, they were grouped into three broader categories: Low (1–3), Mid (4–7), and High (8–10). The combination of four aspects across three polarity levels yields a large number of potential configurations, many of which occur only sporadically. To maintain interpretability, only configurations with at least 15 occurrences were retained, and from these the ten most frequent were selected for visualisation. Figure 13 displays the results, showing both absolute frequencies of each configuration across rating intervals and row-normalised proportions to emphasise relative tendencies.

The heatmaps reveal a consistent alignment between joint polarity and rating level. Fully positive configurations concentrate overwhelmingly in the High interval. Combinations in which Plot and Cast are both positive, with or without Ambience and Directing also positive, account for the largest masses and allocate the great majority of their occurrences to the High interval. The mirror pattern is observed for fully negative configurations: when Cast and Plot are both negative, frequently with Ambience also negative, occurrences cluster in the Low interval and are rare in High.

Mixed configurations tend to cluster in the Mid range of ratings. For instance, combinations such as a positive evaluation of Plot alongside a negative evaluation of Cast are distributed mainly across Mid scores, with some presence in the High range. This suggests that favourable perceptions of the storyline can partially offset criticism of acting, preventing ratings from falling to the lowest levels. By contrast, when negative judgements concern both Plot and Cast, positive assessments of peripheral dimensions such as Ambience or Directing rarely suffice to elevate ratings beyond the Mid range.

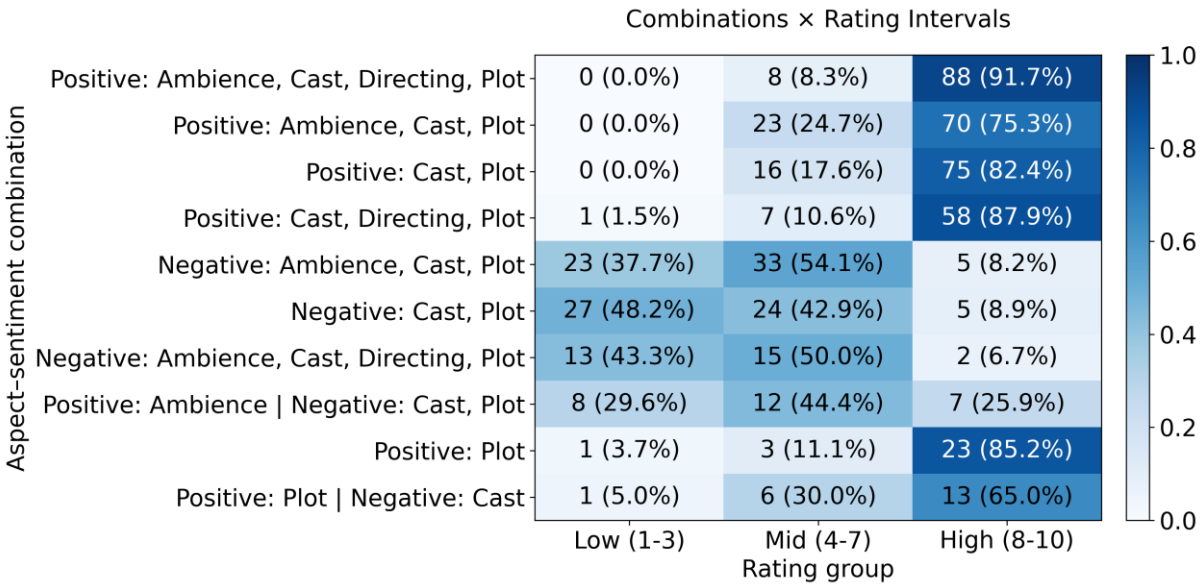


Figure 13 - Aspect combinations across rating intervals (counts and row-normalised proportions; shading reflects proportions)

Overall, the interaction results indicate that ratings arise from the configuration of aspects rather than from reliance on isolated dimensions. Plot and Cast exert the greatest impact: when they move in the same direction, the overall rating follows, clustering at the top when positive and at the bottom when negative. When their polarities diverge, ratings concentrate in the Mid range, showing that disagreement between these core aspects tends to produce intermediate evaluations. Ambience and Directing act primarily as modulators, amplifying or dampening the outcome depending on their alignment with Plot and Cast. This pattern is consistent with the results of Section 4.2, where Plot and Cast also emerged as the most frequent and explanatory aspects, while Ambience and Directing played more limited but contextually meaningful roles. Taken together, the three analytical layers provide convergent evidence that storyline and acting are the most influential drivers of ratings, while other dimensions reinforce or moderate these evaluations depending on context.

Chapter 5 – Conclusions and future research

This dissertation set out to address the research question: *Which cinematic aspects identified in movie reviews have the greatest impact on their respective ratings?* The central aim was to move beyond general sentiment classification and provide a fine-grained understanding of how specific dimensions of films contribute to audience evaluations.

To achieve this, a systematic methodology was adopted. A dataset of 1,000 IMDb reviews (Benlahbib, 2019) was selected, encompassing film titles, ratings and manually annotated sentiment polarity labels. In addition, a gold standard for aspect identification was manually created for a subset of reviews to ensure evaluation reliability. Three approaches to ABSA were implemented: two pipeline methods combining keyword-based aspect detection with either VADER or DistilBERT for sentiment classification, and an end-to-end approach using large language models (GPT-4o mini and GPT-4.1 mini). These methods were rigorously evaluated on precision, recall, F1, and accuracy, leading to the selection of the Keyword-based + DistilBERT pipeline as the most balanced and reliable framework for subsequent analyses.

5.1. Main conclusions

The findings provide clear evidence that Plot and Cast are the dominant drivers of audience ratings. These aspects were not only the most frequently mentioned across reviews but also showed the strongest correlations and regression coefficients, confirming their central role in shaping evaluations. By contrast, Directing and Ambience emerged as secondary but meaningful contributors. Although referenced less often, they consistently displayed positive effects, particularly in reviews with high ratings, where they acted as reinforcing dimensions that amplified the overall evaluative tone.

The analyses further revealed that film ratings are multidimensional outcomes shaped by configurations of aspects rather than by isolated dimensions. When Plot and Cast aligned in sentiment, ratings tended to cluster at the extremes (very high or very low), whereas divergences between them led to more intermediate evaluations. Directing and Ambience modulated these effects, often strengthening positive assessments when aligned with favourable views of Plot and Cast. This pattern demonstrates that audience evaluations are best understood as the interaction of multiple cinematic aspects rather than the dominance of a single factor.

In conclusion, the dissertation confirms that narrative quality and acting performance are the most influential determinants of ratings, while directing style and audiovisual ambience play important supporting roles. By systematically applying and comparing ABSA methods,

the study not only identified which cinematic aspects matter most to audiences but also highlighted the value of aspect-based approaches for uncovering the nuanced mechanisms that underlie film evaluations. These conclusions provide a comprehensive response to the research question and establish a solid foundation for both the academic contributions and practical implications outlined in the following sections.

5.2. Contributions to the scientific and business community

This research advances the literature on ABSA by applying a systematic, multi-phase approach to online movie reviews, a domain where sentiment analysis is both highly relevant and methodologically challenging. The study makes three main contributions. First, it offers a methodological contribution by comparing different ABSA approaches and demonstrating the importance of empirically evaluating model performance before selecting the most suitable pipeline for the domain under study. Second, it provides empirical evidence that Plot and Cast are the strongest determinants of ratings, consistently exerting the greatest influence on audience evaluations, while Directing and Ambience play secondary but meaningful roles. Third, it shows that the explanatory power of aspects depends not only on their individual frequency but also on their evaluative interactions, underscoring the importance of considering combinatory configurations in explanatory analyses.

At the business level, the results offer actionable insights for practitioners in the film industry, streaming platforms, and related sectors. The consistent finding that Plot and Cast drive ratings more strongly than other aspects provides guidance for decision-making in content production and promotion. Producers can prioritise investments in script development and casting decisions, while distributors and streaming services can refine recommendation algorithms by assigning greater weight to viewer sentiment on these dimensions. By systematically linking cinematic aspects to audience evaluations, the study provides practical knowledge that can help industry stakeholders align creative and commercial choices more closely with viewer expectations.

5.3. Research limitations

While the study provides valuable contributions, several limitations must be acknowledged. The dataset consisted of 1,000 reviews covering only ten films, all sourced from IMDb, which may limit generalisability across genres, languages, and cultural contexts. In methodological terms, aspect identification relied on predefined keyword lists and named-entity matching, which, although effective, cannot fully capture implicit or nuanced references. Sentiment

analysis was restricted to models trained primarily on English data and general-domain corpora, which may not fully account for the stylistic particularities of film reviews. Finally, the analysis focused on correlations and regressions without exploring causal mechanisms, meaning that findings should be interpreted as associations rather than definitive causal explanations.

5.4. Future research proposals

Building on these findings, several avenues for future research can be proposed. Expanding the dataset to include a larger number of reviews across multiple platforms, genres, and languages would strengthen the generalisability of results. Future work could also explore more recent end-to-end ABSA models, including state-of-the-art large language models, which may further improve the detection of implicit aspects and subtle sentiments. Another promising direction involves examining the temporal evolution of aspect-level sentiments, for instance by analysing reviews before and after a film's release or by comparing theatrical and streaming contexts. Incorporating multimodal data, such as trailers or social media content, could enrich the analysis by linking textual sentiment with visual and auditory cues. Finally, future studies may investigate causal relationships between aspects and ratings, for example through experimental structural equation modelling, thereby providing deeper insights into how cinematic dimensions shape audience evaluations.

References

Ali, N. M., Hamid, M. M., & Youssif, A. (2019). Sentiment analysis for movies reviews dataset using deep learning models. *International Journal of Data Mining & Knowledge Management Process*, *9*(2–3), 19–27. https://doi.org/10.5121/ijdkp.2019.9302

Babbar, I. (2024). Evolution of cinema. *International Journal for Multidisciplinary Research (IJFMR), 6*(2), 1-4. https://doi.org/10.36948/ijfmr.2024.v06i02.17578

Baid, P., Gupta, A., & Chaplot, N. (2017). Sentiment analysis of movie reviews using machine learning techniques. *International Journal of Computer Applications, 179*(7), 45-49. https://doi.org/10.5120/ijca2017916005

Benlahbib, A. (2019). *1000 Movie Reviews (Review + Attached rating + Sentiment polarity) for Reputation Generation [Dataset]*. Mendeley Data. https://doi.org/10.17632/38j8b6s2mx.1

Battaglia, J. (2010). *Everyone's a critic: Film criticism through history and into the digital age* (Senior Honors Thesis, The College at Brockport). The College at Brockport. http://hdl.handle.net/20.500.12648/6777

Bird, S., Klein, E., & Loper, E. (2009). *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media.

Chen, J. (2023). Examining the transformation of film in the digital age through the lens of post-classical film theory. *Frontiers in Art Research*, *5*(18), 11-17. https://doi.org/10.25236/FAR.2023.051803

Chen, T., Samaranayake, P., Cen, X., Qi, M., & Lan, Y.-C. (2022). The impact of online reviews on consumers' purchasing decisions: Evidence from an eye-tracking study. *Frontiers in Psychology, 13*, 865702. https://doi.org/10.3389/fpsyg.2022.865702

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1), 37–46. https://doi.org/10.1177/001316446002000104

Creswell, J. W., & Creswell, J. D. (2017). *Research design: Qualitative, quantitative, and mixed methods approaches* (5th ed.). SAGE Publications.

Curran, J., & Hesmondhalgh, D. (2019). *Media and society* (6th ed.). Bloomsbury Publishing.

Danyal, M. M., Khan, S. S., Khan, M., Ullah, S., Mehmood, F., & Ali, I. (2024). Proposing sentiment analysis model based on BERT and XLNet for movie reviews. *Multimedia Tools and Applications, 83*, 64315–64339. https://doi.org/10.1007/s11042-024-18156-5

Dellarocas, C., Zhang, X. (M.), & Awad, N. F. (2007). Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *Journal of Interactive Marketing*, *21*(4), 23-45. https://doi.org/10.1002/dir.20087

Devi, B. L., Bai, V. V., Ramasubbareddy, S., & Govinda, K. (2020). Sentiment analysis on movie reviews. In P. V. Krishna & M. S. Obaidat (Eds.), *Emerging research in data engineering systems and computer communications* (*Advances in Intelligent Systems and Computing*, Vol. 1054, pp. 321–328). Springer. https://doi.org/10.1007/978-981-15-0135-7_31

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers),* 4171–4186. https://doi.org/10.48550/arXiv.1810.04805

Duan, W., Gu, B., & Whinston, A. B. (2008). Do online reviews matter? — An empirical investigation of panel data. *Decision Support Systems, 45*(4), 1007-1016. https://doi.org/10.1016/j.dss.2008.04.001

Esuli, A., & Sebastiani, F. (2006). Determining term subjectivity and term orientation for opinion mining. In *Proceedings of the 11th Conference of the European Chapter of the*

Association for Computational Linguistics (pp. 193–200). Association for Computational Linguistics.

Fan, Z., Wu, Y., Zeng, Y., & Sun, L. (2019). Target-oriented opinion words extraction with target-fused neural sequence labeling. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2509–2518. https://doi.org/10.18653/v1/N19-1255

Filieri, R., McLeay, F., Tsui, B., & Lin, Z. (2018). Consumer perceptions of information helpfulness and determinants of purchase intention in online consumer reviews of services. *Information & Management, 55*(8), 956-970. https://doi.org/10.1016/j.im.2018.04.010

Gupta, S., Deodhar, S. J., Tiwari, A. A., Gupta, M., & Mariani, M. (2024). How consumers evaluate movies on online platforms? Investigating the role of consumer engagement and external engagement. *Journal of Business Research, 176*, 114613. https://doi.org/10.1016/j.jbusres.2024.114613

Horsa, O. G., & Tune, K. K. (2023). Aspect-based sentiment analysis for Afaan Oromoo movie reviews using machine learning techniques. *Applied Computational Intelligence and Soft Computing, 2023*, Artigo 3462691. https://doi.org/10.1155/2023/3462691

Hu, M., & Liu, B. (2004). Mining and summarizing customer reviews. *Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 168–177). Association for Computing Machinery. https://doi.org/10.1145/1014052.1014073

Hutto, C. J., & Gilbert, E. (2014). VADER: *A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the Eighth International Conference on Weblogs and Social Media* (ICWSM-14) (pp. 216–225). AAAI Press. https://doi.org/10.1609/icwsm.v8i1.14550

Kim, J. M., Park, K., & Mariani, M. M. (2023). Do online review readers react differently when exposed to credible versus fake online reviews? *Journal of Business Research, 154*, 113377. https://doi.org/10.1016/j.jbusres.2022.113377

Kim, S. H., Park, N., & Park, S. H. (2013). Exploring the effects of online word of mouth and expert reviews on theatrical movies' box office success. *Journal of Media Economics*, *26*(2), 98–114. https://doi.org/10.1080/08997764.2013.785551

Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *Proceedings of the International Conference on Learning Representations* (ICLR). https://doi.org/10.48550/arXiv.1609.02907

Kiss, T., & Strunk, J. (2006). *Unsupervised multilingual sentence boundary detection*. Computational Linguistics, 32(4), 485–525. https://doi.org/10.1162/coli.2006.32.4.485

Kit, H. S. B., & Joseph, M. H. (2023). Aspect-based sentiment analysis on movie reviews. *2023 15th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 237-243). IEEE. https://doi.org/10.1109/DeSE58274.2023.10099815

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics,* 33(1), 159–174. https://doi.org/10.2307/2529310

Li, J., Zhao, Y., Jin, Z., Li, G., Shen, T., Tao, Z., & Tao, C. (2022). SK2: Integrating implicit sentiment knowledge and explicit syntax knowledge for aspect-based sentiment analysis. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 1114–1123). Association for Computing Machinery. https://doi.org/10.1145/3511808.3557452

Li, X., Ma, B., & Bai, R. (2020). Do you respond sincerely? How sellers' responses to online reviews affect customer relationship and repurchase intention. *Frontiers of Business Research in China, 14*, 18. https://doi.org/10.1186/s11782-020-00086-2

Liu, P., Joty, S., & Meng, H. (2015). Fine-grained opinion mining with recurrent neural networks and word embeddings. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 1433–1443. https://doi.org/10.18653/v1/D15-1168

Liu, Y. (2006). Word of mouth for movies: Its dynamics and impact on box office revenue. *Journal of Marketing*, *70*(3), 74–89. https://doi.org/10.1509/jmkg.70.3.074

Luo, H., Li, T., Liu, B., Wang, B., & Unger, H. (2019). Improving aspect term extraction with bidirectional dependency tree representation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing, 27*(7), 1201–1212. https://doi.org/10.1109/TASLP.2019.2913094

Maas, A. L., Daly, R. E., Pham, P. T., Huang, D., Ng, A. Y., & Potts, C. (2011). Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (pp. 142–150). Association for Computational Linguistics.

Maitama, J. Z., Idris, N., Abdi, A., Shuib, L., & Fauzi, R. (2020). A systematic review on implicit and explicit aspect extraction in sentiment analysis. *IEEE Access, 8*, 194166–194191. https://doi.org/10.1109/ACCESS.2020.3031217

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

Mäntylä, M. V., Graziotin, D., & Kuutila, M. (2018). The evolution of sentiment analysis - A review of research topics, venues, and top cited papers. *Computer Science Review, 27*, 16-32. https://doi.org/10.1016/j.cosrev.2017.10.002

Mensah, S., Sun, K., & Aletras, N. (2021). An empirical study on leveraging position embeddings for target-oriented opinion words extraction. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 9174–9179. https://doi.org/10.18653/v1/2021.emnlp-main.722

Mir, J., & Mahmood, A. (2020). Movie aspects identification model (MAIM) for aspect-based sentiment analysis. *Information Technology and Control, 49*(4), 564-582. https://doi.org/10.5755/j01.itc.49.4.25350

Nazir, A., Rao, Y., Wu, L., & Sun, L. (2022). Issues and challenges of aspect-based sentiment analysis: A comprehensive survey. *IEEE Transactions on Affective Computing, 13*(2), 845–863. https://doi.org/10.1109/TAFFC.2020.2970399

Oh, C., Roumani, Y., Nwankpa, J. K., & Hu, H.-F. (2017). Beyond likes and tweets: Consumer engagement behavior and movie box office in social media. *Information & Management, 54*(1), 25-37. https://doi.org/10.1016/j.im.2016.03.004

Onalaja, S., Romero, E., & Yun, B. (2021). Aspect-based sentiment analysis of movie reviews. *SMU Data Science Review, 5*(3), Article 10. https://scholar.smu.edu/datasciencereview/vol5/iss3/10

Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 79–86). Association for Computational Linguistics. https://doi.org/10.3115/1118693.1118704

Patil, D. R., & Rane, N. L. (2023). Customer experience and satisfaction: Importance of customer reviews and customer value on buying preference. *International Research Journal of Modernization in Engineering Technology and Science, 5*(3), 3437-3447. https://doi.org/10.56726/IRJMETS36460

Pocchiari, M., Proserpio, D., & Dover, Y. (2024). Online reviews: A literature review and roadmap for future research. *International Journal of Research in Marketing.* https://doi.org/10.1016/j.ijresmar.2024.08.009

Pontiki, M., Galanis, D., Pavlopoulos, J., Papageorgiou, H., Androutsopoulos, I., & Manandhar, S. (2014). SemEval-2014 Task 4: Aspect Based Sentiment Analysis. *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, 27–35. Association for Computational Linguistics. https://doi.org/10.3115/v1/S14-2004

Powers, D. M. W. (2011). Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1), 37–63

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv. https://doi.org/10.48550/arXiv.1910.01108

Satyarthi, S., & Sharma, S. (2023). Identification of effective deep learning approaches for classifying sentiments at aspect level in different domain. 2023 *IEEE International Conference on Paradigm Shift in Information Technologies with Innovative Applications in Global Scenario (ICPSITIAGS)* (pp. 496-508). IEEE. https://doi.org/10.1109/ICPSITIAGS59213.2023.10527695

Sharma, N. A., Ali, A. B. M. S., & Kabir, M. A. (2024). A review of sentiment analysis: Tasks, applications, and deep learning techniques. *International Journal of Data Science and Analysis*. https://doi.org/10.1007/s41060-024-00594-x

Stilwell, S. (2024). *Explainable prompt learning for movie review sentiment analysis* (Master's thesis, University of Ottawa). University of Ottawa Research Repository. http://hdl.handle.net/10393/46044

Thakur, R. (2018). Customer engagement and online reviews. *Journal of Retailing and Consumer Services, 41*, 48-59. https://doi.org/10.1016/j.jretconser.2017.11.002

Thet, T. T., Na, J.-C., & Khoo, C. S. G. (2010). Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of Information Science, 36*(6), 823-848. https://doi.org/10.1177/0165551510388123

Tsao, W.-C. (2014). Which type of online review is more persuasive? The influence of consumer reviews and critic ratings on moviegoers. *Electronic Commerce Research*, *14*, 559–583. https://doi.org/10.1007/s10660-014-9160-5

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics* (pp. 417–424). Association for Computational Linguistics. https://doi.org/10.3115/1073083.1073153

Vaniuha, L., Kyreia, M., Lemishka, N., Spolska, O., & Patron, I. (2024). History of the evolution of cinema in the context of considering the stages of development of science and technology: The first steps to the birth of cinema. *History of Science and Technology / History of Technology*, *14*(2), 513-538. https://doi.org/10.32703/2415-7422-2024-14-2-513-538

Wang, Q., Wen, Z., Zhao, Q., Yang, M., & Xu, R. (2021). Progressive self-training with discriminator for aspect term extraction. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 257–268. https://doi.org/10.18653/v1/2021.emnlp-main.23

Wang, Y., Shen, G., & Hu, L. (2020). Importance evaluation of movie aspects: Aspect-based sentiment analysis. *2020 5th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)* (pp. 2444-2448). IEEE. https://doi.org/10.1109/ICMCCE51767.2020.00527

Wankhade, M., Rao, A. C. S., & Kulkarni, C. (2022). A survey on sentiment analysis methods, applications, and challenges. *Artificial Intelligence Review, 55*, 5731–5780. https://doi.org/10.1007/s10462-022-10144-1

Wu, Y., Ngai, E. W. T., Wu, P., & Wu, C. (2020). Fake online reviews: Literature review, synthesis, and directions for future research. *Decision Support Systems, 132*, 113280. https://doi.org/10.1016/j.dss.2020.113280

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2018). Double embeddings and CNN-based sequence labeling for aspect extraction. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers),* 592–598. https://doi.org/10.18653/v1/P18-2094

Xu, H., Liu, B., Shu, L., & Yu, P. S. (2019). BERT post-training for review reading comprehension and aspect-based sentiment analysis. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2324–2335. https://doi.org/10.18653/v1/N19-1242

Yin, Y., Wei, F., Dong, L., Xu, K., Zhang, M., & Zhou, M. (2016). Unsupervised word and dependency path embeddings for aspect term extraction. *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence* (IJCAI-16), 2979–2985. https://doi.org/10.48550/arXiv.1605.07843

Zhang, W., Li, X., Deng, Y., Bing, L., & Lam, W. (2023). A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering, 35*(11), 11019–11038. https://doi.org/10.1109/TKDE.2022.3230975

Zhou, J., Huang, J. X., Chen, Q., Hu, Q. V., Wang, T., & He, L. (2019). Deep learning for aspect-level sentiment classification: Survey, vision, and challenges. IEEE Access, 7, 78454–78483. https://doi.org/10.1109/ACCESS.2019.2920075

Appendices

Appendix A

To enhance the accuracy of aspect detection, the names of actors and directors associated with the films in the dataset were retrieved using the OMDb API. These names were incorporated into the keyword lists of the Cast and Directing aspects, ensuring that explicit references to individuals were correctly identified during the aspect detection process.

A.1 Actor Names (Cast)

Al Pacino, Anthony Perkins, Carrie-Anne Moss, Chiwetel Ejiofor, Cyril Cusack, Ed Harris, Elizabeth Berridge, Ewan McGregor, F. Murray Abraham, James Caan, Jason Statham, Jennifer Connelly, Jet Li, John Cusack, Keanu Reeves, Laurence Fishburne, Liam Neeson, Marlon Brando, Natalie Portman, Ralph Fiennes, Richard Jordan, Russell Crowe, Sam Worthington, Sigourney Weaver, Sylvester Stallone, Thandiwe Newton, Tom Hulce, Zoë Saldaña.

A.2 Director Names (Directing)

Francis Ford Coppola, George Lucas, Glenn Jordan, James Cameron, Lana Wachowski, Lilly Wachowski, Louis Leterrier, Milos Forman, Roland Emmerich, Ron Howard, Sylvester Stallone.

Appendix B

B.1 Full instruction (exact text used)

```
In this movie review, identify the presence of the aspects below:

Aspect definitions (keywords are hints, not hard rules):

• Plot: refers to the story of the movie (e.g. keywords: plot, story,
storyline, ending, storytelling, drama, writing, twist, script, writing, end,
movie).

• Cast: refers to the actors and their performance as mentioned in the review
(e.g. keywords: acting, role, character, act, actress, actor, villain,
protagonist, antagonist, performance, performed, play, played, playing,
casting, cast, crew, artist, portray). Also include the following names as Cast
mentions if detected:

al pacino, anthony perkins, carrie-anne moss, chiwetel ejiofor, cyril cusack,
ed harris, elizabeth berridge, ewan mcgregor, f. murray abraham, james caan,
jason statham, jennifer connelly, jet li, john cusack, keanu reeves, laurence
fishburne, liam neeson, marlon brando, natalie portman, ralph fiennes, richard
jordan, russell crowe, sam worthington, sigourney weaver, sylvester stallone,
thandiwe newton, tom hulce, zoe saldaña.

• Directing: refers to the flow of the movie and the way it was directed (e.g.
keywords: direct, directing, direction, filming, cinematography, filmmaker,
cinematic, director). Also include the following names as Directing mentions if
detected:

francis ford coppola, george lucas, glenn jordan, james cameron, lana
wachowski, lilly wachowski, louis leterrier, milos forman, roland emmerich, ron
howard, sylvester stallone.

• Ambience: refers only to immersion elements such as visual effects and sound
effects (keywords: visual, effect, animation, CGI, graphics, scenery, stunt,
design, audio, sound, music, track).

– Mentions of scene(s), destruction, or action should not be classified as
Ambience unless they are explicitly linked to visuals or sound (e.g.,
“spectacular scenes with CGI,” “soundtrack,” “special effects,” “amazing
audio”). Otherwise, classify them as Plot.

– Generic adjectives such as “epic,” “intense,” “fun,” or “boring” should not
activate Ambience unless directly tied to visuals or sound.

• General: use this when no specific aspect is clearly mentioned in the review.

Key rule (very important)

• Aspect detection must NOT depend on sentiment. If an aspect is present but
there is no clear evaluative cue, leave its *_sentiment cell blank.

Counting rules

• Mark at most one mention per aspect per sentence (even if multiple keywords
or multiple actors/directors of the same aspect appear).
```

```
• Aspect_count = the number of distinct sentences in the review that contain explicit lexical evidence for that aspect. If an aspect is mentioned in 3 different sentences, the count must be 3.
• If Aspect_count = 0, leave the corresponding Aspect_sentiment cell blank.
• General_count = 1 only if Plot_count = Cast_count = Directing_count = Ambience_count = 0; otherwise General_count = 0 and General_sentiment blank.
Sentiment
• For each aspect with count > 0, assign one overall Aspect_sentiment in [-1, 1].
• Provide Overall_sentiment in [-1, 1] for the entire review.
• Overall_label rules:
– If Overall_sentiment ≥ 0.05 → positive
– If Overall_sentiment ≤ -0.05 → negative
– If -0.05 < Overall_sentiment < 0.05 → neutral
Format your answer as a table with:
• One column for each aspect's count (e.g. Plot_count)
• One column for each aspect's sentiment (e.g. Plot_sentiment)
• One column for the overall sentiment score (e.g. Overall_sentiment)
• One column for the overall sentiment label (e.g. Overall_label)
Respond only with the table.
```

B.2 Model and decoding parameters

- Models: GPT-4o mini and GPT-4.1 mini
- Input: raw review text concatenated to the instruction above (no additional preprocessing)
- Temperature: *0.0*, to ensure deterministic outputs
- Number of responses: 1
- Post-processing: The model returned the results in a tabular text format, which was subsequently parsed to ensure that each field was correctly mapped into the predefined schema {Aspect_counts, Aspect_sentiments, Overall_sentiment, Overall_label}.