Sentiment Beyond Text: A Multimodal Framework for Movie Review Classification

A thesis
Submitted in partial fulfillment of the requirements for the Degree of
Bachelor of Science in Computer Science and Engineering

Submitted by

Sanjana Akhter	**20210104150**
Md. Mostafizur Rahman	**20210104151**
Mahajabin Haque Upoma	**20210104154**

Supervised by

Md. Reasad Zaman Chowdhury

Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Dhaka, Bangladesh

December 2025

CANDIDATES' DECLARATION

We, hereby, declare that the thesis presented in this report is the outcome of the investigation performed by us under the supervision of Md. Reasad Zaman Chowdhury, Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh. The work was spread over two final year courses, CSE4100: Project and Thesis I and CSE4250: Project and Thesis II, in accordance with the course curriculum of the Department for the Bachelor of Science in Computer Science and Engineering program.

It is also declared that neither this thesis nor any part thereof has been submitted anywhere else for the award of any degree, diploma or other qualifications.

Sanjana Akhter
20210104150

Md. Mostafizur Rahman
20210104151

Mahajabin Haque Upoma
20210104154

CERTIFICATION

This thesis titled, **"Sentiment Beyond Text: A Multimodal Framework for Movie Review Classification"**, submitted by the group as mentioned below has been accepted as satisfactory in partial fulfillment of the requirements for the degree B.Sc. in Computer Science and Engineering in December 2025.

Group Members:

Sanjana Akhter	**20210104150**
Md. Mostafizur Rahman	**20210104151**
Mahajabin Haque Upoma	**20210104154**

Md. Reasad Zaman Chowdhury
Lecturer & Supervisor
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

Prof. Dr. Md. Shamim Akhter
Professor & Head
Department of Computer Science and Engineering
Ahsanullah University of Science and Technology

ACKNOWLEDGEMENT

Firstly, we would like to express our sincere gratitude to our supervisor, Md. Reasad Zaman Chowdhury, for his invaluable guidance, patience, and continuous support throughout this research. His insightful feedback and thoughtful critiques shaped not only the technical directions of our work but also our overall approach to research. He consistently encouraged us to think critically and challenge our assumptions, which strengthened the quality of this study.

We are also grateful to Ms. Nawshin Tabassum Tanny for her early guidance and meaningful input during the formative stages of this work. Her constructive feedback played an important role in shaping how we approached the problem.

We extend our heartfelt appreciation to the members of the defense committee for their time, insightful comments, and encouraging remarks, all of which contributed to improving the final outcome of this research.

Finally, we would like to express our deep appreciation to everyone in the Department of Computer Science and Engineering at Ahsanullah University of Science and Technology—our teachers, classmates, and staff. Their openness to questions, willingness to support, and the academic environment they fostered gave us the space, resources, and motivation needed to complete this project successfully.

Dhaka
December 2025

Sanjana Akhter

Md. Mostafizur Rahman

Mahajabin Haque Upoma

ABSTRACT

User-written movie reviews have become a powerful way to understand public opinion, which makes sentiment analysis essential. This project focuses on both binary and 3 class sentiment analysis and aims to classify movie reviews accurately using a multimodal approach. For this task, a dataset from Rotten Tomatoes was utilized. Thorough preprocessing techniques were applied to handle not only raw reviews but also structured metadata, along with addressing class imbalance and selecting features according to their ranks, using various ranking methods based on their relationship with the target label. To evaluate model performance, traditional machine learning algorithms such as Logistic Regression, XGBoost, and LightGBM were implemented alongside advanced transformer-based models like DistilBERT, RoBERTa, and DeBERTa. The study extends beyond binary classification by introducing a third "Neutral" sentiment class, derived via a data-driven approach based on the Point-Biserial Correlation between normalized user scores and binary sentiment labels, which identifies reviews with weak alignment and reclassifies them as Neutral. All models were trained and evaluated across three distinct environments: using all selected features, a subset of top-ranked features, and only review text. Each of the models was evaluated using accuracy, F1-score, and other key performance metrics, and comparisons were made among the models for both binary and three-class problems, with performance further optimized through hyperparameter tuning. This research aims to improve understanding of audience sentiment and contribute to natural language processing applications through a combination of careful data preparation and a wide range of modeling techniques..

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Introduction/Overview

Online platforms such as Rotten Tomatoes have evolved into the primary arena for measuring audience reaction and public sentiment regarding films. These user-generated reviews serve a dual purpose as they reflect individual viewer opinions and significantly influence collective consumption behavior, directly impacting box office performance and long-term movie popularity [1]. As the volume of digital content expands exponentially, the ability to automatically interpret and categorize these reviews has become a critical necessity for modern media platforms and recommendation engines.

However extracting accurate sentiment from movie reviews presents unique challenges. Unstructured text is often mixed with sarcasm, slang, and implicit meaning, which traditional analysis methods struggle to interpret correctly. Conversely, relying solely on numerical ratings fails to capture the emotional nuance and context provided by the written word. To address these limitations, this research implements a multimodal approach, integrating unstructured textual data with structured metadata such as scores, genres, directors and language to enhance predictive accuracy.

While earlier studies predominantly focused on binary classification (Positive vs. Negative) using standard Machine Learning algorithms, this project advances the field by incorporating state-of-the-art Transformer architectures, specifically DistilBERT, RoBERTa, and DeBERTa. Furthermore, the study moves beyond the limitations of binary sentiment by introducing a rigorous, data-driven methodology for 3 class classification by a statistically grounded "Neutral" class identified through the Point-Biserial Correlation between normalized user ratings and binary sentiment labels. By benchmarking these advanced deep learning models against traditional baselines like Logistic Regression, XGBoost, and LightGBM across varying feature environments, this research aims to provide a comprehensive frame-

work for understanding the interplay between metadata and language in sentiment analysis.

1.2 Problem Statement

The massive volume of user comments on movie platforms has made it impossible to analyze public opinion manually. While automation is necessary it faces several significant hurdles. First, reviews are often written in casual language filled with sarcasm, slang, and mixed emotions which are difficult for standard computer models to interpret correctly. Second, real-world datasets are rarely balanced as they frequently contain far more positive reviews than negative ones which can cause models to become biased during training.

A critical limitation in many existing systems is the attempt to force every review into a simple "positive" or "negative" box. This binary approach fails to handle "neutral" or average opinions effectively, often misclassifying mixed reviews because they do not fit neatly into either category. Finally, movie data is complex because it comes in different formats like unstructured text, numerical scores, and categorical tags like genre which must be carefully harmonized to work together.

This project addresses these challenges by developing a flexible system capable of both binary and 3 class classification. By using advanced Transformer models to better understand language and by combining text with metadata, this research aims to overcome the limitations of basic classifiers and build a more accurate, balanced framework for predicting sentiment.

1.3 Motivation

Movie reviews act as the collective voice of the audience, possessing the power to shape public perception and determine a film's financial success. While platforms like Rotten Tomatoes are rich with this feedback, much of the true insight remains locked away in unstructured text which is difficult for standard computers to decode.

This project is driven by the challenge of transforming that raw noise into usable data. Despite progress in sentiment analysis, traditional models still struggle with the complexities of human language, such as sarcasm, slang, or conflicting emotions. Furthermore, treating every opinion as simply "positive" or "negative" ignores the reality of the "average" moviegoer. This binary approach often forces complex, mixed feelings into the wrong category, reducing the reliability of the results.

The core motivation, therefore, is to move beyond these limitations. By adopting advanced

Transformer architectures that can truly understand context, and by introducing a "neutral" category to capture the middle ground, this research seeks to build a framework capable of handling the messy reality of user reviews. The goal is to create a system that utilizes both text and metadata to provide a deeper, more accurate understanding of audience sentiment.

1.4 Objectives

The primary aim of this research is to construct a holistic sentiment analysis framework that evaluates the efficacy of modern Transformer architectures against traditional machine learning baselines. This study seeks to determine how well these models perform across both binary and 3 class environments when presented with varying combinations of unstructured text and structured metadata. To achieve this, the project focuses on the following specific goals:

- **Multimodal Data Curation:** To execute a rigorous preprocessing pipeline that harmonizes heterogeneous data types, effectively cleaning and integrating unstructured textual inputs (reviews and titles) with structured numeric and categorical metadata (such as genres, runtime, and scores).
- **Data-Driven Target Formulation:** To resolve the ambiguity inherent in raw, binary-labeled datasets by leveraging the Point-Biserial Correlation between normalized user ratings and binary sentiment labels. This correlation is used to construct a weighted composite index that identifies reviews with weak alignment between score and label—thereby enabling the data-driven derivation of a "Neutral" sentiment class and supporting more nuanced 3 class analysis.
- **Feature Importance Analysis:** To apply ensemble-based ranking techniques to quantify the predictive value of different metadata attributes, systematically identifying which features contribute most significantly to model accuracy while filtering out noise.
- **Model Implementation:** To deploy and fine-tune a diverse suite of classifiers, benchmarking established machine learning algorithms (Logistic Regression, XGBoost, LightGBM) against state-of-the-art Transformer models (DistilBERT, RoBERTa, DeBERTa).
- **Stratified Benchmarking:** To conduct a comparative performance analysis across three distinct feature environments,the full feature set, a selected subset of top-ranked predictors, and only review text data to assess each model's reliance on metadata versus linguistic understanding.

- **Optimization and Evaluation:** To enhance predictive performance through systematic hyperparameter tuning and to evaluate generalization capabilities using standard metrics (Accuracy, Precision, Recall, and F1-Score) and thereby identifying the optimal trade-offs between model complexity and accuracy.

1.5 Organization of the Book

This thesis is organized into eight chapters that systematically present the research framework and findings. Chapter 1 introduces the study by outlining the problem statement, motivation, and specific research objectives. Chapter 2 establishes the theoretical foundation, reviewing statistical baselines like Logistic Regression and XGBoost alongside Transformer architectures, and identifies gaps in existing literature. Chapter 3 details the methodology, describing the dataset curation, the proposed dual-stream comparative framework, and the implementation of the Transformer-based late fusion architecture. Chapter 4 presents the experimental results, offering a comprehensive analysis of binary and 3 class classification performance, followed by an assessment of impact and sustainability and comparison with existing works. Chapter 5 covers the project management aspects, including cost analysis and project scheduling. Chapter 6 discusses the ethical and professional responsibilities inherent in the research. Chapter 7 maps the study to complex engineering problems and activities, demonstrating its technical depth. Finally, Chapter 8 concludes the thesis and suggests potential directions for future work.

Chapter 2

Background Study and Literature Review

2.1 Introduction/Overview

This section provides an overview of prior studies and scholarly approaches to sentiment analysis in the domain of movie reviews. Existing research predominantly focuses on applying machine learning and deep learning algorithms to classify reviews as positive or negative, often using benchmark datasets such as IMDb and Rotten Tomatoes. Key themes emerging from the literature include the comparison of traditional models (e.g., Naive Bayes, SVM) with more advanced neural architectures (e.g., LSTM, BERT, XLNet), as well as the role of preprocessing techniques in improving model accuracy. Despite these advancements, a critical trend observed is the reliance on unidimensional input, primarily textual data—while rich metadata such as genre, runtime, and director remain underutilized. This chapter explores these foundational patterns, highlighting the strengths of prior methods and the gaps that warrant further investigation.

2.2 Background Study

This section outlines the theoretical foundations of the algorithms employed in this research. The study utilizes a multimodal approach, benchmarking efficient traditional machine learning classifiers against state-of-the-art Transformer architectures. Additionally, it details the statistical method used for the 3 class target generation.

2.2.1 Logistic Regression

Logistic Regression serves as the foundational baseline for this study, providing a robust statistical framework for classification. While traditionally used for binary tasks, this research utilizes it for both the binary (Positive/Negative) and 3 class (Negative/Neutral/Positive) problems. Unlike linear regression, which predicts continuous values, Logistic Regression maps the weighted sum of input features to a probability score between 0 and 1, making it ideal for sentiment probability estimation.

Logistic Regression Formula

Depending on the classification task, the model employs one of two activation functions:

1. Binary Classification (Sigmoid):
For the standard two-class problem, the model uses the sigmoid function to estimate the probability $\hat{y}$ of the positive class:

$$\hat{y} = \sigma(w^T x + b) \tag{2.1}$$

Where:

- $\hat{y}$: Predicted probability (0 to 1) of the target class.
- $\sigma(z)$: The Sigmoid function, defined as $\frac{1}{1+e^{-z}}$.
- w: The weight vector associated with the input features.
- b: The bias intercept.

2. 3 class Classification (Softmax):
For the three-class problem, the model is extended to Multinomial Logistic Regression. Instead of a single value, it computes a probability distribution over the K classes (where $K = 3$) using the Softmax function:

$$P(y = k \mid x) = \frac{e^{w_k^T x + b_k}}{\sum_{j=1}^{K} e^{w_j^T x + b_j}} \tag{2.2}$$

This ensures that the probabilities for Negative, Neutral, and Positive sum to exactly 1.

Logistic Regression Workflow

The training process minimizes the error between predicted probabilities and actual labels by optimizing specific loss functions—Binary Cross-Entropy for the two-class problem and Categorical Cross-Entropy for the 3 class variant. The model utilizes the `lbfgs` solver to iteratively adjust weights and biases, while explicitly applying class weights to penalize misclassifications in the minority "Neutral" class. Final predictions are generated by converting probabilities into discrete labels, using a standard threshold for binary tasks and the `argmax` function to select the highest probability class for the 3 class problem.

2.2.2 XGBoost (Extreme Gradient Boosting)

XGBoost is a highly efficient implementation of the gradient boosting framework, designed for speed and performance. Unlike traditional bagging methods (like Random Forest) that build trees independently, XGBoost builds an ensemble of decision trees sequentially, where each new tree aims to correct the residual errors of the previous ones. It is widely adopted for its ability to handle sparse data and its built-in regularization (L1 and L2), which is critical for preventing overfitting on high-dimensional metadata.

XGBoost Formula

The final prediction is the weighted sum of scores from K individual regression trees.

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i), \quad f_k \in \mathscr{F} \tag{2.3}$$

Where f_k represents the k-th tree in the ensemble $\mathscr{F}$.

Objective Functions: To convert this raw sum into a probability for classification, XGBoost minimizes a specific objective function based on the task:

- **Binary:** Uses the `binary:logistic` objective, which applies the Sigmoid function to the output logits to minimize the log-loss.
- **3 class:** Uses the `multi:softprob` objective, applying the Softmax function to generate a probability distribution across the Negative, Neutral, and Positive classes.

XGBoost Workflow

The training process follows an additive strategy. At each step, a new tree is added that maximally reduces the loss function (gradient descent in function space). The algorithm utilizes a pre-sorted algorithm and histogram-based splitting to find optimal split points efficiently. For the 3 class problem, the model builds distinct trees for each class in a "One-vs-Rest" fashion internally or uses a vectorized approach to output probabilities via Softmax, ensuring that the model learns the distinct boundaries for the "Neutral" class.

2.2.3 LightGBM

Light Gradient Boosting Machine (LightGBM) is a distributed gradient boosting framework that uses tree-based learning algorithms. It is specifically designed to handle massive datasets with lower memory usage and faster training speeds than XGBoost. Its primary innovation is the use of Gradient-based One-Side Sampling (GOSS) and Exclusive Feature Bundling (EFB), which allow it to ignore data instances with small gradients (small errors) and focus on the difficult cases.

LightGBM Tree Growth Formula

Unlike most decision tree algorithms that grow trees level-wise (depth-first), LightGBM grows trees leaf-wise (best-first). It chooses the leaf with the maximum delta loss to grow:

$$\text{Gain} = \frac{1}{2}\left[\frac{(\sum g_L)^2}{\sum h_L + \lambda} + \frac{(\sum g_R)^2}{\sum h_R + \lambda} - \frac{(\sum g)^2}{\sum h + \lambda}\right] - \gamma \tag{2.4}$$

Where g and h are the first and second-order gradients of the loss function, and subscripts L and R denote the left and right children.

LightGBM Workflow

The learning workflow focuses on optimizing the loss function efficiently.

- **Leaf-Wise Growth:** The model iteratively splits the leaf node that results in the largest decline in loss, leading to deeper, more complex trees that generally achieve higher accuracy.
- **Objective Optimization:** Similar to XGBoost, it minimizes Binary Log-loss (using Sigmoid) for binary tasks and 3 class Log-loss (using Softmax) for the three-class problem.

- **Categorical Handling:** LightGBM natively handles categorical features (like genre encodings) by finding optimal split points without requiring one-hot encoding, preserving information structure.

2.2.4 Transformer Architectures

This study moves beyond statistical baselines by implementing three variants of the BERT (Bidirectional Encoder Representations from Transformers) architecture. These models utilize the mechanism of Self-Attention to weight the significance of different words in a sentence relative to one another, regardless of their positional distance.

DistilBERT

DistilBERT is a distilled version of BERT that retains 97% of the performance while being 40% smaller and 60% faster. It is trained using knowledge distillation, where a "student" model mimics the soft target probabilities of a larger "teacher" model. It serves as an efficient baseline for extracting semantic context without the heavy computational cost of full-scale models.

RoBERTa

RoBERTa (Robustly Optimized BERT Approach) modifies the pre-training procedure of BERT to improve stability and performance. It removes the Next Sentence Prediction (NSP) task, trains on larger mini-batches, and utilizes dynamic masking—where the masked tokens change during training epochs. This allows the model to learn more robust representations of informal language and sarcasm found in movie reviews.

DeBERTa

DeBERTa (Decoding-enhanced BERT with Disentangled Attention) introduces a disentangled attention mechanism. Unlike BERT, which sums content and position embeddings, DeBERTa encodes them separately using two distinct vectors.

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{2.5}$$

By computing attention weights based on content-to-content and content-to-position matrices separately, DeBERTa captures nuanced syntactic structures and long-range dependencies more effectively than its predecessors.

Transformer Classification Workflow

All three Transformer models follow a unified classification workflow in this study:

- **Encoding:** The input text is tokenized and passed through the Transformer layers. The hidden state of the special `[CLS]` token (h_{CLS}) is extracted as the aggregate semantic representation of the entire review.
- **Late Fusion:** For the multimodal approach, this h_{CLS} vector is concatenated with the normalized metadata vector.
- **Classification Head:** The combined vector is passed through a fully connected (Dense) layer. Finally, an activation function is applied to generate probabilities: Sigmoid for the binary task and Softmax for the 3 class task.

2.2.5 Point-Biserial Correlation

The Point-Biserial Correlation coefficient (r_{pb}) is a special case of the Pearson Product-Moment Correlation, designed specifically to measure the strength and direction of the association between a continuous variable and a binary variable. In the context of sentiment analysis, this statistical metric is essential for quantifying the relationship between a user's numerical rating (continuous) and the platform's assigned sentiment label (binary).

Mathematical Definition

Mathematically, the Point-Biserial Correlation is calculated as follows:

$$r_{pb} = \frac{M_1 - M_0}{s_n} \sqrt{\frac{n_1 n_0}{n^2}} \tag{2.6}$$

Where:

- M_1: The mean of the continuous variable for the group where the binary variable is 1 (Positive).
- M_0: The mean of the continuous variable for the group where the binary variable is 0 (Negative).
- s_n: The standard deviation of the continuous variable across the entire dataset.
- n_1 and n_0: The number of data points in group 1 and group 0, respectively.
- n: The total sample size ($n = n_1 + n_0$).

Application in Feature Weighting

The value of r_{pb} ranges from -1 to +1. A high positive value indicates that higher numerical scores are strongly associated with the positive binary class. In hybrid classification frameworks, this coefficient provides a data-driven metric for "trust." By treating r_{pb} as a weight, models can dynamically assign importance to the numerical score versus the binary label, allowing for the construction of weighted composite indices that reveal underlying data patterns not visible through simple binary categorization.

2.3 Literature Review

The evolution of sentiment analysis in the movie domain has progressed through distinct phases from traditional machine learning and deep learning architectures to modern transformer-based approaches. This section synthesizes key studies from each era, with a particular focus on dataset characteristics and classification scope.

Traditional Machine Learning Approaches

Early research heavily relied on statistical classifiers applied to varying dataset sizes. Steinke et al. [2] utilized the Stanford dataset to evaluate decision trees and random forests, concluding that Support Vector Machines (SVM) served as a more effective baseline (F1=0.86). However, dataset scale significantly influences model choice. In a smaller study using only 2,000 Kaggle reviews, Dey et al. [3] found that Random Forest achieved the highest accuracy (93.95%), outperforming Naïve Bayes. Conversely, when scaling up to larger datasets, such as the 50,000 IMDb reviews analyzed by Lahase and Deshmukh [4], SVM proved superior, achieving 81.7% accuracy compared to K-Nearest Neighbors (KNN).

Reinforcing the continued relevance of traditional methods, Soni et al. [5] conducted a comparative study on a Rotten Tomatoes dataset of 10,215 sentences. They benchmarked Multinomial Naïve Bayes (MNB) and Support Vector Machines (SVM) against a deep learning model, finding that the optimized SVM achieved the highest accuracy (77.5%) followed by MNB with bigrams (76.1%), both significantly outperforming the deep learning baseline (58.6%).

The dataset balance is also critical. Dahir and Alkindy [6] utilized a balanced dataset of 10,000 IMDb reviews, reporting that Logistic Regression paired with TF-IDF achieved 89.20% accuracy. This finding was corroborated by Tomal [7] on a larger balanced set of 50,000 reviews, where Logistic Regression yielded the highest F1-score (0.898) against decision tree variants. In the context of recommendation systems, Wu [8] applied KNN to

a massive Netflix dataset containing 17,800 movies, achieving 89.5% accuracy in rating prediction.

Crucially, some studies attempted to move beyond binary classification. Mishra et al. [9] developed a model to classify reviews into Positive, Negative, and Neutral categories. While the Random Forest classifier performed well on polarized sentiments, it struggled significantly with the Neutral class, yielding F1-scores around only 0.5. This highlights a persistent gap in traditional ML's ability to handle ambiguous sentiment.

Deep Learning and Neural Networks

The shift toward neural networks allowed for better handling of large-scale text data. Azahree and Alias [10] utilized the IMDb 50k dataset to demonstrate that LSTM deep learning models (87.12% accuracy) significantly outperform lexicon-based approaches like VADER (70.00%). Similarly, Lu et al. [11] applied Word2Vec embeddings with an LSTM classifier on the same 50,000-review IMDb dataset, achieving 92.65% accuracy and demonstrating strong generalization capabilities.

Advanced ensemble architectures have also been explored. Das et al. [12] investigated a weighted average ensemble of CNN-GRU, BiLSTM-CNN, and 2D-CNN models, achieving a test accuracy of 90.6%. Meanwhile, Subedi et al. [13] benchmarked SVM against Naïve Bayes on 50,000 Kaggle reviews, confirming that while deep learning is powerful, optimized SVMs can still achieve competitive accuracies ($\approx 88\%$).

Transformer-Based Architectures

The current state-of-the-art is dominated by Transformer architectures, which have been tested on diverse and extensive corpora. Khan et al. [14] benchmarked models on 50,000 Rotten Tomatoes reviews, finding that XLNet (87.68%) and BERT (82.24%) consistently outperformed traditional CNN-LSTM networks.

Hybrid approaches have yielded further gains. Nkhata et al. [15] tested a fine-tuned BERT+BiLSTM model across multiple large-scale datasets—including IMDb, SST-2, and Amazon reviews achieving accuracies as high as 98.76%. Hua [16] validated a similar hybrid architecture on a combined corpus of 50,000 English and Chinese reviews, achieving 0.91 accuracy.

Most relevant to the current study, Gao [17] explored a three-class sentiment problem using a Kaggle dataset of approximately 35,000 reviews. While their hybrid BERT+BiLSTM model achieved only ($\approx 65\%$) accuracy on the multiclass task, it highlighted the "accuracy-efficiency trade-off" and the significant challenge of correctly classifying non-binary senti-

ments compared to pure binary tasks.

Table 2.1: Literature Review Comparison Table

Paper	Dataset Used	Dataset Size	Classes	Models Used	Best Performance
Steinke et al. [2]	Stanford Movie Review / IMDb	50k	Binary	Decision Tree, Random Forest, SVM	SVM (86% F1 Score)
Dey et al. [3]	Kaggle Movie Reviews	2,000	3 Classes	Naïve Bayes, Decision Tree, Random Forest	Random Forest (93.95% Accuracy)
Lahase & Sachin [4]	IMDb	50,000	Binary	SVM, KNN	SVM (81.7% Accuracy)
Soni et al. [5]	Rotten Tomatoes (AMT-annotated)	10,215 sentences	Binary	MNB, SVM, Deep Learning	Optimized SVM (77.5% Accuracy)
Dahir and Alkindy [6]	IMDb 10k	10,000	Binary	Logistic Regression, SVM, Random Forest, TF-IDF	Logistic Regression (89.2% Accuracy)
Tomal [7]	IMDb (Kaggle)	50,000	Binary	Logistic Regression, Random Forest, Decision Tree	Logistic Regression (0.898 F1 Score)
Wu [8]	Netflix Open Dataset	17.8k movies	Rating Prediction	KNN + Collaborative Filtering	89.5% Accuracy
Mishra et al. [9]	IMDb	Not specified	3 Classes	Random Forest, Logistic Regression, SVM	Random Forest (80% Accuracy)
Azahree & Alias [10]	IMDb 50k	50,000	Binary	VADER, LSTM	LSTM (87.12% Accuracy)
Lu et al. [11]	IMDb 50k	50,000	Binary	Word2Vec (CBOW) + LSTM	92.65% Accuracy
Das et al. [12]	IMDb	Not specified	Binary	CNN–GRU, BiLSTM–CNN, 2D–CNN	Weighted Ensemble (90.6% Accuracy)
Subedi et al. [13]	IMDb (Kaggle)	50,000	Binary	SVM, Naïve Bayes	SVM (≈88% Accuracy)
Sarwar Shah Khan et al. [14]	Rotten Tomatoes	50,000	Binary	XLNet, BERT, CNN–LSTM	XLNet (87.68% Accuracy)
Nkhata et al. [15]	IMDb, MR, SST-2, Amazon	50k+	Binary	BERT + BiLSTM	BERT+BiLSTM (98.76% Accuracy)
Mingze Hua [16]	IMDb + Chinese Corpus	50,000	Binary	BERT + CNN	BERT+CNN (91.5% Accuracy)
Rongjun Gao [17]	Kaggle IMDb	32,745 train / 2,681 test	3 Classes	Encoder, BERT, BERT+BiLSTM	BERT+BiLSTM (65% Accuracy)

2.3.1 Gap Analysis

Based on the review of existing literature, this study identifies gaps that current research has yet to fully address:

Limited Dataset Scope: A significant portion of prior research relies heavily on standard, relatively small datasets, such as the IMDb 50k review set. While useful for initial testing, these datasets often lack the scale and variety needed to train robust models for real-world applications. There is a need for studies that utilize larger, more diverse datasets such as extensive Rotten Tomatoes archives to better represent the complexity of general audience sentiment.

Reliance on Text-Only Analysis: Most existing sentiment analysis frameworks treat movie reviews as a text-only problem. They focus almost exclusively on processing the written review, ignoring the rich structured metadata that usually accompanies it, such as directors, genres, and language. This "unimodal" approach overlooks valuable context that could significantly improve prediction accuracy, especially when the text itself is sarcastic or brief.

Oversimplified Binary Classification: The dominant trend in the field is to classify reviews into just two categories: positive or negative. This binary approach oversimplifies human opinion, forcing "average" or mixed reviews into a polarized category where they don't truly fit. This leads to a loss of nuance and higher misclassification rates for films that are neither excellent nor terrible.

Ambiguity in Defining "Neutral": Even when studies attempt to include a third "Neutral" class, the definition is often arbitrary. Researchers typically use manual rules (e.g., "3 stars is neutral") rather than a data-driven method. There is a clear lack of rigorous, statistical methodologies like correlation-based weighting to mathematically define where "Neutral" begins and ends, which creates noisy and unreliable labels in multiclass datasets.

2.4 Summary

This chapter provided a comprehensive review of the evolution of sentiment analysis, tracing the progression from traditional machine learning algorithms to state-of-the-art Transformer architectures. It also established the theoretical background of the specific models employed, detailing the mathematical principles behind both statistical classifiers and advanced neural networks. Furthermore, the review identified critical limitations in current methodologies, specifically the pervasive reliance on text-only analysis and the lack of robust, data-driven definitions for neutral sentiment. These gaps directly motivate the proposed research framework, which integrates multimodal features with a mathematically derived target variable to build a more accurate and nuanced classification system.

Chapter 3

Methodology

3.1 Introduction/Overview

This chapter outlines the methodological framework implemented to advance the classification of movie reviews through a multimodal approach. The methodology is specifically designed to resolve the limitations identified in the literature, namely the reliance on unimodal data and the ambiguity inherent in binary sentiment labels. The primary objective is to evaluate how integrating unstructured textual data with structured metadata can enhance predictive performance across both traditional machine learning and advanced Transformer architectures.

To achieve this, the research adopts a data-centric design that prioritizes the rigorous harmonization of heterogeneous features, ensuring that both text and metadata are optimized for model consumption. A key innovation of this framework is the shift from arbitrary manual labeling to a statistically grounded target formulation, which uses correlation metrics to scientifically define neutral sentiment. Furthermore, the study employs a comprehensive comparative strategy, benchmarking various model architectures across differing feature environments to isolate the specific impact of metadata on classification accuracy.

3.2 Dataset

This study utilizes the "Massive Rotten Tomatoes Movies & Reviews" dataset hosted on Kaggle [18], which serves as a rich repository of audience and critic reactions. The original data comprises two separate files: a metadata collection covering approximately 143,000 films and a reviews archive containing over 1.4 million entries. This dataset was chosen for its multimodal nature, offering a crucial combination of unstructured textual feedback and

structured metadata—including features such as `originalScore`, `audienceScore`, `tomatoMeter`, and `genre`-which allows for a comparative analysis between text-only and hybrid modeling approaches.

3.2.1 Data Selection

To construct a high-quality dataset suitable for robust model training, a rigorous selection and filtering process was applied to the raw data.

- **Integration:** The movie metadata and review datasets were merged using the movie id as a primary key to associate each review with its film-specific attributes.
- **Linguistic Filtering:** To ensure compatibility with the pre-trained Transformer models (which are optimized for English language patterns), a language detection algorithm was applied to the `reviewText`. Only reviews identified as English were retained, while entries in other languages (such as Spanish, French, or Portuguese) were filtered out.
- **Popularity Capping:** A significant issue in review datasets is popularity bias, where a handful of blockbuster movies contribute a disproportionate number of reviews. To mitigate this, a strict cap was enforced, limiting the number of reviews per movie title (16 reviews max). This ensured that the model learned from a diverse range of films rather than over-fitting to a few popular titles.
- **Balancing and Target Definition:** To prevent model bias, the final dataset was down-sampled to approximately 195,000 records, ensuring a balanced distribution between the two primary classes (Positive and Negative) for the binary classification task. This binary ground truth was established using the platform's provided `scoreSentiment` label. For the 3 class experiment, a more granular target variable was generated. The `originalScore` was utilized as the foundational metric to mathematically derive a third "Neutral" class (interfaced between Negative and Positive), effectively transforming the balanced binary dataset into a three-class structure for more nuanced sentiment analysis.

	A	B	C	D	E	F	G	H	I	J
1	title	reviewText	scoreSentiment	originalScore	audienceScore	tomatoMeter	runtimeMinutes	genre	originalLanguag	director
2	Nekrotronic	Ghostbusters meets The Matrix in	NEGATIVE	2/5	66	39	99	Comedy, Horror,	English (Australi	Kiah Roache-Tu
3	Goodnight Mom	Goodnight Mommy cannot avoid c	NEGATIVE	5/10	35	40	91	Mystery & thrille	English	Matt Sobel
4	California Split	...a distressingly erratic endeavor t	POSITIVE	3/4	83	87	108	Comedy, Drama	English	Robert Altman
5	Midsommar	High-art horror that won't suit all ta	POSITIVE	3/4	63	83	145	Horror, Mystery	English	Ari Aster
6	The Leather Boy	something rather different and muc	POSITIVE	3.5/5	79	77	108	Drama	English	Sidney J. Furie
7	Eric Clapton: Lif	This Clapton-approved portrait feel	NEGATIVE	2/5	74	69	135	Documentary, M	English	Lili Fini Zanuck
8	Listen Up Philip	A spiky movie about the pointlessn	POSITIVE	3/4	54	82	109	Comedy, Drama	English	Alex Ross Perry
9	Loveling	"Great Brazilian melodrama. One c	POSITIVE	3.5/4	77	95	95	Drama, Comedy	Portuguese (Bra	Gustavo Pizzi
10	Boys and Girls	It's a paint-by-numbers romantic cc	NEGATIVE	1/5	50	11	94	Romance, Come	English	Robert Iscove
11	Fly Away	A defiantly unsentimental look at th	POSITIVE	4/5	62	85	85	Drama	English	Janet Grillo
12	Save Me	As a movie, it doesn't amount to m	NEGATIVE	C	66	68	96	Drama	English	Robert Cary
13	Mary Poppins	The entire set was constructed ind	POSITIVE	71/100	86	96	140	Kids & family, M	English	Robert Stevensc
14	The Army of Crir	A sombre and gripping account of	POSITIVE	4/5	68	88	139	History, Drama, '	French (Canada	Robert Guédigui
15	Femme Fatale	...De Palma takes great care in not	POSITIVE	2.5/4	49	49	110	Mystery & thrille	English	Brian De Palma
16	A Million Ways t	Well, I'm not sure the world neede	NEGATIVE	2.5/5	41	33	115	Comedy, Weste	English	Seth MacFarlane
17	Troy	As summer-movie entertainment, '	POSITIVE	3/4	73	53	162	Adventure, War,	English	Wolfgang Peters
18	The Unseen	... an interesting take on the invisib	POSITIVE	4.5/5	51	100	104	Action, Horror, S	English	Geoff Redknap
19	Tootsie	It turns out to be a touching love st	POSITIVE	4/4	81	90	111	Comedy, Lgbtq+	English	Sydney Pollack
20	And While We W	Imagine a boring, gender-swappec	NEGATIVE	C-	37	43	83	Drama, Romanc	English	Kat Coiro
21	Heaven Can Wa	Frivolous and amusing in spots.	POSITIVE	B-	68	87	101	Comedy, Fantas	English	Warren Beatty,B
22	The Sower	The period piece is an alluring first	POSITIVE	B	57	100	99	History, Drama,	French (France)	Marine Francen
23	The Haunting	It's just a conglomeration of cheap	NEGATIVE	1.5/5	28	17	114	Horror, Mystery	English	Jan de Bont
24	Boy Erased	A tad too safe and sanitary. But it's	POSITIVE	3.5/5	74	80	115	Biography, Dran	English	Joel Edgerton
25	The Cocksure La	I've watched cartoons show more l	NEGATIVE	1.5/5	82	40	96	Comedy, Drama	English	Murray Foster

Figure 3.1: Sample of the Final Dataset

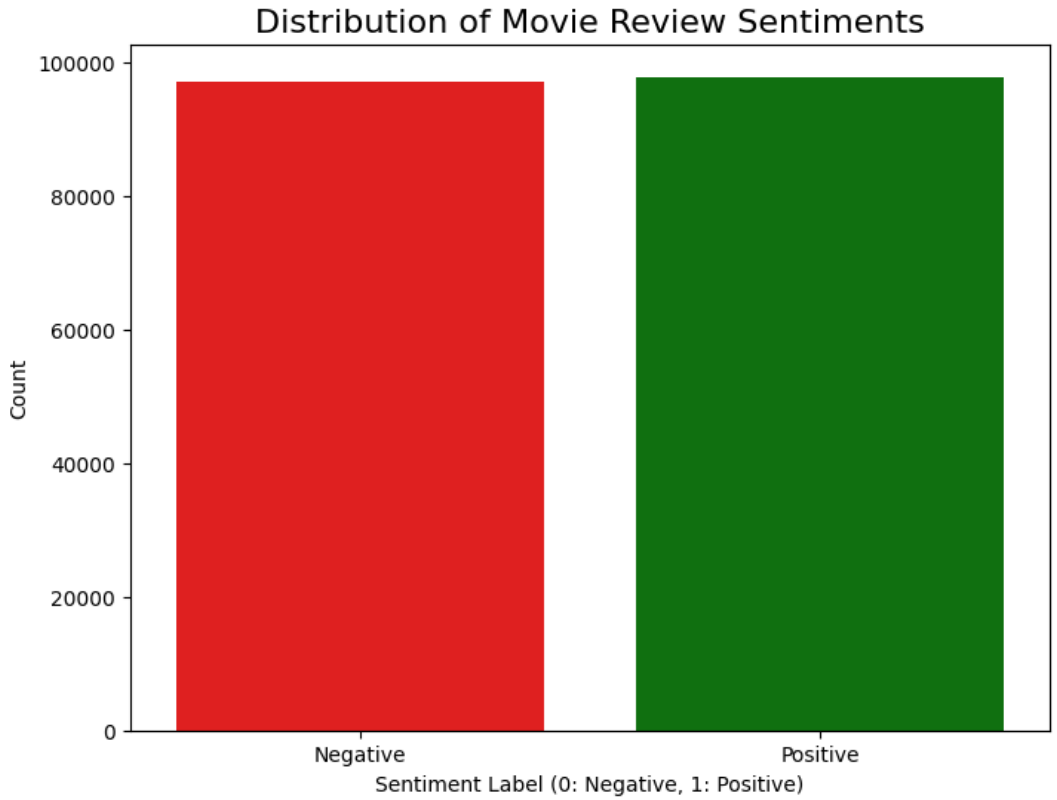

Figure 3.2: Distribution of Movie Review Sentiments

3.2.2 Data Preprocessing

Raw web-scraped data contains significant noise, formatting inconsistencies, and heterogeneous data types that hinder effective model training. To resolve this, a comprehensive preprocessing pipeline was engineered to transform the raw input into a structured, machine-readable format. This process involved linguistic cleaning, numerical standardization, and a novel statistical approach for target variable formulation.

A.Textual and Categorical Cleaning Strategy

To optimize performance across disparate model architectures, a bifurcated preprocessing strategy was implemented. While traditional machine learning algorithms benefit from dimensionality reduction (stemming/lemmatization), deep learning models require the preservation of semantic context.

- **Universal Text Normalization:** All unstructured text fields (`reviewText` and `title`) underwent an initial cleaning stage to remove noise without altering sentence structure. This included the removal of URLs, HTML tags, and non-alphanumeric artifacts.
- **Contextual Preservation (For Transformer):** For the deep learning models, the text was kept in this semi-raw state (post-normalization). Stopwords and punctuation were retained to preserve the grammatical fidelity required for self-attention mechanisms and sequence modeling.
- **Lexical Standardization (For Traditional Models):** To support the Bag-of-Words and TF-IDF approaches used in the baseline models, a rigorous NLP pipeline was applied. This involved tokenization, the removal of English stopwords (using NLTK), lemmatization (via WordNet), and stemming (using the Snowball Stemmer) to reduce high-dimensional vocabulary to base roots.
- **Genre and Language Standardization:** Categorical features required unification. The genre field, originally inconsistent comma-separated strings, was parsed, trimmed, and sorted alphabetically. Simultaneously, the originalLanguage field was normalized by bucketing regional dialects (e.g., "English (UK)" $\rightarrow$ "English") to reduce sparsity before encoding.

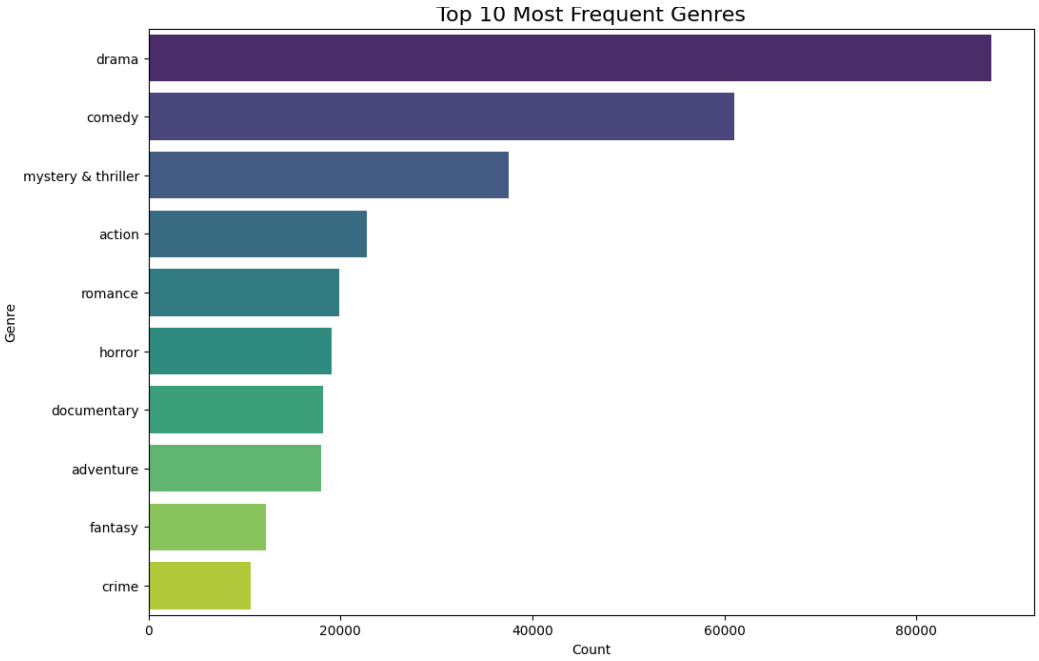

Figure 3.3: Top 10 most frequent genre

- **Director Encoding:** The `director` column exhibited high cardinality. To manage dimensionality without losing information, a frequency-based grouping strategy was applied. The top K most frequent directors were retained as distinct categories, while rare occurrences were grouped under a generic "Other" label.

B.Numerical Feature Normalization (Original Score Parsing)

A significant challenge in this dataset was the heterogeneity of the `originalScore` field, which contained user ratings in various inconsistent formats. A custom parsing algorithm was developed to categorize and normalize these scores into a uniform scale.

- **Score Type Classification:** The raw `originalScore` entries were first classified into distinct types:
  - **Fractions:** Strings representing a ratio (e.g., “3/5”, “10/20”).
  - **Letter Grades:** Academic-style grades (e.g., “A”, “B+”, “C-”).
  - **Numeric:** Raw integers or floats (e.g., “80”, “4.5”).
  - **Other:** Unusable or noisy data which was subsequently filtered out.
- **Fraction Conversion:** Entries identified as fractions were parsed by splitting the numerator and denominator. The value was then mathematically normalized to a stan-

dard $0-10$ scale using the formula:

$$Score_{norm} = \left(\frac{\text{Numerator}}{\text{Denominator}}\right) \times 10 \tag{3.1}$$

- **Grade Conversion:** Letter grades were mapped to numerical equivalents based on a standard scale converted to a 10-point system (e.g., 'A+' $\rightarrow$ 10, 'A' $\rightarrow$ 9, ..., 'F' $\rightarrow$ 1). This allowed categorical grade data to be integrated seamlessly with numerical scores.

- **Rescaling:** Other numerical features, such as `audienceScore` and `tomatoMeter`, which were originally on a $0-100$ scale, were divided by 10. This ensured all score-based features operated on a consistent $0-10$ magnitude, preventing scale-induced bias during model training.

  To visualize the relationships between these standardized variables, a correlation matrix of the numerical features was generated, as shown below

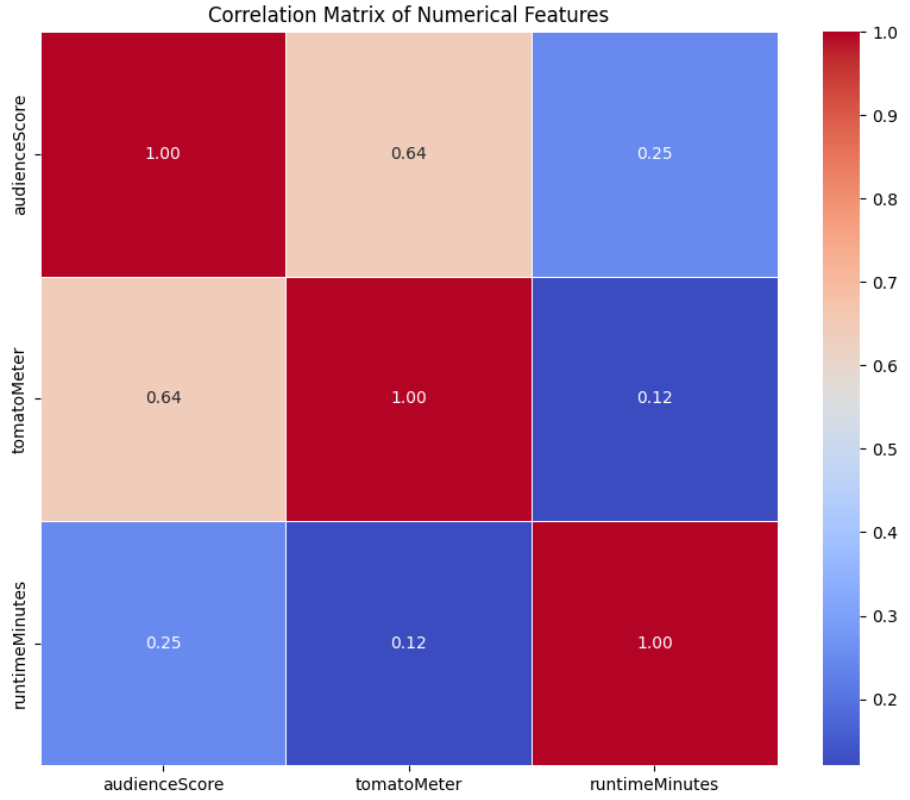

Figure 3.4: Correlation matrix of numeric features

C. 3 class Target Generation (Correlation-Weighted Index)

3 class Target Generation (Composite Formula)

While the dataset provided binary sentiment labels (Positive/Negative), these fail to capture the nuance of "Neutral" or mixed reviews. To address this, a data-driven methodology was developed to mathematically derive a third "Neutral" class rather than relying on arbitrary manual thresholds. This method generates a Composite Index for every review.

1. Weight Calculation via Point-Biserial Correlation : To determine how much "trust" to place in the numerical score versus the binary label, the Point-Biserial Correlation (r) was calculated between the normalized score (S_{norm}) and the binary sentiment label (L).

- The analysis yielded a correlation of $r \approx 0.79$.
- Consequently, weights were assigned dynamically: the Score Weight (w_S) was set to r, and the Label Weight (w_L) was set to $1-r$.

This ensures that the continuous score, which contains more granular information, drives the classification more than the binary label.

$$w_S = r \quad \text{and} \quad w_L = 1 - r \tag{3.2}$$

2. Composite Index Formulation : For every review i, a Weighted Composite Index ($I_{weighted}$) was calculated as the linear combination of the weighted score and label:

$$I_{weighted} = (w_S \cdot S_{norm}) + (w_L \cdot L) \tag{3.3}$$

3. Dynamic Thresholding (Linear Interpolation) : Rather than creating static cutoffs (e.g., 5.0 to 6.0), "Edge Case" anchors were defined to identify where the model should be uncertain (e.g., a Negative label combined with a mid-range Score of 5). We defined four reference anchors ($A_{1..4}$):

- A_1: Negative Boundary (Score 4, Label 0)
- A_2: Neutral Start (Score 5, Label 0)
- A_3: Neutral End (Score 6, Label 1)

- A_4: Positive Boundary (Score 7, Label 1)

Linear interpolation was used between these anchors to calculate precise Lower (C_{low}) and Upper (C_{high}) cutoffs:

$$C_{low} = \frac{A_1 + A_2}{2}, \quad C_{high} = \frac{A_3 + A_4}{2} \tag{3.4}$$

4. Final Classification Function : Using the calculated index and dynamic thresholds, the final 3 class target was assigned using the following piecewise function:

$$Class(x) = \begin{cases} 0 \text{ (Negative)} & \text{if } I_{weighted} < C_{low} \\ 1 \text{ (Neutral)} & \text{if } C_{low} \leq I_{weighted} \leq C_{high} \\ 2 \text{ (Positive)} & \text{if } I_{weighted} > C_{high} \end{cases} \tag{3.5}$$

This ensures that there are no negative reviews in the 3-class positive bin and also no positive reviews in negative bin. The neutral bin consists of both positive and negative reviews. Figure shows the distribution of target labels in three class.

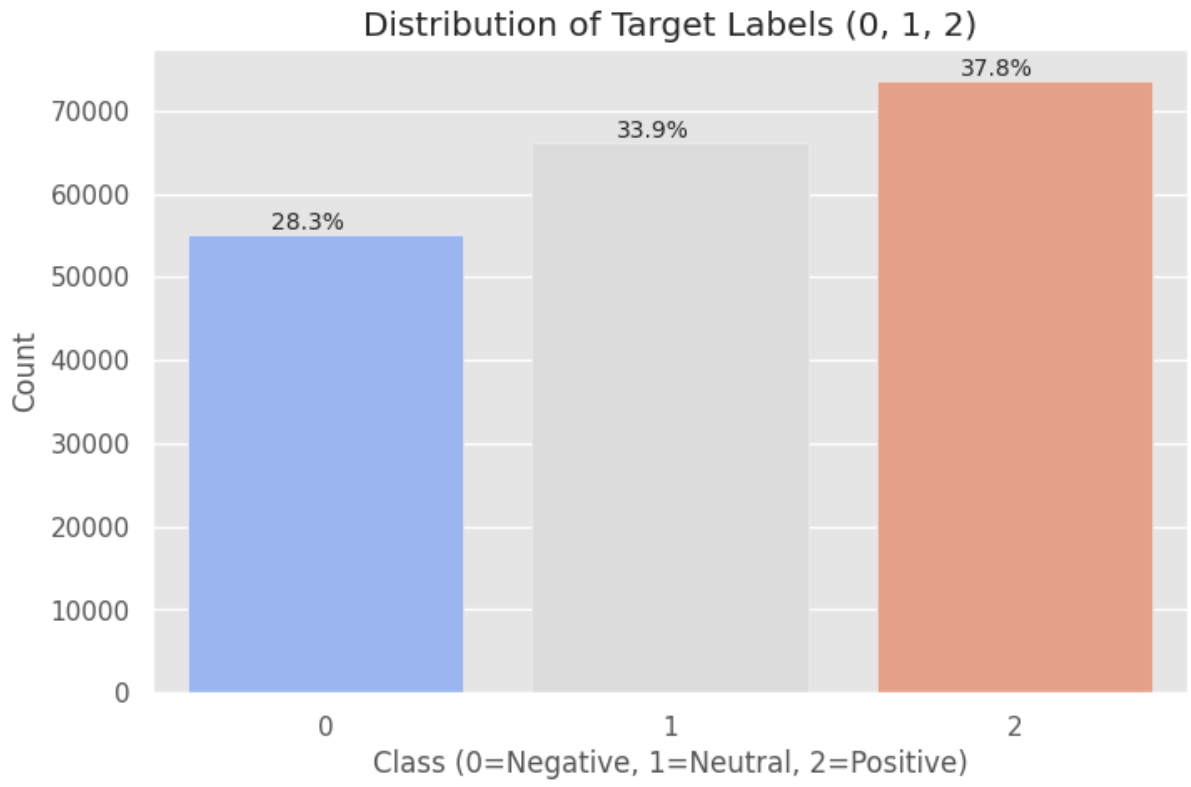

Figure 3.5: Distribution of Target labels in 3 class

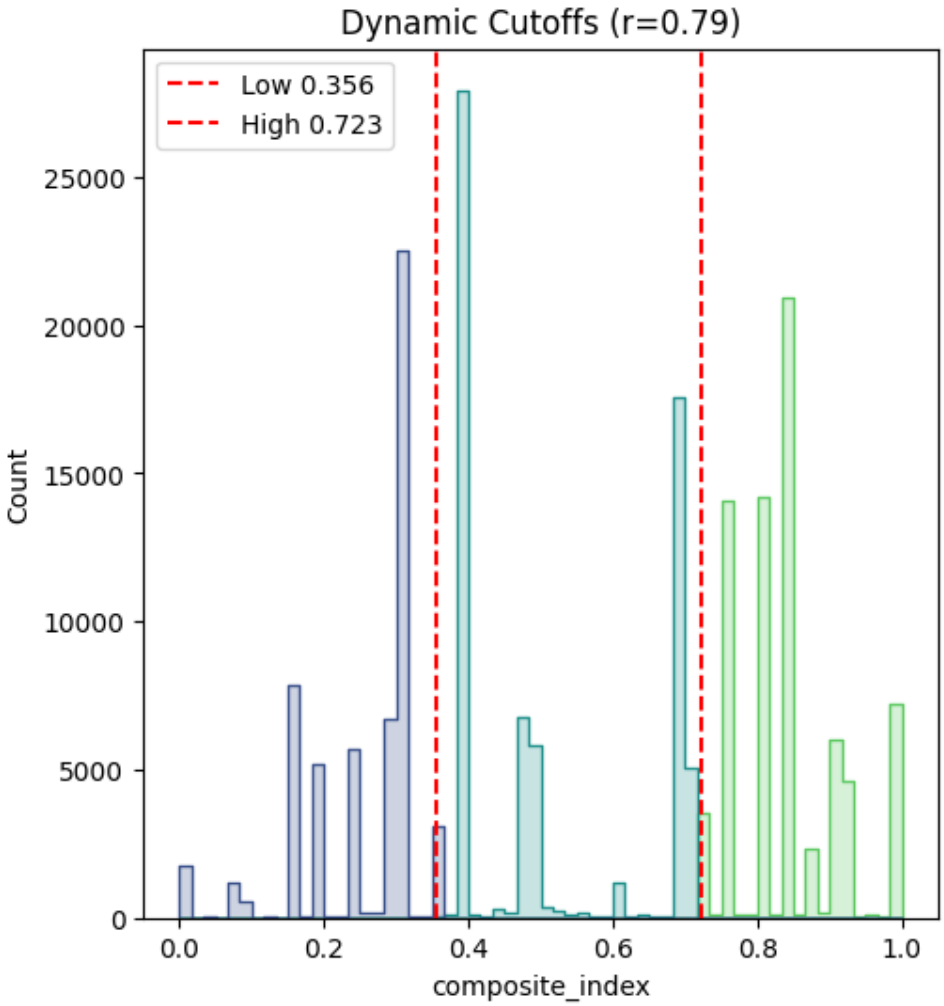

Figure 3.6: Distribution of the Weighted Composite Index with Calculated Decision Boundaries (C_{low} and C_{high}).

D. Feature Selection and Importance Analysis

To identify the most informative features for sentiment clas sification, a robust selection process was employed. This step improved performance, reduced training time, and mitigated overfitting. Feature importance was computed using Random Forest, XGBoost, and LightGBM, and the combined ranking is presented in Figure 3.7.

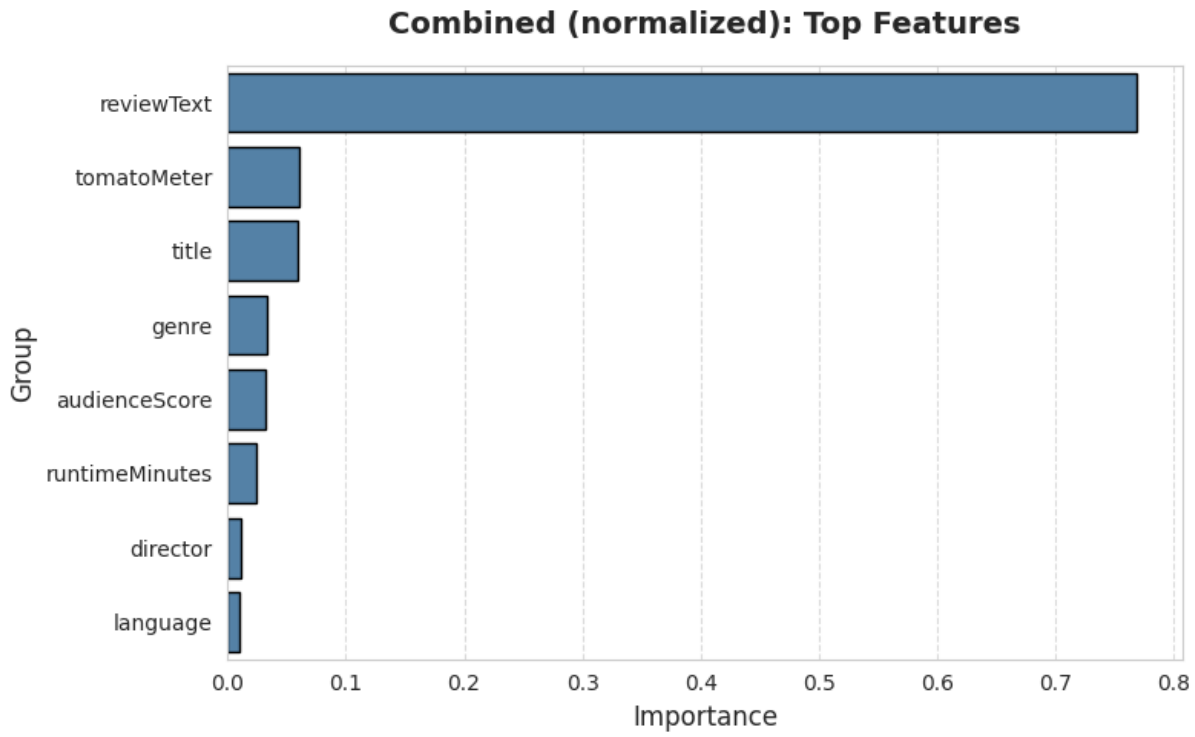

Figure 3.7: Combined Feature Importance Ranking

- **Feature Transformation :** Text features (`reviewText, title`) were transformed into TF-IDF vectors. Categori cal features (`genre, language`) were one-hot or multi hot encoded, and numeric features were scaled.
- **Importance Ranking :** Feature importance was computed from the scores from Random Forest, XGBoost, and LightGBM. This ensemble-based ranking provided robust selection across diverse feature types.

  Based on this ranking, three final feature sets were defined:

  1. All 8 Features (full set of metadata + text)
  2. Top 4 Features (highest-ranking predictors)
  3. Text-Only (`reviewText`)

3.3 Proposed Methodology and Design

3.3.1 Overview of the Experimental Framework

This research proposes a Dual-Stream Comparative Framework designed to systematically evaluate the impact of multimodal data on sentiment classification. Unlike unimodal ap-

proaches that rely solely on textual input, this framework treats sentiment analysis as a composite problem requiring the harmonization of linguistic signals and metadata context.

To achieve a comprehensive evaluation, the methodology is divided into two distinct architectural streams:

1. **The Statistical Stream:** Utilizes traditional Machine Learning (ML) algorithms (Logistic Regression, XGBoost, LightGBM) grounded in frequency-based text representation (TF-IDF).
2. **The Contextual Stream:** Utilizes Transformer-based Deep Learning architectures (DistilBERT, RoBERTa, DeBERTa) employing self-attention mechanisms and latent feature fusion.

Both streams are subjected to the same experimental conditions, operating across three feature environments (Review Text-Only, Top-4 Features, All Features) and two classification tasks (Binary and 3 class).

3.3.2 Stream A: Statistical Baseline Architecture

The first stream establishes a robust baseline using established ML algorithms. As these models cannot process raw text sequences, a feature engineering pipeline is designed to transform unstructured data into numerical vectors.

- **Text Representation:** The `reviewText` and `title` are vectorized using Term Frequency-Inverse Document Frequency (TF-IDF). This converts variable-length reviews into fixed-size sparse vectors, capturing the importance of specific keywords while filtering out common stop words.
- **Feature Fusion:** The sparse TF-IDF vectors are horizontally stacked (concatenated) with the standardized numerical metadata (e.g., `audienceScore`) and Encoded categorical metadata (e.g., `genre`).
- **Classifiers:** The fused vectors serve as input for three distinct classifiers:
  - **Logistic Regression:** Acts as a linear baseline.
  - **XGBoost:** Captures non-linear relationships via gradient boosting.
  - **LightGBM:** Optimized for speed and handling large-scale tabular data.

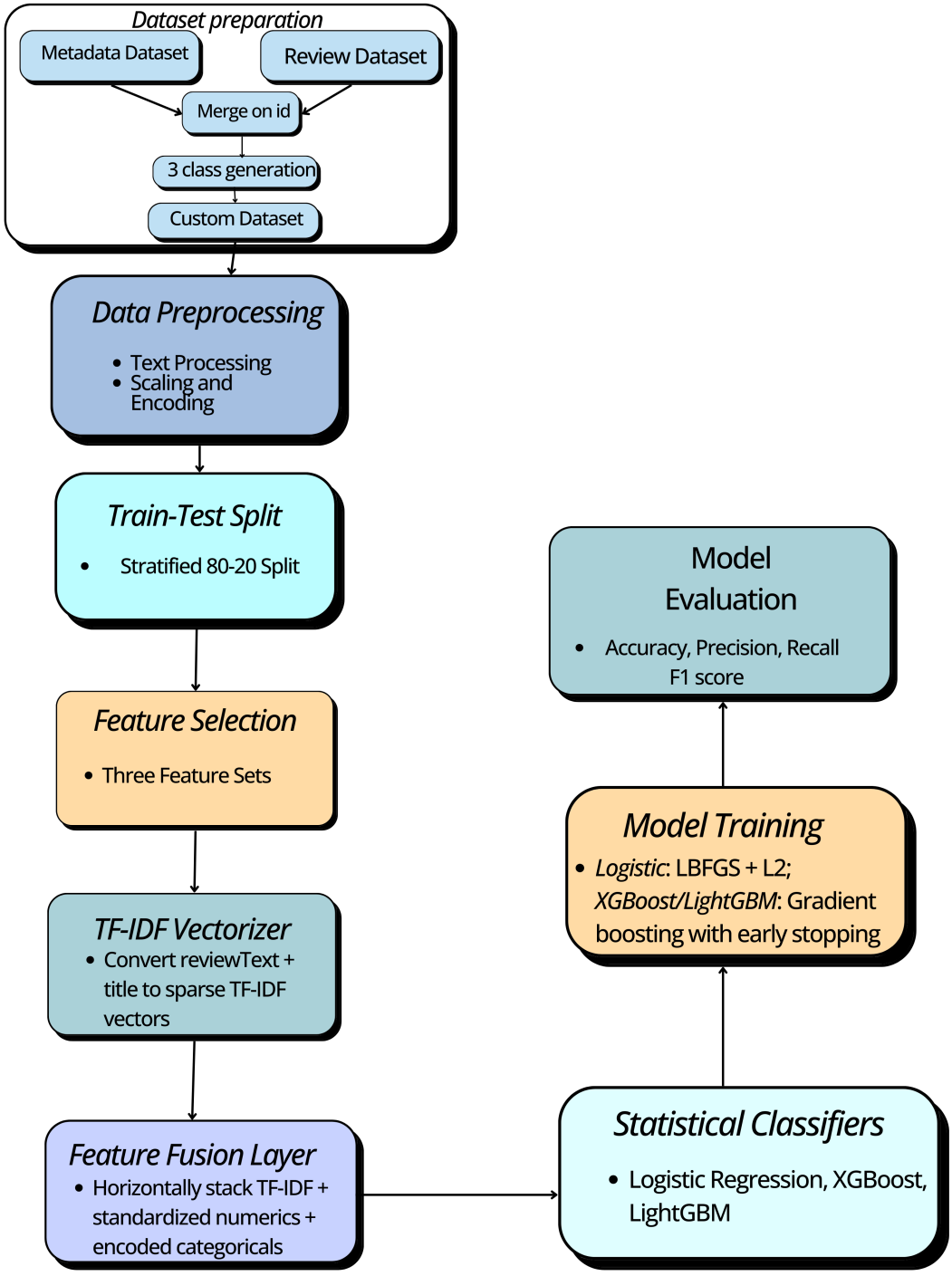

Figure 3.8: Statistical (ML) model pipeline

3.3.3 Stream B: Transformer-Based Late Fusion Architecture

The core contribution of this methodology is the Late Fusion Transformer Architecture. This design allows the model to learn semantic representations from text independently before integrating them with structured metadata.

The architecture consists of three sequential stages:

- **Semantic Encoding:** A pre-trained Transformer backbone (e.g., DistilBERT, RoBERTa, or DeBERTa) processes the tokenized `reviewText` and `title`. The model outputs a sequence of hidden states, from which the pooled output representing the `[CLS]` token is extracted. This vector, $h_{\text{text}} \in \mathbb{R}^d$, encapsulates the global semantic context of the review.
- **Multimodal Fusion:** The semantic vector h_{text} is concatenated with the structured metadata vector v_{meta} (containing normalized scores and encoded categories). This operation is formally defined as:

$$v_{\text{combined}} = \text{Concat}(h_{\text{text}}, v_{\text{meta}})$$

- **Classification Head:** The combined vector v_{combined} is passed through a fully connected dense layer (Feed-Forward Network). To prevent overfitting, a Dropout layer is applied prior to the final prediction.

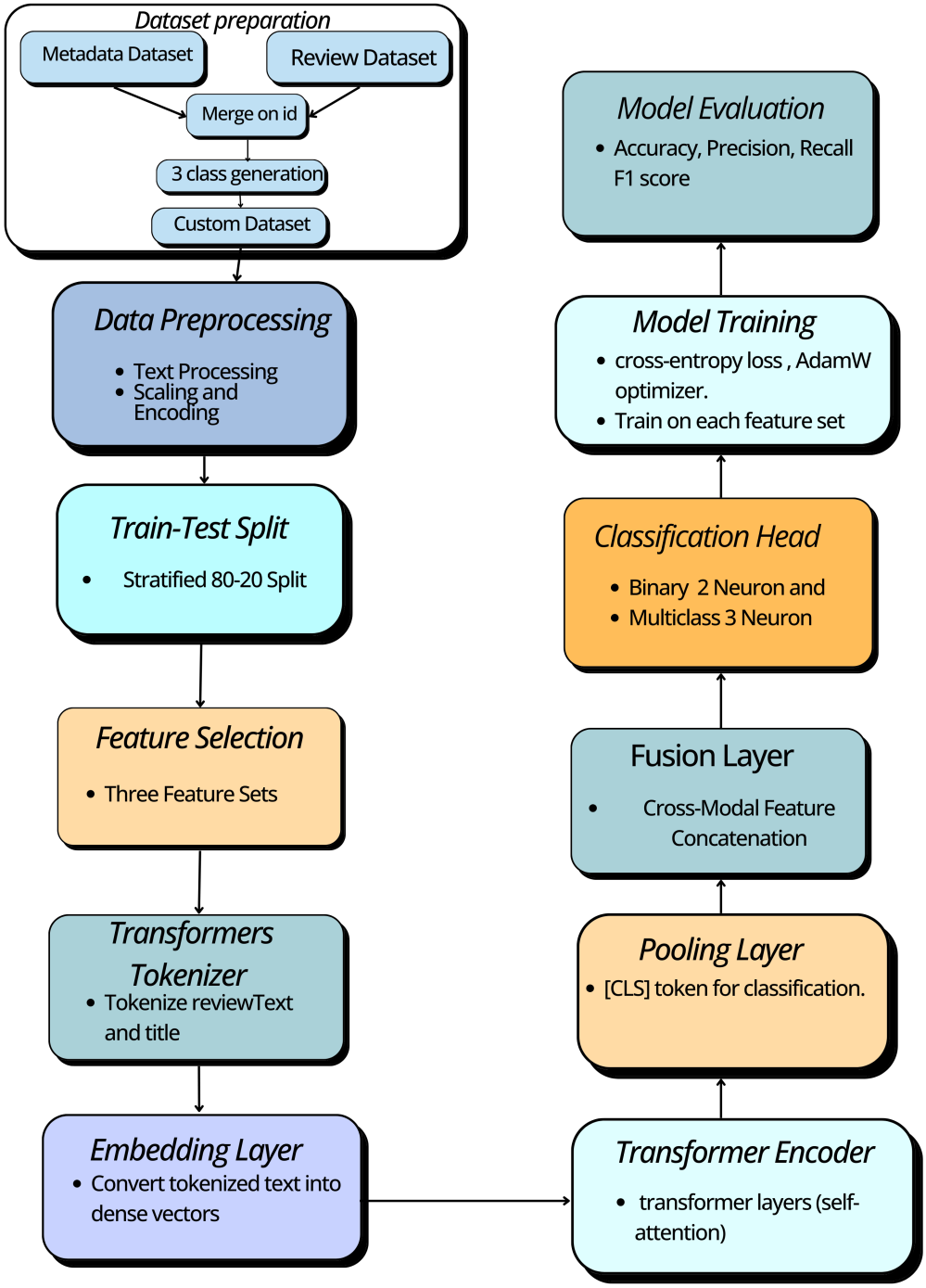

Figure 3.9: Transformer model pipeline

3.3.4 Experimental Design

To isolate the contribution of specific features, the analysis is conducted across a stratified experimental matrix. All six models (3 ML, 3 Transformers) are trained and evaluated on the following configurations:

- **Environment 1: Review Text-Only:** The model relies exclusively on the review content. This establishes the linguistic baseline.
- **Environment 2: Top-4 Metadata:** The model integrates text with the four highest-ranking features identified in the Feature Selection phase (`reviewText,tomatoMeter, title, genre`). This tests the efficiency of a reduced feature set.
- **Environment 3: All Features:** The model utilizes the complete set of text, numeric scores, and categorical tags.

3.4 Implementation

3.4.1 Computational Environment

The implementation was executed using Python 3.10 within a GPU-accelerated environment (Kaggle Kernels/Google Colab) utilizing NVIDIA Tesla T4/P100 GPUs to support the computational demands of Transformer fine-tuning. The primary frameworks employed were PyTorch and Transformers (Hugging Face) for deep learning, and Scikit-learn and XGBoost/LightGBM libraries for traditional machine learning tasks.

3.4.2 Implementation of Statistical Models (ML)

The implementation utilized a hybrid feature set where textual and metadata features were harmonized before training.

- **Feature Construction:** Textual inputs were converted into fixed-dimensional sparse representations using TF-IDF vectorization with high-capacity vocabulary limits to preserve lexical richness. These text embeddings were then fused with standardized numerical scores and encoded categorical metadata into a unified feature matrix, serving as the input for all statistical classifiers.
- **Logistic Regression:** The model was initialized with the `lbfgs` solver. For the multiclass problem, the `multi_class='multinomial'` parameter was set to utilize Softmax regression (estimating probabilities for all classes simultaneously) rather

than a One-vs-Rest approach. To address the class imbalance, particularly for the "Neutral" minority, `class_weight='balanced'` was applied to automatically adjust weights inversely proportional to class frequencies. L2 regularization was applied with a strength of C=0.5.

- **XGBoost & LightGBM:** Both ensemble methods were implemented using their native Scikit-learn APIs (`XGBClassifier` and `LGBMClassifier`).
  - **XGBoost:** Utilized the `binary:logistic` objective for binary tasks and `multi:softprob` for the 3 class task to generate probability distributions.
  - **LightGBM:** Was configured with the standard `gbdt` (Gradient Boosting Decision Tree) boosting type.
  - **Early Stopping:** To prevent overfitting, early stopping was enforced during the training loop, halting iterations if the validation loss failed to improve for 10 consecutive rounds.

3.4.3 Implementation of Multimodal Transformer Architectures

To effectively synthesize textual semantics with metadata features, custom `nn.Module` architectures were implemented using the PyTorch and Hugging Face libraries. The multimodal framework consisted of three distinct stages: input processing, feature fusion, and classification.

1. **Tokenization and Input Configuration:** Textual inputs (comprising the movie title and review text) were processed using the AutoTokenizer corresponding to each pre-trained backbone (distilbert-base-uncased, roberta-base, and microsoft/deberta-v3-base). To accommodate the computational constraints while retaining semantic context, input sequences were truncated or padded to a fixed maximum length of 256 tokens. This ensured uniform tensor dimensions for efficient batch processing.
2. **Network Architecture:** A custom class, MultimodalClassifier, was defined to integrate the disparate data streams. This architecture was consistent across all three transformer backbones:
   - **Text Branch:** The backbone utilized pre-trained weights from DistilBERT, RoBERTa, and DeBERTa to extract semantic features. For every review, the model computed the contextual embedding of the `[CLS]` token (or the pooled output) from the final hidden state to represent the document summary.

- **Tabular Branch:** Numerical features (`Score, Audience Score,Tomatometer`) were normalized using Standard Scaling to ensure magnitude consistency. Categorical features (`Genre, Language, Director`) were encoded and concatenated with the numerical vector.
- **Fusion & Classification:** The textual embedding vector was concatenated with the tabular feature vector. This fused representation was passed through a Dropout layer (rate = 0.2) for regularization, followed by a final Linear classification layer that mapped the combined features to the output logic space (2 dimensions for binary tasks, 3 for 3 class tasks).

3. **Training Configuration and Hyperparameters:** All models were trained using the AdamW optimizer, selected for its decoupling of weight decay from gradient updates.
   - **Hyperparameters:** A learning rate of $2e^{-5}$ was applied with a batch size of 16 and a weight decay of 0.01 to prevent overfitting. Training ran for 3 epochs, with the best model checkpoint saved based on the validation weighted F1-score.
   - **Loss Functions:**
     - **Binary Task:** Standard `CrossEntropyLoss` was utilized across all models.
     - **3 class Task:** To address the class imbalance inherent in the derived "Neutral" class, a custom trainer was implemented. This injected computed class weights into the `CrossEntropyLoss` function, penalizing the model more heavily for misclassifying minority classes.

3.4.4 Evaluation Metrics

To ensure a fair comparison across balanced and imbalanced tasks, the models were evaluated using a suite of standard performance metrics. While Accuracy provided a general measure of correctness, primary emphasis was placed on the Weighted F1-Score. This metric calculates the harmonic mean of Precision and Recall for each class and then computes the average weighted by support (the number of true instances for each label). This ensures that the model's performance on the minority "Neutral" class is adequately represented in the final score.

3.5 Summary

This chapter outlines a framework that integrates traditional machine learning with modern Transformer-based methods. The goal is to determine if fusing structured metadata

with raw text improves sentiment classification beyond what language alone can achieve. Instead of using arbitrary thresholds, this study uses a correlation-weighted scoring system to define a statistically valid 'Neutral' class, creating a more robust labeling strategy. The late-fusion architecture—applied across DistilBERT, RoBERTa, and DeBERTa preserves the semantic meaning of the reviews while integrating contextual details like genre and director. By systematically comparing text-only models against those enriched with metadata across multiple feature environments, this framework is designed to isolate the contribution of structured context particularly in linguistically ambiguous reviews and to empirically identify which metadata features most meaningfully enhance classification performance.

Chapter 4

Experiments and Results Analysis

4.1 Introduction/Overview

This chapter details the experimental evaluation of the proposed multimodal framework, benchmarking the performance of statistical baselines against state-of-the-art Transformer architectures. The core objective was to empirically determine how different combinations of unstructured text and structured metadata influence sentiment classification accuracy. To achieve this, a comprehensive suite of experiments was conducted across two distinct classification tasks: a standard Binary Classification (Positive/Negative) and a more complex 3 class Classification (Positive/Neutral/Negative).

Crucially, each model was tested in three stratified feature environments—Text-Only, Top-4 Features, and All Features—to isolate the predictive value of metadata. The results presented here are significant as they quantify the "accuracy gap" between traditional and modern methods and validate the necessity of the "Neutral" class for capturing the nuanced reality of audience opinion.

4.2 Modern Tools

To execute these experiments effectively, a stack of modern, industry-standard tools and frameworks was utilized. The relevance and effectiveness of each are described below:

- **PyTorch & Hugging Face Transformers:** These libraries served as the backbone for the Deep Learning stream. Hugging Face provided access to pre-trained weights for DistilBERT, RoBERTa, and DeBERTa V3, allowing us to leverage transfer learning. This was highly effective, as it enabled the models to start with a rich understanding of English syntax rather than training from scratch, significantly reducing computational

cost and time.

- **Scikit-learn:** Utilized for the statistical baselines (Logistic Regression) and evaluation metrics. Its robust implementation of TF-IDF vectorization and cross-validation ensured that our baseline comparisons were mathematically rigorous and reproducible.
- **XGBoost & LightGBM:** These gradient-boosting frameworks were chosen for their speed and state-of-the-art performance on tabular data. Their ability to handle feature interactions made them ideal for testing the limit of non-neural architectures.
- **GPU Acceleration (NVIDIA T4/P100):** Given the computational intensity of fine-tuning Transformers, experiments were conducted in GPU-accelerated environments (Kaggle Kernels/Google Colab). This hardware was essential for processing large batches of 256-token sequences in parallel, making the extensive grid of experiments feasible.

4.3 Result Analysis

This section presents a detailed analysis of the findings, organized by the complexity of the classification task. The primary metric for comparison is the Weighted F1-Score, which accounts for class imbalance and ensures the minority "Neutral" class is fairly represented.

4.3.1 Binary Classification Analysis (Positive vs. Negative)

The binary task served as a foundational benchmark to test the models' ability to distinguish broadly polarized sentiments.

A.Statistical Models (ML Baselines)

The traditional models (Logistic Regression, XGBoost, LightGBM) showed a heavy dependency on structured metadata. As seen in Table 4.1, relying on text alone limited their performance, but fusing metadata triggered substantial gains.

- **The Metadata Jump:** When trained on `reviewText` alone, ML models hovered around 74–76% accuracy and F1-score. However, adding the full metadata suite boosted performance to approximately 84%. This ∼8–10% jump proves that for models using simple frequency counts (TF-IDF), metadata provides the critical context needed to clarify sentiment.

- **Best Performer:** Logistic Regression performed comparably well in the "All Features" environment (∼84%), suggesting that for binary tasks, the relationship between features like `TomatoMeter` and `sentiment` is largely linear.

Table 4.1: Comparative Performance of ML Models (Binary)

Features	Model	Accuracy	Precision	Recall	F1
All	**Logistic Regression**	**84.13%**	**84.13%**	**84.13%**	**84.12%**
	XGBoost	80.03%	80.03%	80.03%	80.03%
	LightGBM	83.21%	83.22%	83.21%	83.21%
Top 4	Logistic Regression	80.82%	79.94%	82.47%	81.19%
	XGBoost	79.02%	78.83%	79.56%	79.19%
	LightGBM	80.47%	80.33%	80.88%	80.60%
Review Text	Logistic Regression	76.52%	76.59%	76.65%	76.62%
	XGBoost	71.79%	73.32%	68.84%	71.01%
	LightGBM	74.32%	74.77%	73.70%	74.23%

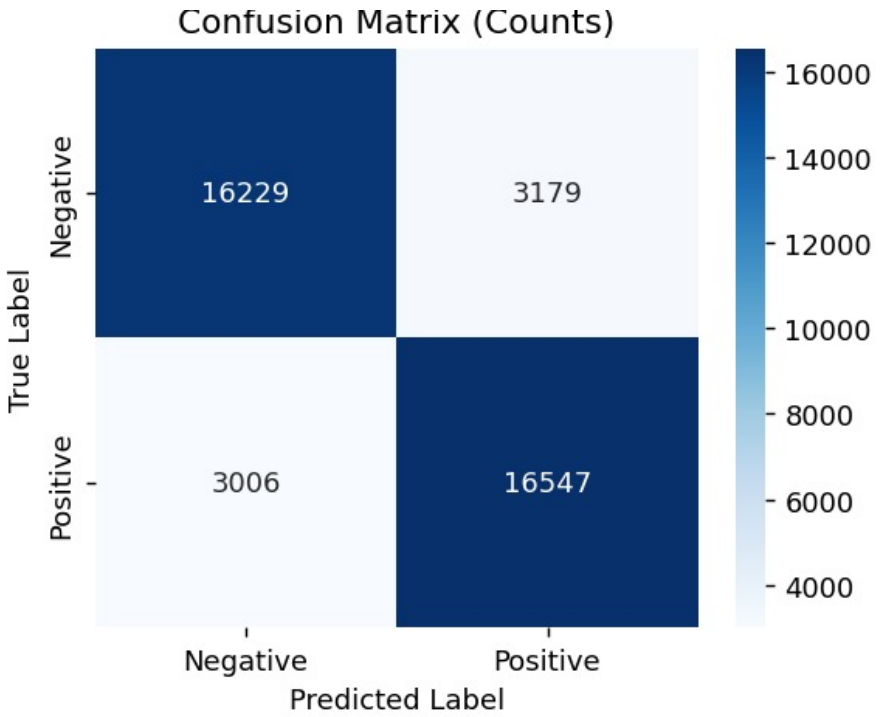

Figure 4.1: Confusion Matrix of Best Binary ML Model (Logistic Regression - All Features)

The strong diagonal indicates high precision in distinguishing positive from negative reviews.

Integrating structured metadata (All Features) yields a significant improvement in F1-Score compared to the text-only baseline.

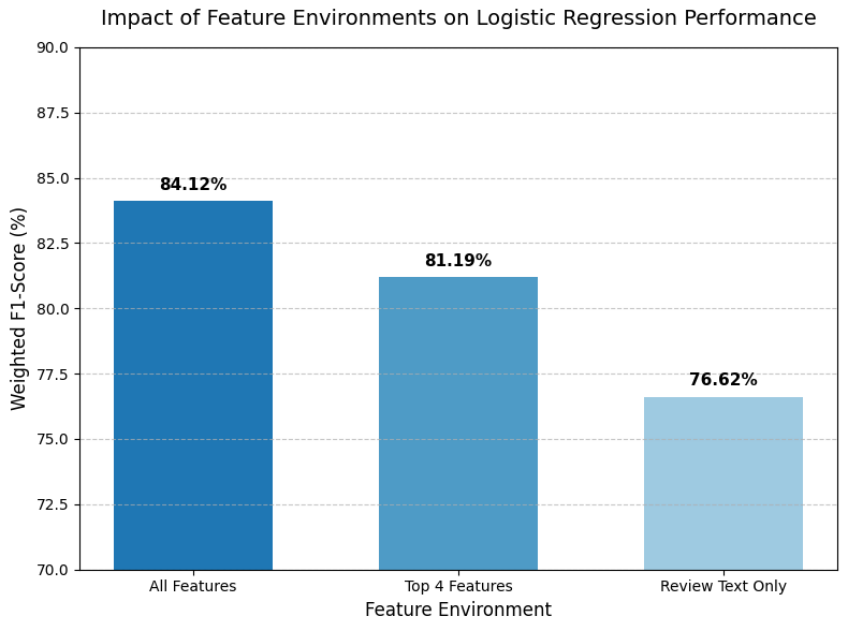

Figure 4.2: Impact of Feature Environments on Binary Logistic Regression.

B.Transformer Models (Deep Learning)

The Transformer architectures demonstrated superior semantic understanding, significantly outperforming the baselines.

- **DeBERTa Dominance: DeBERTa V3** emerged as the clear leader, achieving a 91.50% F1-score in the "All Features" setting. Even with text alone, it maintained 90.54%, proving its ability to disentangle complex language without needing metadata "hints."
- **Other Models Performance :** RoBERTa also performed well, achieving 90.36% F1 score in All feature environment beating DistillBERT's 88.73%.

Table 4.2: Comparative Performance of Transformer Models (Binary)

Features	Model	Accuracy	Precision	Recall	F1
All	DistillBERT	88.42%	88.41%	89.06%	88.73%
	DeBERTa	**91.50%**	**91.50%**	**91.50%**	**91.50%**
	RoBERTa	90.36%	90.38%	90.36%	90.36%
Top 4	DistillBERT	87.50%	88.87%	85.85%	87.32%
	DeBERTa	91.47%	91.47%	91.47%	91.47%
	RoBERTa	90.23%	90.24%	90.23%	90.23%
Review Text	DistillBERT	86.44%	86.65%	86.27%	86.46%
	DeBERTa	90.54%	90.57%	90.54%	90.54%
	RoBERTa	89.48%	89.48%	89.48%	89.48%

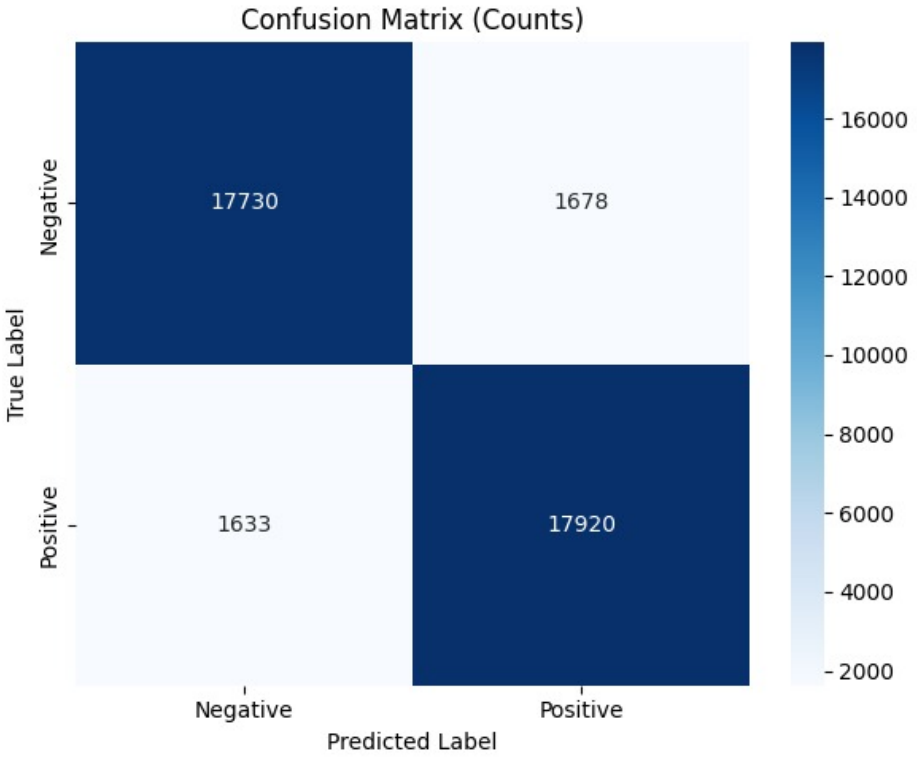

Figure 4.3: Confusion Matrix of Best Binary Transformer (DeBERTa- All Features)

The minimal false positives/negatives illustrate the model's robustness.

Integrating structured metadata (All Features) does not yield a significant improvement in F1-Score compared to the text-only baseline.

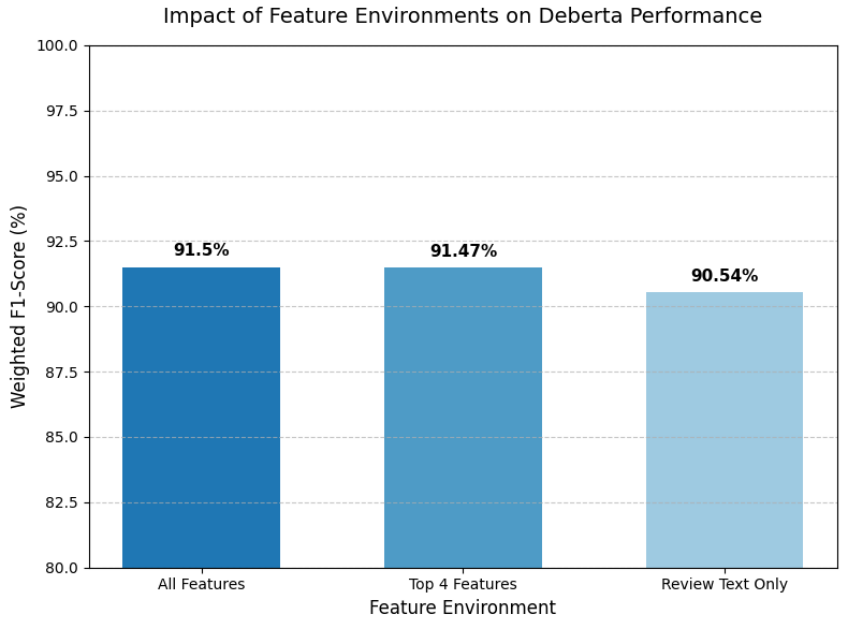

Figure 4.4: Impact of Feature Environments on Binary DeBERTa.

4.3.2 3 class Classification Analysis (Negative / Neutral / Positive)

The introduction of the "Neutral" class significantly increased difficulty, exposing the limitations of simpler models.

A. Statistical Models (ML Baselines)

Traditional models struggled to identify the "Neutral" sentiment, often confusing it with Positive or Negative.

- **The Neutral Bottleneck:** The best-performing ML model capped at a weighted F1-score of roughly 64%. Without metadata (`reviewText` Only), performance plummeted to ∼56–58%. Logistic Regression performed best here too in the All Features environment.
- **Feature Dependency:** This confirms that without the “anchor” of a numerical score, bag-of-words models cannot easily detect the subtle, mixed phrasing typical of neutral reviews.

Table 4.3: Comparative Performance of ML Models (3 class)

Features	Model	Accuracy	Precision	Recall	F1
All	**Logistic Regression**	**64.18%**	**63.36%**	**64.18%**	**63.51%**
	XGBoost	63.41%	62.92%	63.41%	63.07%
	LightGBM	63.49%	62.98%	63.49%	63.13%
Top 4	Logistic Regression	63.94%	63.06%	63.94%	63.20%
	XGBoost	62.94%	62.49%	62.94%	62.63%
	LightGBM	62.83%	62.29%	62.83%	62.46%
Review Text	Logistic Regression	58.42%	57.87%	58.42%	57.90%
	XGBoost	56.21%	56.01%	56.21%	55.92%
	LightGBM	56.38%	56.11%	56.38%	56.07%

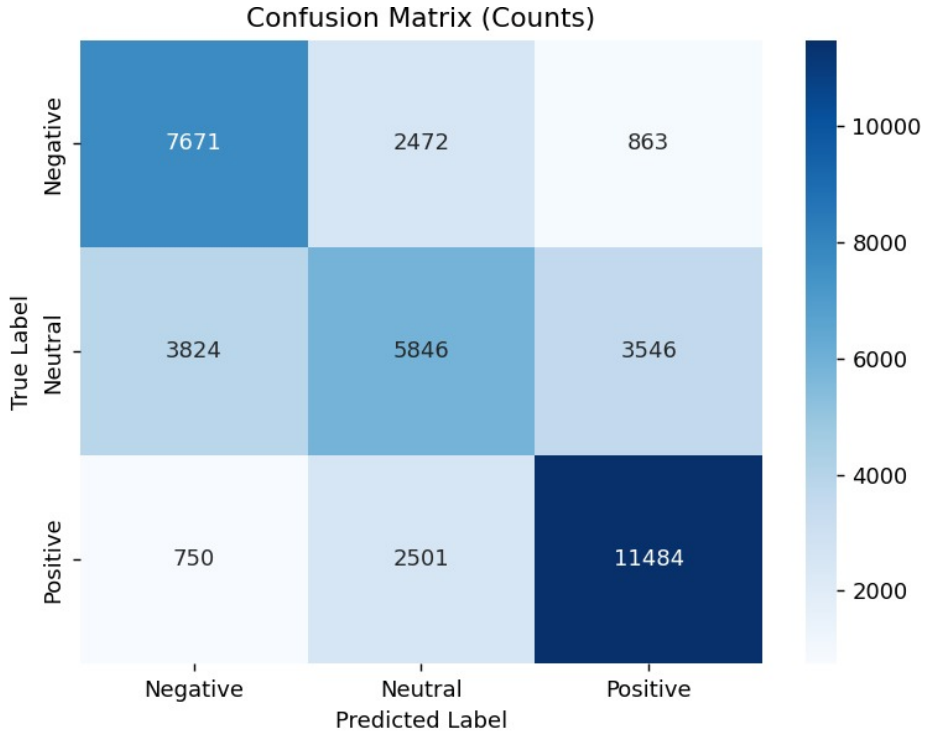

Figure 4.5: Confusion Matrix of Best 3 class ML Model (Logistic Regression - All Features)

The dispersion in the center indicates difficulty in classifying the Neutral class.

Integrating structured metadata (All Features) yields a significant improvement in F1-Score compared to the text-only baseline just like it did in binary for logistic regression

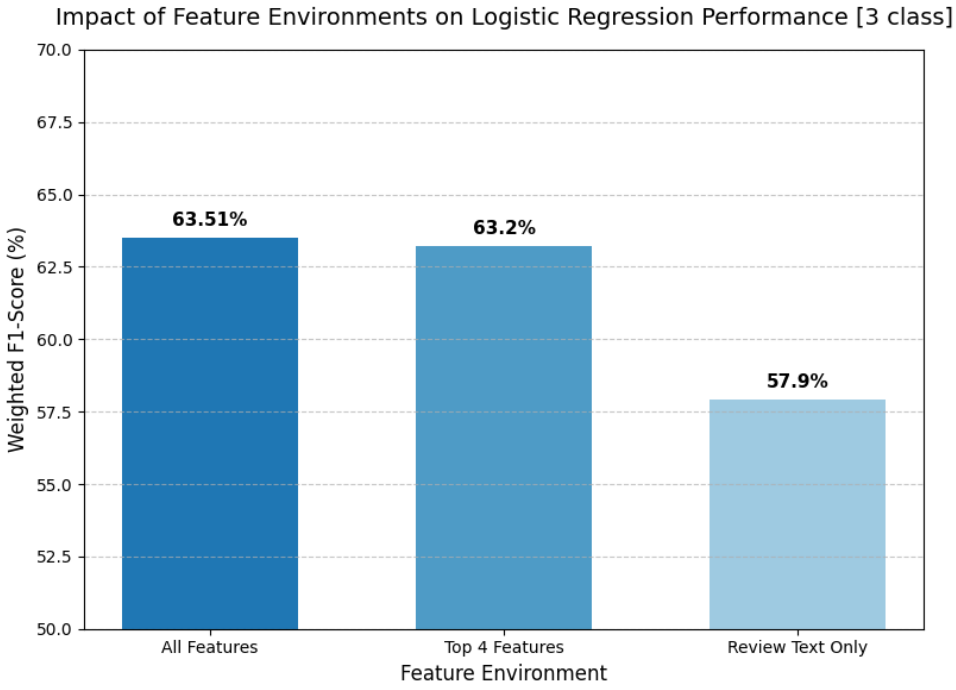

Figure 4.6: Impact of Feature Environments on 3 class Logistic Regression.

B. Transformer Models (Deep Learning)

This is where the Multimodal Transformer approach proved essential.

- **DeBERTa's Superiority:** DeBERTa (All Features) achieved a Weighted F1-score of 71.48%. This is an improvement over the best ML baseline, validating that attention mechanisms are required to capture "mixed feelings."
- **Other Models Performance :** RoBERTa also performed decent, achieving 70.30% F1 score in All feature environment beating DistillBERT's 67.52%.

Table 4.4: Comparative Performance of Transformer Models (3 class)

Features	Model	Accuracy	Precision	Recall	F1
All	DistillBERT	67.73%	67.36%	67.73%	67.52%
	DeBERTa	**71.94%**	**71.29%**	**71.94%**	**71.48%**
	RoBERTa	70.71%	70.11%	70.71%	70.30%
Top 4	DistillBERT	67.73%	66.99%	67.73%	67.14%
	DeBERTa	71.73%	71.19%	71.73%	71.38%
	RoBERTa	70.52%	69.91%	70.52%	70.11%
Review Text	DistillBERT	66.50%	66.17%	66.50%	66.31%
	DeBERTa	70.86%	70.26%	70.86%	70.45%
	RoBERTa	69.63%	69.01%	69.63%	69.18%

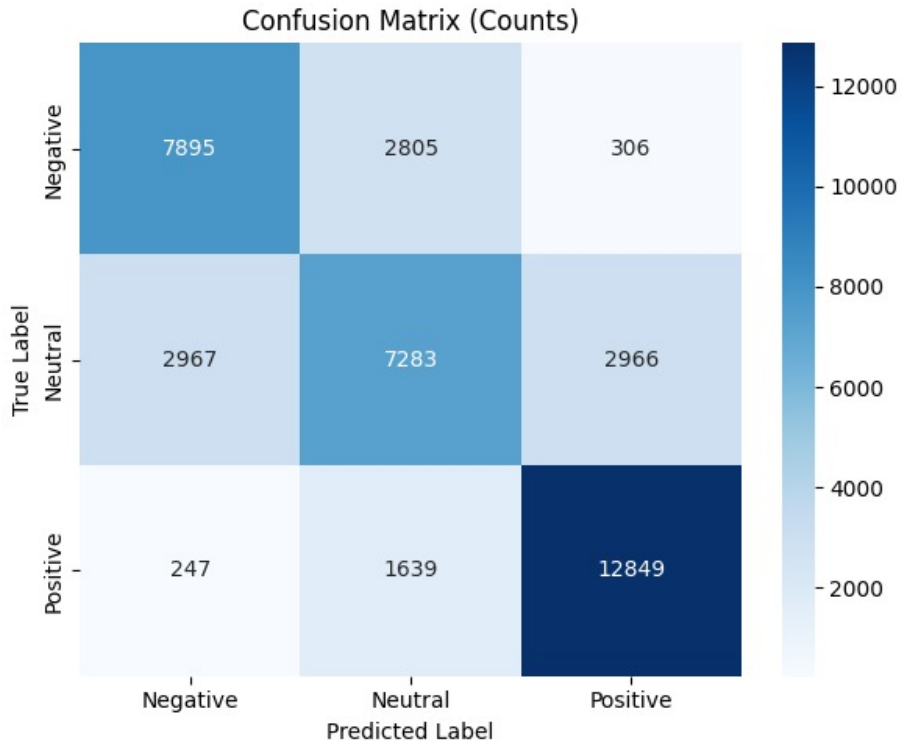

Figure 4.7: Confusion Matrix of Best 3 class Transformer (DeBERTa-All Features)

The sharper diagonal proves the model successfully learned the boundaries of the Neutral sentiment.

While integrating structured metadata (All Features) yields a modest improvement of 1.03% over the text-only baseline, this gain is meaningful in the context of a 3-class classification task. Given that multi-class performance is inherently lower than binary classification, this margin demonstrates the additive value of the feature environment.

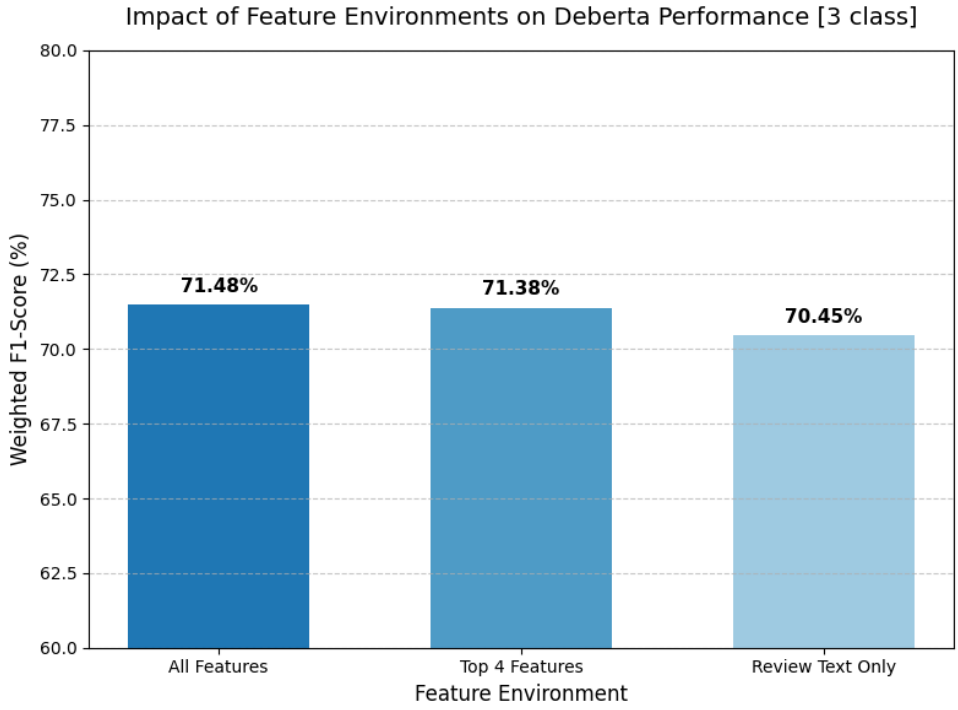

Figure 4.8: Impact of Feature Environments on 3 class DeBERTa

4.3.3 Impact of Feature Environments

A cross-cutting analysis of the experiments reveals consistent patterns regarding feature utility:

1. **"All Features" is Optimal:** Across every experiment, the "All Features" environment yielded the highest scores. This empirically validates that sentiment is multimodal; combining explicit signals (metadata) with implicit signals (text) creates the most robust classifier.

2. **Efficiency of "Top-4":** The "Top-4" environment (utilizing `ReviewText, Title, genre , TomatoMeter`) performed nearly as well as the full set, often within 1-2%. This finding is critical for efficiency, suggesting that just a few key metadata points are sufficient for high performance.

4.3.4 Comparative Analysis with Existing Literature

To validate the effectiveness of the proposed multimodal framework, we benchmarked our results against recent studies. The comparison highlights two key advantages: the power of metadata in binary classification and the robustness of our data-driven "Neutral" class in 3 class tasks.

A. Binary Classification on Rotten Tomatoes

For the standard positive/negative task, our DeBERTa V3 (All Features) model achieved a Weighted F1-score of 91.50%, outperforming recent benchmarks on similar Rotten Tomatoes datasets.

- **Khan et al.** [14]benchmarked advanced Transformers on 50,000 reviews, reporting that XLNet achieved 87.68% accuracy. Our framework surpassed this by approximately 3.8%, demonstrating that adding structured metadata (like `TomatoMeter`) provides critical context that text-only Transformers miss. On `reviewText` only environment also, our model outperformed it.
- **Soni et al.** [5] previously concluded that statistical classifiers like SVM (77.5%) offer superior accuracy over basic neural networks (58.6%), our findings reframe this narrative. We demonstrate that the integration of structured metadata with state-of-the-art Transformer architectures (DeBERTa V3) effectively reverses this performance gap.

B. 3 class Classification and the "Neutral" Gap

A major limitation in existing research is the lack of a standardized "Neutral" class. Most studies, such as Mishra et al. [9], rely on arbitrary manual labeling. Consequently, their models often fail to detect neutral sentiment, yielding F1-scores as low as 0.50. Similarly, Gao [17] achieved only $\sim$ 65% overall accuracy on a three-class IMDb task using a hybrid BERT model. In contrast, our framework achieved a Weighted F1-score of 71.48%. This superior performance is largely attributed to our novel target formulation. Instead of using arbitrary manual labels, we utilized Point-Biserial Correlation to mathematically derive the "Neutral" class based on the actual statistical relationship between user scores and text. This data-driven approach reduced label noise, allowing the model to correctly identify nuanced, mixed sentiments that previous methods struggled to capture.

4.4 Impact Analysis and Sustainability

4.4.1 Impact on Industry and Practice

The findings of this research have direct implications for digital media platforms and recommendation engines. By demonstrating that a 3 class Transformer (specifically DeBERTa) can accurately identify sentiment with ∼72% F1-score, we offer a solution to the problem of "rating inflation," where mediocre movies are pushed into positive or negative extremes. A system based on this framework would provide users with more transparent and nuanced summaries of public opinion, potentially reducing churn and increasing trust in platform ratings.

4.4.2 Sustainability and Efficiency

From a sustainability perspective, this research highlights a critical trade-off between accuracy and computational cost.

- **Feature Efficiency:** The success of the "Top-4" environment contributes to data sustainability. It demonstrates that models do not need massive, diverse features to perform well; a small, curated set of high-value metadata is sufficient. This reduces the storage and processing burden for future deployments.
- **Model Selection:** While DeBERTa achieved the highest accuracy, DistilBERT (a smaller, faster model) achieved respectable binary performance with significantly lower energy consumption. For resource-constrained environments, deploying a DistilBERT + Top-4 Metadata model represents a sustainable alternative that balances performance.

4.5 Summary

This chapter presented a rigorous empirical evaluation of the proposed multimodal framework. The results confirmed that DeBERTa V3 is the superior architecture for handling complex sentiment, particularly when tasked with identifying the nuanced "Neutral" class. However, the experiments also revealed that traditional ML models, while less accurate, benefit disproportionately from metadata integration.

Key takeaways include:

1. **Multimodal Synergy:** Combining text with metadata consistently outperforms text-only approaches, particularly for identifying neutral sentiment.

2. **The Neutral Challenge:** While ML models struggle to define neutrality (F1 ∼64%), Transformer models can effectively learn this boundary (F1 ∼72%) when supported by metadata.

3. **Data Efficiency:** The "Top-4" features capture the vast majority of the predictive signal, offering a roadmap for building efficient, sustainable classification systems in the future.

Chapter 5

Project Management and Cost Analysis

5.1 Project Management

Effective project management was fundamental to the successful execution of this research, which aimed to develop a robust multimodal sentiment analysis framework. The project was structured into four major phases: planning, execution, monitoring, and control.

Planning

During the initial stage, the research scope was defined to address the limitations of unimodal text analysis by integrating structured metadata. The project requirements were mapped out, identifying the critical need for high-performance computational resources. Specifically, the planning phase secured access to GPU-accelerated environments (NVIDIA T4/P100) to ensure the feasibility of fine-tuning large Transformer architectures like DeBERTa and RoBERTa within the project timeline.

Execution

This phase focused on the rigorous implementation of the "Dual-Stream Comparative Framework." We executed a complex preprocessing pipeline to harmonize heterogeneous data, normalizing inconsistent numerical scores and mathematically deriving the "Neutral" target class using Point-Biserial Correlation. Development involved coding custom `nn.Module` classes in PyTorch to construct the "Late Fusion" architecture, effectively concatenating text embeddings with metadata vectors for the final classification tasks.

Monitoring

Continuous monitoring was carried out to track model performance, specifically prioritizing the Weighted F1-Score over simple accuracy to ensure the minority "Neutral" class was adequately represented. Training loops were closely observed for signs of overfitting using validation loss metrics. Furthermore, feature importance rankings were generated using ensemble methods to monitor which metadata attributes, such as `TomatoMeter`, were contributing most significantly to the model's predictive power other than `reviewText`.

Control

Corrective actions were taken to address data biases and computational bottlenecks encountered during training. To control for "Popularity Bias," where blockbuster movies dominated the dataset, a strict cap was enforced on the number of reviews per movie to ensure generalization. Additionally, when class imbalance threatened model fairness, specific class weights were injected into the loss functions, and Early Stopping was implemented to halt training when validation performance stagnated, optimizing resource usage.

5.2 Cost Analysis

Although this research was conducted as an academic thesis, estimating the financial resources required is critical for understanding the feasibility of reproducing the study or transitioning it into a commercial application. The cost breakdown focuses primarily on the direct expenses related to high-performance computing and specialized software tools.

5.2.1 Direct Computational and Software Expenses

The primary financial expenditure stemmed from the need for specialized hardware to fine-tune large-scale Transformer architectures (DeBERTa, RoBERTa) and premium tools for architectural visualization.

- **Google Colab Pro (GPU Acceleration):** To manage the massive review dataset and perform fine-tuning without memory errors, the project required access to high-RAM runtimes and NVIDIA Tesla T4 GPUs. A subscription was maintained strictly during the intensive experimental phase.
  - *Rate:* $9.99 per month.
  - *Duration:* 5 Months.

  - *Total:* $\sim$ \$50 USD ($\approx$ 6,000 BDT).

- **Canva Pro (Visual Design):**Premium design tools were utilized to create system architecture diagrams (e.g., the Late Fusion Pipeline) and high-quality presentation assets for the defense, pre defense and proposal.

  - *Rate:* \$12.99 per month.
  - *Duration:* 3 Months.
  - *Total:* $\sim$ \$39 USD ($\approx$ 4,680 BDT).

- **Cloud Storage (Google One):**Storing multiple checkpoints of large language models (which can exceed several gigabytes each) and the uncompressed dataset required expanding cloud storage beyond the standard free tiers.

  - *Rate:* $\sim$ \$2.00 per month (100GB Plan).
  - *Duration:* 12 Months.
  - *Total:* $\sim$ \$24 USD ($\approx$ 2,880 BDT).

5.2.2 Human Capital and Operational Utilities

Beyond the direct software costs, the project required a significant investment in intangible assets and operational utilities. These represent the bulk of the true "cost" of the research, particularly when viewed through the lens of strict project deadlines.

Human Capital

The research was executed by a team of three members working intensively over a rigid one-year timeline (two academic semesters). This represents a substantial allocation of human resources, as the team had to adhere to strict milestones while navigating distinct phases such as literature review, algorithm design, coding, and documentation. The time constraints inherent in an undergraduate thesis required a focused and accelerated work pace to ensure all experimental validations were completed before the final defense.

Operational Utilities

A continuous and reliable high-speed internet connection was indispensable, particularly given the time-sensitive nature of the project. Reliability was critical, as any connectivity downtime would have jeopardized the timely downloading of massive datasets (Rotten Tomatoes archives) or the syncing of large model weights with the Hugging Face hub. When

combined with the electricity consumption required for long-duration local testing and debugging, these utility costs represent a consistent operational overhead that was essential for meeting the project's tight submission schedules.

5.3 Project Scheduling

A structured and realistic schedule was vital for ensuring that the research progressed smoothly and met all academic deadlines without compromising the depth of the experimental analysis. The project was executed over a 12-month timeline (December 2024 to November 2025), divided into seven well-defined phases. Each phase had specific objectives to maintain focus and facilitate effective resource management. The following phases constituted the project lifecycle:

- **Phase 1: Planning and Literature Review (Dec '24 - Jan '25):** The initial two months were dedicated to defining the research scope and identifying gaps in existing sentiment analysis studies. The team conducted a comprehensive review of unimodal vs. multimodal approaches, specifically analyzing the limitations of binary classification in prior works. The "Dual-Stream Comparative Framework" was conceptualized during this phase.
- **Phase 2: Data Collection and Preprocessing (Feb '25 - Mar '25):** This phase focused on acquiring the "Massive Rotten Tomatoes" dataset. Heterogeneous data types were harmonized: the inconsistent "`originalScore`" attributes (fractions, letter grades) were normalized to a standard 0-10 scale, and text data was cleaned.
- **Phase 3: Feature Selection and Ranking (Apr '25):** To ensure model efficiency, a rigorous feature importance analysis was conducted. Ensemble methods (Random Forest, XGBoost) were utilized to rank metadata attributes, identifying the "Top-4" features that would be used in the stratified experiments.
- **Phase 4: Binary Model Development (May '25 - Jun '25):** The team developed and fine-tuned the initial binary classifiers (Positive/Negative). This involved implementing the TF-IDF baselines (eg., Logistic Regression) and the first iteration of Transformer models (DistilBERT, RoBERTa) on the physically balanced dataset to establish a performance benchmark.
- **Phase 5: 3 class Generation and Training (Jul '25 - Aug '25):** This was the most computationally intensive phase. The "Neutral" class was mathematically generated using the weighted composite index. The "Late Fusion" architecture was finalized, and

the DeBERTa V3 models were trained on the 3 class dataset using Google Colab Pro, applying class weights to handle the new 3-class distribution.

- **Phase 6: Evaluation and Result Analysis (Sep '25 - Oct '25):** The results from all six architectures across the three feature environments (Text-Only, Top-4, All Features) were consolidated. Confusion matrices were generated to analyze misclassifications, and the Weighted F1-Scores were computed to validate the hypothesis that metadata integration improves "Neutral" sentiment detection.

- **Phase 7: Thesis Writing and Defense Preparation (Nov '25):** The final month was dedicated to compiling the comprehensive thesis book using Overleaf. System architecture diagrams were finalized using Canva Pro, and the final presentation slides were prepared for the project defense.

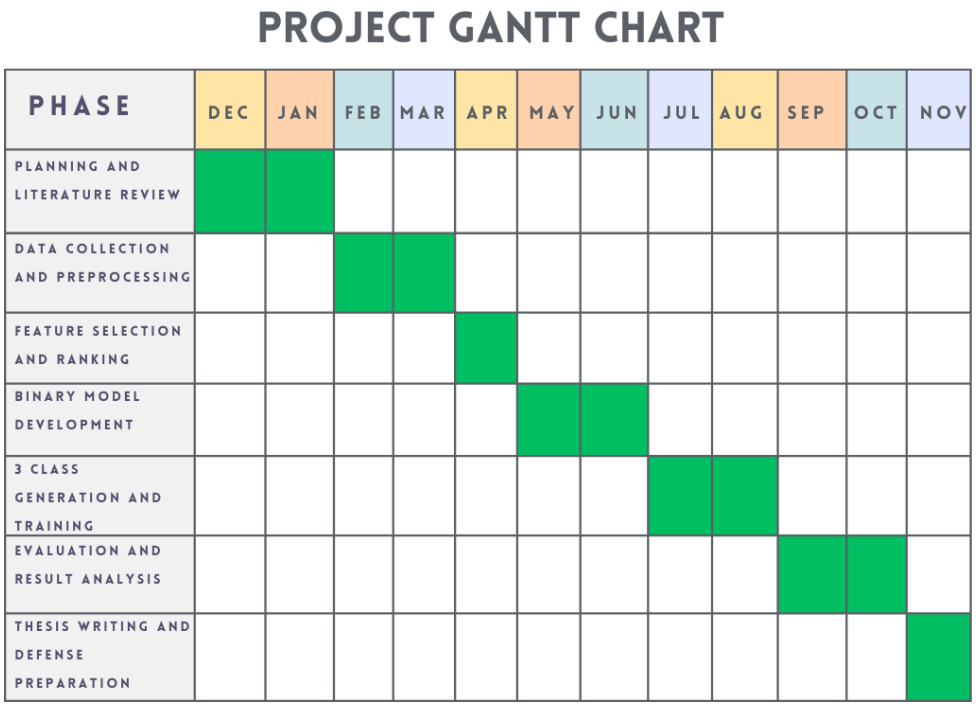

Figure 5.1: Gantt Chart of project Schedule

Chapter 6

Ethics and Professional Responsibilities

6.1 Introduction/Overview

In real-world engineering, ethics isn't just a box to check but it's woven into how we build and deploy technology. Nowhere is this more true than in NLP, where the models we design don't just process data, they interpret human expression. As these systems increasingly influence what people see, believe, or buy from movie recommendations to public sentiment summaries—they carry real responsibility. If we're not careful, they can amplify biases, ignore nuance, or present skewed views as fact.

Therefore, in this sentiment analysis project, we treated ethics as part of the design process itself. Movie reviews aren't just positive or negative rather they're often mixed, hesitant, or context-dependent. Reducing them to a binary label doesn't just lose meaning—it risks misrepresenting how real people feel. So instead of defaulting to the easiest technical solution, we built a framework that respects complexity. By combining transparent baselines with deep learning and grounding our labels in data we aimed to create a model that's not only accurate but also fair and interpretable.

6.2 Identify and Apply Ethical and Professional Responsibilities

From the first line of code to the final evaluation, we kept ethical and professional standards at the forefront. Here's how that played out in practice:

- **Responsible Data Use and Privacy:** We used the "Massive Rotten Tomatoes Movies & Reviews" dataset from Kaggle—a public resource shared under open terms. All

reviews are user-generated and already publicly visible, so our analysis stayed within ethical boundaries. We never tried to identify individuals, track behavior, or extract personal details. Our focus was solely on the text and its associated metadata (like genre or scores), not the people behind the words.

- **Resisting Oversimplification:** Rather than force every review into "positive" or "negative," we introduced a rigorously defined Neutral class using a correlation-weighted composite of scores and labels. It acknowledges that mixed opinions are valid and deserve their own category, not forced misclassification. We also applied class weighting to ensure the model paid proper attention to this underrepresented group.

- **Transparency Over Black Boxes:** Transformer models are powerful, but they're often opaque. To counter that, we didn't just rely on RoBERTa or DeBERTa in isolation—we compared them side-by-side with interpretable models like Logistic Regression and XGBoost. We documented every preprocessing step, shared feature importance rankings, and used consistent evaluation protocols.

- **Guarding Against Popularity Bias:** Blockbuster movies dominate most review datasets, which can make models blind to indie films or niche genres. To avoid this, we capped the number of reviews per movie and used stratified sampling during evaluation. This helped ensure our models generalize across a wide range of films, not just the most hyped ones, making the system more equitable for diverse content.

- **Honest Academic Practice:** All tools, libraries, and pre-trained models (like those from Hugging Face) were properly cited. We reported results truthfully—including where models struggled—and avoided overstating capabilities.

Chapter 7

Identification of Complex Engineering Problems and Activities

7.1 Complex Engineering Problem

7.1.1 Complex Problem Solving

In this section, we map the challenges encountered during the development of the multimodal sentiment analysis framework against standard complex problem-solving categories. The research addresses the ambiguity of human sentiment by integrating Transformer architectures (DeBERTa, RoBERTa) with statistical metadata analysis. The project maps onto these characteristics as follows:

Table 7.1: Mapping with complex problem solving.

P1	P2	P3	P4	P5	P6	P7
Depth of Knowledge	Range of Conflicting Requirements	Depth of Analysis	Familiarity of Issues	Extent of Applicable Codes	Extent of Stakeholder Involvement	Interdependence
✓	✓	✓	✓	✓		✓

Depth of Knowledge Required (P1)

Addressing this problem required knowledge beyond fundamental engineering principles, necessitating an integration of advanced mathematics, specialist computer science topics, and modern engineering practices.

WK2 (Conceptually based mathematics & statistics): The project utilized the Point-Biserial Correlation coefficient to statistically derive a weighted composite index for target generation, ensuring the "Neutral" class was mathematically grounded rather than arbitrarily defined.

WK3 (Engineering Fundamentals): The research established a systematic Dual-Stream Comparative Framework, applying fundamental engineering principles to harmonize heterogeneous data (text and metadata) within a rigorous preprocessing and fusion pipeline to ensure structural consistency.

WK4 (Specialist Knowledge): The implementation required in-depth knowledge of Natural Language Processing (NLP) and state-of-the-art Transformer architectures (Self-Attention mechanisms in DeBERTa/RoBERTa) which represent the forefront of the discipline.

WK5 (Engineering Design): The study addressed design optimization and efficient resource use by evaluating feature sustainability, demonstrating that a reduced Top-4 feature set could achieve near-optimal performance while minimizing computational overhead and storage requirements.

WK6 (Engineering Practice): The project applied modern engineering tools and standards, utilizing PyTorch, Hugging Face Transformers, and GPU acceleration (NVIDIA T4) to implement scalable deep learning solutions in line with current industry practices.

WK8 (Research Literature): The study required critical evaluation of current literature to identify the gap in multimodal analysis and the limitations of binary-only classification in existing benchmarks.

Range of Conflicting Requirements (P2)

The system design presented significant conflicting requirements between predictive accuracy, computational efficiency, and model transparency:

- **Accuracy vs. Complexity:** While the DeBERTa V3 model provided the highest F1-score of 91.50% , it is computationally heavy and resource-intensive. Conversely, the "Top-4 Feature" subset offered high efficiency with minimal accuracy loss , creating a trade-off between maximizing raw performance and ensuring system sustainability.
- **Class Imbalance vs. Minority Detection:** The "Neutral" class is naturally underrepresented in the dataset. Maximizing overall accuracy typically favors the majority classes (Positive/Negative), while properly detecting the "Neutral" minority requires applying heavy class weights. This creates a conflict where prioritizing the minority class prevents the model from simply defaulting to the majority labels for higher "easy"

accuracy.

Depth of Analysis Required (P3)

The problem has no obvious solution and required abstract thinking to resolve the "unimodal blindness" of traditional models.

- The analysis involved creating a novel Late Fusion architecture to harmonize heterogeneous data types (unstructured text vs. structured numerical scores).
- It required a rigorous feature importance analysis using ensemble ranking to determine that metadata (like `TomatoMeter`) is decisive in resolving linguistic ambiguity in sarcastic or short reviews.
- The generation of the "Neutral" class demanded deep statistical analysis to move beyond arbitrary manual labeling. This involved formulating a mathematically derived composite index using Point-Biserial Correlation and linear interpolation to dynamically define class boundaries, requiring a fundamental rethinking of how sentiment targets are defined.

Familiarity of Issues (P4)

The research navigated issues that are not frequently encountered in standard undergraduate sentiment analysis projects.

- Standard projects typically treat sentiment as a binary (Positive/Negative) text problem.
- This project dealt with the "Neutral" class, which is rarely defined mathematically in literature. Resolving the ambiguity of mixed reviews using a correlation-weighted index places this problem outside the scope of routine coding tasks.

Extent of Applicable Codes (P5)

The solution required more than just using standard tools. We had to design custom logic to make the system work correctly.

- **Custom Model Structure:** We could not use a standard "off-the-shelf" model because they only handle text. We had to write custom code (using PyTorch `nn.Module`) to build a special layer that combines text with numbers (scores and genres).

- **Custom Training Logic:** We could not use the standard error calculation. We implemented a custom loss function (Weighted Cross-Entropy) to force the model to pay attention to the "Neutral" class, ensuring it wasn't ignored just because it had fewer data points.

Interdependence (P7)

The problem involves a high level of interdependence between different sub-disciplines.

- The system relies on the interplay between Linguistics (semantic understanding of sarcasm/slang) and Statistical Data Science (numerical correlation of metadata).
- A failure in the preprocessing of metadata (e.g., incorrect normalization of `original_Score`) would directly degrade the performance of the Transformer model, proving the tight coupling between the text and tabular data streams.

7.1.2 Engineering Activities

This research aligns with complex engineering activities defined by their innovative nature, use of diverse resources, and societal consequences. The mapping is detailed below:

Table 7.2: Mapping with complex engineering activities.

A1	A2	A3	A4	A5
Range of Sources	Level of Interaction	Innovation	Consequences for Society and Environment	Familiarity
✓	✓	✓	✓	✓

Range of Resources (A1)

The project utilized a diverse range of resources to execute the experimental framework:

- **Data Resources:** The "Massive Rotten Tomatoes" dataset from Kaggle.
- **Computational Hardware:** Utilization of NVIDIA T4/P100 GPUs via cloud environments (Colab/Kaggle) to handle the massive matrix operations required for fine-tuning Transformers.
- **Software Libraries:** Integration of distinct libraries including PyTorch and Hugging Face Transformers for deep learning, and Scikit-learn and XGBoost for statistical benchmarking.

Level of Interaction (A2)

The project required managing interactions between conflicting technical disciplines and objectives.

- It bridged the gap between Natural Language Processing (NLP) and Tabular Data Analysis.
- It involved managing the interaction between conflicting user signals. The system had to reconcile instances where a user's numerical rating (e.g., 5/10) did not align with their binary category (Positive/Negative), using statistical correlation to merge them into a unified "Neutral" target.

Innovation (A3)

The core innovation lies in the Data-Driven Target Formulation.

- Unlike creative manual labeling, this project utilized the Point-Biserial Correlation to create a dynamic thresholding logic. This creative use of engineering principles allowed for the automated discovery of "Neutral" sentiment, applying a statistical solution to a linguistic problem.
- The multimodal fusion strategy specifically improving Transformer performance by injecting external metadata context represents a creative departure from standard text-only analysis.

Consequences for Society and Environment (A4)

The engineering activity has significant consequences for digital media consumption and data ethics.

- **Reducing Bias:** By successfully identifying "Neutral" and mixed reviews (71.48% F1-score), the model prevents the polarization of public opinion, where average movies are unfairly destroyed or hyped.
- **Transparency:** The framework combats "rating inflation," providing users and recommendation engines with a more honest, nuanced reflection of audience sentiment.
- **Sustainability:** The "Top-4 Features" finding demonstrates that high-performance models can be built with reduced data requirements, promoting computational efficiency.

Familiarity (A5)

The activities undertaken differed significantly from routine engineering tasks.

- The process of writing a custom fusion layer to concatenate BERT embeddings with normalized scalar values is not a standard practice in introductory courses.
- The rigorous benchmarking of six different architectures across three stratified environments represents a complexity of experimental design typically found in graduate-level research rather than standard software development.

Chapter 8

Conclusion and Future Works

This research demonstrates that effective sentiment analysis requires more than linguistic patterns, as it fundamentally demands contextual awareness. By introducing a statistically grounded "Neutral" class and evaluating models across varied feature environments, we established that structured metadata is not auxiliary. Rather, it is often decisive, especially when interpreting ambiguous or mixed reviews.

In the binary classification task, the fusion of metadata proved to be a critical equalizer. We observed that integrating numerical signals, such as genres and tomatometer ratings, provided necessary grounding for simple statistical models. This allowed them to overcome the limitations of ambiguous language. Simultaneously, for advanced Transformer architectures, this multimodal approach solidified their ability to detect polar sentiment with exceptional precision and set a new standard for the task.

The impact of this framework was even more consequential in the 3 class setting, where the complexity of identifying "Neutral" sentiment exposed the inherent limits of traditional approaches. While statistical baselines struggled to resolve the ambiguity of mixed reviews, attention-based architectures demonstrated a superior ability to disentangle nuanced opinions when supported by context. Furthermore, the consistent success of reduced feature sets across all experiments confirms that high performance does not require exhaustive complexity. Strategic metadata selection is sufficient to build robust, efficient, and scalable systems.

Looking ahead, we will pursue a hybrid fusion architecture that moves beyond simple late concatenation. Our next step is to develop a cross-attention mechanism where metadata directly modulates the Transformer's internal attention weights—allowing the model to dynamically adjust its linguistic understanding based on contextual cues, closing the gap between surface-level correlation and true semantic grounding. Ultimately, this direction promises not just higher accuracy, but more interpretable and context-aware sentiment systems that respect the richness of human expression.

References

[1] Y.-L. Chiu, J. Du, Y. Sun, and J.-N. Wang, "Do critical reviews affect box office revenues through community engagement and user reviews?," *Frontiers in Psychology*, vol. 13, p. 900360, 2022.

[2] I. Steinke, J. Wier, L. Simon, and R. Seetan, "Sentiment analysis of online movie reviews using machine learning," *International Journal of Advanced Computer Science and Applications*, vol. 13, 01 2022.

[3] D. Dey, "Smart movie review analysis: A data-driven sentiment classification system," 02 2025.

[4] A. Lahase and S. N. Deskmukh, "Sentiment classification of movie review using machine learning approach," *International Journal of Computer Science Engineering and Information Technology*, vol. 8, pp. 29–35, 11 2024.

[5] K. Soni, Rahul and P. Yadav, "Comparative analysis of rotten tomatoes movie reviews using sentiment analysis," 05 2022.

[6] U. Dahir and K. Alkindy Faisal, "Utilizing machine learning for sentiment analysis of imdb movie review data," *International Journal of Engineering Trends and Technology*, vol. 71, pp. 18–26, 05 2023.

[7] s. Y. I. Tomal, "Sentiment analysis of imdb movie reviews," *International Journal of Innovative Science and Research Technology (IJISRT)*, pp. 2338–2343, 06 2024.

[8] Y. Wu, "Movie recommendation system using knn, cosine similarity and collaborative filtering," *Highlights in Science, Engineering and Technology*, vol. 85, pp. 339–346, 03 2024.

[9] R. Mishra, H. Hemant, and P. Bajaj, "Movie review system using machine learning," 04 2024.

[10] A. D. Azahree and N. Alias, "Understanding viewer opinions: Sentiment analysis on movie review using vader and lstm model," 2024.

[11] J. Lu, H. Fan, and Y. Zhang, "Sentiment analysis of imdb movie reviews based on lstm," *Journal of Advances in Engineering and Technology*, vol. 2, 06 2025.

[12] A. Das, S. Chaudhuri, A. Murshed, R. Basak, and P. Singh, "Performance comparison of different deep learning ensemble models for sentiment classification of movie reviews," 08 2025.

[13] D. Subedi, E. N. Lamichhane, and N. Subedi, "Sentiment analysis of imdb movie reviews using svm and naive bayes classifier," *Journal of Engineering and Sciences*, vol. 4, pp. 56–68, 05 2025.

[14] S. S. Khan and Y. Alharbi, "Sentiment analysis of movie review classifications using deep learning approaches," *International Journal of ADVANCED AND APPLIED SCIENCES*, vol. 11, pp. 146–157, 09 2024.

[15] G. Nkhata, U. Anjum, and J. Zhan, "Sentiment analysis of movie reviews using bert," 02 2025.

[16] M. Hua, "Hybrid sentiment analysis model for movie reviews," *Applied and Computational Engineering*, vol. 185, pp. 78–84, 09 2025.

[17] R. Gao, "Bert+bilstm: Boosting performance in movie review sentiment classification," *Applied and Computational Engineering*, vol. 151, pp. 108–117, 05 2025.

[18] A. Villa, "Massive rotten tomatoes movies & reviews." https://www.kaggle.com/datasets/andrezaza/clapper-massive-rotten-tomatoes-movies-and-reviews, 2023. Kaggle, 2023, [Online]. License: CC0 (Public Domain).

Generated using Undergraduate Thesis LaTeX Template, Version 2.1.0. Department of Computer Science and Engineering, Ahsanullah University of Science and Technology, Dhaka, Bangladesh.

This thesis was generated on Friday 12th December, 2025 at 11:22am.