DarazMM: A Multilingual Multimodal E-commerce Product Review Dataset and Benchmark Evaluation

by

Md. Zahid Hossain
24366023

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
M.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
BRAC University
March 2026

Declaration

It is hereby declared that

1. The thesis submitted is my own original work while completing degree at BRAC University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Md. Zahid Hossain
24366023

Approval

The thesis titled "DarazMM: A Multilingual Multimodal E-commerce Product Review Dataset and Benchmark Evaluation"

By:
Md. Zahid Hossain (24366023)

Of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of M.Sc. in Computer Science on March 11, 2026.

Examining Committee:

Examiner:
(External)

Dr. Raihan Ul Islam
Associate Professor
Department of Computer Science and Engineering
East West University

Examiner:
(Internal)

Dr. Md. Ashraful Alam
Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Supervisor:
(Member)

Dr. Md. Golam Rabiul Alam
Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Program Coordinator:
(Member)

Dr. Md Sadek Ferdous
Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Chairperson:
(Chair)

Sadia Hamid Kazi, Ph.D.
Associate Professor
Department of Computer Science and Engineering
School of Data and Sciences
BRAC University

Ethics Statement

This study utilizes a multimodal e-commerce review dataset constructed exclusively from publicly available product pages and user-generated content on the Daraz platform. All data were collected in accordance with the platform's terms of service and restricted to information that is openly accessible to the public. The study does not involve human subjects, clinical data, or any form of direct interaction with individuals.

To protect user privacy, personally identifiable information (e.g., usernames, profile details, or contact information) was removed or anonymized during preprocessing. The dataset contains only review text, product-related images, and associated metadata necessary for research purposes. No attempts were made to infer or disclose sensitive personal attributes.

The dataset is released solely for academic and research use under the **Creative Commons CC BY-NC-ND 4.0 (Attribution–NonCommercial–NoDerivatives)** license, which permits sharing with attribution for non-commercial purposes while prohibiting redistribution of modified versions. Researchers are expected to use the data responsibly, respect user privacy, and avoid any misuse or re-identification attempts. Additionally, models developed using this dataset are intended for research and decision-support applications in e-commerce analytics and should not be deployed in ways that could unfairly impact users or consumers without proper evaluation and oversight.

Abstract

Ordinal sentiment classification from customer reviews using star rating labels is challenging due to subtle distinctions between adjacent rating levels, particularly in low-resource and multilingual settings. Existing benchmarks are predominantly English-centric and text-only, limiting their ability to reflect real-world e-commerce scenarios. We introduce DarazMM, a large-scale Bangla-centric multilingual multimodal dataset for five-class star rating prediction, covering Bangla, English, code-mixed Bangla-English and Romanized Bangla. The dataset is released in two settings: a text-only corpus of 257,817 reviews and a multimodal subset of 25,452 reviews that includes review text, customer-uploaded images, advertised product images and corresponding product descriptions. We benchmark zero-shot, few-shot and supervised models, and propose two neuro-symbolic frameworks that model rating ordinality and cross-modal interactions. Our best supervised framework achieves 0.87 accuracy and 0.83 macro-F1, while the best zero-shot and few-shot models reach 0.71 and 0.79 accuracy, respectively. The dataset will be made publicly available.

Keywords: Multimodal Sentiment Analysis; Multilingual and code-mixed NLP; Neuro-symbolic AI; E-commerce product reviews

Dedication

This work is for my wonderful parents, whose unwavering support has been the foundation of my journey; for my teachers and mentors, who have inspired me to pursue knowledge with passion and integrity; and for everyone working to advance research in multilingual and multimodal natural language processing. I hope that this work contributes to a better understanding and modeling of customer reviews.

Acknowledgement

Firstly, all praise to the Almighty Allah, through whom my thesis has been completed without major interruption.
I am deeply grateful to my distinguished supervisor, **Dr. Md. Golam Rabiul Alam**, Professor in BRAC University's Department of Computer Science and Engineering, for his guidance and support throughout this thesis. His unwavering guidance helped me stay focused and on course.
I also thank BRAC University for providing the tools and resources necessary to carry out this research.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

Samples Number of review instances in a dataset

(t) Text-only input setting

(t, i) Joint text–image input setting

$\mathcal{L}_{\text{CE}}$ Categorical Cross-Entropy loss

$\mathcal{L}_{\text{InfoNCE}}$ Bidirectional InfoNCE loss

$\mathcal{L}_{\text{MSE}}$ Feature-space Mean Squared Error loss

Banglish Code-mixed Bangla–English text

BERT Bidirectional Encoder Representations from Transformers

BERT-KAN Hybrid model combining BERT with Kolmogorov–Arnold Networks

BiLSTM Bidirectional Long Short-Term Memory

CE Cross-Entropy loss

CLIP Contrastive Language–Image Pretraining model

CNN Convolutional Neural Network

DL Deep Learning

Gemma-3n Gemma-3n-E4B-it model

GPT-OSS GPT-OSS-20B model

LLaMA-3.2 Llama-3.2-11B-Vision-Instruct model

LLM Large Language Model

ML Machine Learning

NLP Natural Language Processing

POS Part-of-Speech

*Qwen*3 Qwen3-VL-8B-Instruct model

$SMOTE$ Synthetic Minority Over-sampling Technique

SVM Support Vector Machine

$TF\text{-}IDF$ Term Frequency–Inverse Document Frequency

ViT Vision Transformer

VLM Vision-Language Model

$XLMR$ + CLIP XLM-Roberta features concatenated with CLIP ViT-B/16 image embeddings

$XLMR - hybrid$ Multimodal XLM-Roberta + ViT model with hybrid loss (CE + InfoNCE + MSE)

$XLMR - two - stage$ Two-stage hierarchical XLM-Roberta classifier

Chapter 1

Introduction

Customer-written reviews are a primary source for understanding consumer preferences in online marketplaces. They motivate research in computational linguistics and machine learning for tasks such as recommendation, product ranking and market analysis [1]. On most e-commerce platforms, reviews include free-form text, a star rating and sometimes an optional product image [2, 3]. These heterogeneous signals provide complementary information about user satisfaction and purchasing behavior. As a result, review data serve as an important bridge between qualitative opinions and quantitative decision-making. Among opinion mining problems, star rating prediction aims to infer ordinal rating categories from review content [3]. This task is more challenging than binary sentiment classification because adjacent rating levels often differ only subtly [4]. Small linguistic nuances, contextual cues and subjective interpretations can therefore significantly affect prediction accuracy. Recent advances in deep learning and large language models (LLMs) have improved sentiment and rating prediction performance [5, 6]. Modern neural architectures can automatically learn semantic and contextual representations without extensive feature engineering, enabling more robust modeling of complex review patterns. Pretrained models further enhance generalization by transferring knowledge from large corpora to downstream tasks. However, most progress relies on English-centric benchmarks with relatively clean and homogeneous data [7, 8]. In Bangladesh, e-commerce platforms are widely used and customers predominantly write reviews in Bangla or code-mixed Bangla–English, often including Romanized text [9]. These multilingual and informal reviews are difficult to process because Bangla is a low-resource language with limited large-scale annotated corpora [10,11]. Informal spelling variations, non-standard grammar and frequent code-switching further increase linguistic ambiguity. Existing benchmarks therefore fail to reflect the linguistic variability and multimodal characteristics of Bangladeshi e-commerce reviews [12].

Several Bangla-centric datasets have been proposed to support sentiment analysis and rating prediction. Most of them were collected from platforms such as Daraz [13] and Pickaboo [14]. These datasets provide an initial foundation for developing localized language technologies and benchmarking models in realistic settings. They also highlight the growing research interest in addressing underrepresented languages in opinion mining. Early datasets were small and text-only. They were often manually annotated to ensure label quality, but they lacked linguistic and contextual diversity [4,15,16]. More recent resources expanded language coverage to include Banglish

and code-mixed text [17]. Large-scale datasets such as BanglishRev increased size and metadata richness and included customer-uploaded review images [12]. However, none of these datasets provide advertised product images together with product descriptions and review content. This limitation restricts the analysis of expectation mismatch between advertised products and received items. Consequently, important visual cues that influence customer satisfaction remain underexplored in current benchmarks.

1.1 Motivation

Real-world e-commerce reviews contain multiple information sources. Customers express opinions through text, assign star ratings, and often upload product images. These signals are complementary and may improve rating prediction when modeled jointly. Text explains user experience in detail, ratings provide structured supervision and images offer visual evidence of product condition or defects. Considering only one modality may therefore ignore useful contextual cues that influence customer judgment. Integrating these heterogeneous sources allows models to capture both subjective impressions and objective evidence. Such holistic modeling better reflects how humans naturally evaluate products and make purchasing decisions.

In Bangladeshi marketplaces, reviews are frequently multilingual and informal. Users switch between Bangla and English and often rely on Romanized Bangla [9]. Such variation increases vocabulary sparsity and introduces inconsistent spelling. These characteristics reduce the effectiveness of models trained on clean English datasets [10, 11]. They also complicate preprocessing, tokenization and feature extraction. Moreover, informal expressions, abbreviations and phonetic spellings create additional noise in the data. Addressing these challenges requires approaches that are robust to linguistic variability and domain-specific writing styles.

Existing resources do not adequately combine linguistic diversity with multimodal information. Many datasets are text-only, while others include limited visual content. As a result, they do not fully represent the structure of reviews collected from platforms such as Daraz and Pickaboo. This gap motivates the development of a realistic multilingual and multimodal dataset for star rating prediction that better reflects practical e-commerce scenarios. A comprehensive resource can facilitate fair comparison across methods and encourage more generalizable solutions. It can also support future research on cross-modal reasoning and low-resource language understanding.

1.2 Problem Statement

Star rating prediction from reviews is an ordinal classification problem, with ratings mapped to ordered sentiment categories from Very Negative (1) to Very Positive (5) [18]. The model must distinguish fine-grained differences between adjacent rating levels. Small variations in wording may correspond to different ratings. For example, moderate satisfaction and strong satisfaction may both appear positive but belong to separate categories. This requirement makes the decision boundary between classes less distinct than in binary sentiment analysis [4]. Consequently, prediction errors often occur between neighboring classes rather than distant ones. Effective models

must therefore capture subtle semantic shifts and contextual cues to make reliable distinctions.
The task becomes more difficult in low-resource multilingual settings where Bangla, English, code-mixed and Romanized text appear together. Standard language models often assume consistent scripts and grammar. These assumptions do not hold for informal user-generated content [10, 11]. Limited annotated corpora further restrict the training of robust models. Consequently, performance may degrade when models are transferred from English-centric benchmarks. Domain mismatch and vocabulary gaps can lead to unstable predictions and reduced generalization. Specialized preprocessing and adaptation strategies are therefore necessary to maintain model reliability.
Existing Bangla-centric datasets are either text-only or lack complete product-aware modalities such as advertised images and descriptions. Without these inputs, models cannot analyze cross-modal consistency or expectation mismatch between advertised and received products. They also cannot exploit visual cues that may explain low or high ratings. Therefore, current approaches do not fully capture the structure of real-world reviews. Important evidence that influences customer perception may remain unused during training and inference. A comprehensive dataset and suitable modeling framework are required to address these limitations.

1.3 Research Objectives

Based on the motivation and identified research gaps, this study defines a set of objectives that guide the construction of the dataset and the development of modeling approaches. These objectives aim to establish a realistic benchmark and enable consistent evaluation under multiple learning settings. By clarifying these goals, the study provides a structured roadmap for step-by-step experimentation and comparison. Specifically, the objectives are summarized as follows:

- Develop a large-scale Bangla-centric multilingual multimodal dataset for five-class star rating prediction.
- Provide two complementary settings: a text-only corpus and a multimodal subset that includes review text, customer-uploaded image, advertised product images and product descriptions.
- Benchmark zero-shot, few-shot and supervised models across multilingual and multimodal configurations.
- Design neuro-symbolic frameworks that explicitly model rating ordinality and cross-modal interactions.

1.4 Contributions

This work contributes both new data resources and methodological advances for star rating prediction. The contributions focus on realistic dataset design, comprehensive evaluation, and structured modeling of ordinal and multimodal information. The main contributions of this thesis are summarized as follows:

- We introduce a large-scale multilingual e-commerce review dataset for five-class star rating prediction, covering Bangla, English, code-mixed Bangla–English and Romanized Bangla. The dataset is released in two settings: a text-only corpus of 257,817 reviews and a multimodal subset of 25,452 reviews.
- We provide comprehensive baseline evaluations under zero-shot, few-shot and supervised learning paradigms across multilingual and multimodal settings.
- We propose two neuro-symbolic rating prediction frameworks that explicitly model rating ordinality and cross-modal interactions. The proposed frameworks consistently outperform strong baselines on the dataset.

1.5 Organization of the Report

The remainder of this thesis is organized as follows. Chapter 2 reviews the related literature. Chapter 3 describes the experimental methodology. Chapter 4 presents and analyzes the results. Chapter 5 concludes the study, discusses limitations and outlines directions for future work.

Chapter 2

Related Work

Early research on Bangla e-commerce reviews primarily relied on relatively small, text-only datasets. These early efforts were largely motivated by the rapid growth of online retail platforms in Bangladesh, particularly during the COVID-19 pandemic, when consumers increasingly depended on digital marketplaces. As online transactions surged, concerns regarding fraudulent sellers, defective products and misleading information also grew, creating a strong need for automated sentiment analysis systems to monitor customer feedback and detect potential scams. Consequently, researchers focused on building compact but practical corpora that could support classical Natural Language Processing (NLP) pipelines and traditional machine learning approaches. One of the earliest contributions was made by Shafin et al. [19], who compiled over 1,000 customer comments from popular local platforms to examine positive and negative sentiment distributions. Through comparative experiments with multiple classifiers, they found that Support Vector Machines (SVM) offered the most stable performance, achieving an accuracy of 88.81%. Expanding this line of work, Hossain et al. [4] curated a larger Daraz dataset containing 7,874 unique reviews enriched with metadata such as usernames, product categories and timestamps, and framed the task as five-class star rating prediction using TF-IDF features with SVM. To address the heavy dominance of 5-star reviews, they applied random oversampling to expand the dataset to 21,875 samples, ultimately improving accuracy to 90.44%. Similarly, Akter et al. [15] released manually annotated datasets organized by product sub-categories, while Chowdhury et al. [20] constructed a balanced corpus of 13,550 reviews with verified labels and reported strong results for SVM with unigram features (78.63%), with hybrid CNN–BiLSTM models achieving competitive performance (75.25%).

As the field matured, researchers began moving beyond small-scale corpora to develop larger and more linguistically diverse datasets capable of capturing the complex emotional expressions of Bangladeshi consumers. These later efforts recognized that simple binary sentiment labels were insufficient to reflect the nuanced opinions present in real-world reviews, prompting the inclusion of multilingual content and richer affective annotations. By incorporating broader coverage in both scale and language, these datasets enabled more robust and generalizable modeling approaches. For instance, Rashid et al. [12] introduced a large-scale collection of 78,130 reviews gathered from Daraz and Pickaboo across 152 product categories, combining 28,912 Bangla and 49,218 English reviews. Beyond sentiment polarity, they provided expert-validated annotations for five emotion classes—Happiness, Sadness,

Fear, Anger and Love—following Shaver's psychological framework, thereby offering deeper cultural insight. Hossain et al. [17] further extended this direction by incorporating Bangla, English and Romanized Bangla (Banglish) text with five-level sentiment labels to better support low-resource multilingual settings. Zeng et al. [21] demonstrated the effectiveness of few-shot synthetic data augmentation in code-mixed Spanish-English and Malayalam-English reviews, using GPT-4 to generate naturalistic synthetic sentences that improved F1 scores by 9.32% in low-resource scenarios. Outside the Bangla context, several large-scale review datasets have served as comparative benchmarks [22–25]. Notably, Fang et al. [22] analyzed 5.1 million Amazon reviews and proposed a POS-based sentiment scoring framework that handled linguistic phenomena such as negation, combining it with classifiers like Naïve Bayes, Random Forest and SVM to achieve strong sentence-level and document-level results. Barbieri et al. [26] contributed XLM-T and the Unified Multilingual Sentiment Analysis Benchmark (UMSAB), showing that massive multilingual pre-training on social media data can significantly improve zero-shot cross-lingual transfer for low-resource languages.

The transition from text-only analysis to multimodal learning marked another important milestone for Bangla review research. While earlier datasets relied solely on written feedback, real-world e-commerce interactions frequently include complementary signals such as product images and user engagement metadata. Recognizing this gap, researchers began constructing multimodal corpora to better reflect authentic consumer behavior and to enable models that leverage both textual and visual information for improved sentiment understanding. A prominent example is BanglishRev [27], the first large-scale multimodal Bangla e-commerce dataset, containing 1.74 million reviews derived from 3.2 million ratings. In addition to text, the dataset includes customer-uploaded images and extensive metadata such as purchase timestamps, likes/dislikes and seller responses, providing a richer contextual foundation for analysis. Linguistically, it mirrors regional habits, with 37.7% Banglish and 5.9% code-mixed content. Given the impracticality of manual annotation at this scale, the authors adopted numerical ratings as proxy sentiment labels and trained a BanglishBERT model, demonstrating the effectiveness of large-scale pretraining by achieving 94% accuracy and an F1-score of 0.94 on manually labeled benchmarks. Ling et al. [28] proposed VLP-MABSA, a task-specific vision-language pretraining framework using BART to jointly model textual and visual aspect-opinion alignments, achieving significant gains on multimodal sentiment benchmarks, particularly in low-resource settings. Zhang et al. [29] introduced the Adaptive Language-guided Multimodal Transformer (ALMT) to suppress sentiment-irrelevant information from auxiliary modalities, improving stability and convergence across MOSI, MOSEI and CH-SIMS datasets.

Beyond dataset construction, methodological improvements have also focused on addressing common challenges such as class imbalance and noisy annotations. E-commerce reviews often exhibit highly skewed rating distributions, with positive feedback dominating, which can bias classifiers and reduce generalization performance. To counteract this issue, researchers have explored data balancing strategies and more sophisticated architectures capable of capturing the syntactic and semantic complexity of multilingual and code-mixed text. Zaman et al. [1] systematically evaluated the Synthetic Minority Over-sampling Technique (SMOTE), generating synthetic minority samples through a k-nearest neighbor strategy to equalize

class distributions. Their experiments showed a substantial boost in SVM accuracy from 63.76% to 92.80%, highlighting the effectiveness of balanced training data. In parallel, Rashid et al. [7] proposed BERT-KAN, a hybrid architecture combining BERT's contextual embeddings with Kolmogorov–Arnold Networks that employ polynomial expansions and polynomial attention to model complex non-linear relationships. This fusion of deep transformer representations with functional approximation mechanisms achieved state-of-the-art results, including 95.3% precision and a 96.1% F1-score. Similarly, Koto et al. [30] leveraged multilingual sentiment lexicons to enhance low-resource language modeling without requiring sentence-level supervision, outperforming large models like GPT-3.5 and BLOOMZ in code-mixed and zero-shot settings. Quoc et al. [31] introduced ViCloABSA for Vietnamese clothing reviews, providing a rigorous benchmark for aspect-based sentiment analysis and demonstrating that LLM-based models consistently outperform smaller baselines as training data increases. Zhang et al. [32] presented SENTIEVAL, highlighting where LLMs excel in few-shot scenarios and where specialized models remain superior in structured output tasks.

Multimodal sentiment analysis (MSA) surveys [33] emphasize that while text features often dominate, integrating textual, visual, and acoustic cues is essential for capturing deeper emotional polarity. These surveys highlight the evolution of fusion techniques, categorizing them into early feature-level fusion, medium-term model-based fusion, and late-stage decision fusion. They identify critical remaining challenges, such as the detection of hidden emotions like sarcasm, and the lack of diverse, large-scale multilingual datasets that include physiological signals like brain waves or pulses. Looking forward, researchers advocate for the development of unified, large-scale architectures capable of handling hidden emotions and multilingual content to bridge the current gap between human and artificial intelligence.

Collectively, these studies illustrate several clear trends in sentiment analysis research. Early efforts focused on building small, text-only corpora and applying classical machine learning models, gradually expanding toward larger, multilingual datasets that capture nuanced emotional expressions. Concurrently, the field has embraced multimodal learning, integrating textual, visual and auxiliary signals to reflect realistic e-commerce interactions. Methodological innovations, including data balancing, lexicon-based pretraining and hybrid transformer architectures, have further improved performance, particularly in low-resource and code-mixed settings. Despite these advances, existing Bangla datasets either emphasize text-only reviews or incorporate limited multimodal signals, leaving a gap in comprehensive resources that combine diverse modalities across languages.

To the best of our knowledge, no prior Bangla-centric dataset combines review text, customer-uploaded images, advertised product images and product descriptions [4, 12, 17, 27]. Our dataset fills this gap, supporting five-class star rating prediction in multilingual and multimodal settings.

Table 2.1 summarizes the main approaches and methods reported in the literature, highlighting the datasets, techniques and trends relevant to this study.

Table 2.1: Overview of Existing Approaches, Datasets and Methods for Sentiment and Star Rating Prediction

Aspect	**Details**
Datasets	Manually labeled datasets with metadata [15]; balanced corpus of 13,550 Bangla reviews [20]; 78,130 Bangla-English reviews with emotion labels [12]; 5.1M Amazon reviews [22]; BanglishRev multimodal dataset with 1.74M reviews [27].
Sentiment analysis in Non-Bangla Languages	Large-scale modeling in various domains [23–25]; POS-based sentiment scoring and negation handling on Amazon reviews [22].
Sentiment analysis/star rating prediction in Bangla	SVM applied to small datasets [19]; five-class rating prediction using TF-IDF and SVM with random oversampling [4]; performance comparison on balanced corpora [20]; rating prediction using SMOTE for class balancing [1].
Multilingual sentiment analysis	Analysis involving Bangla, English and Romanized Bangla (Banglish) [12,17]; hybrid BERT-KAN architecture for code-mixed text [7].
Multimodal sentiment analysis/star rating prediction	Integration of review images and metadata in BanglishRev [27]; surveys emphasizing the integration of textual and visual cues [33].
Machine Learning Approaches	SVM for sentiment and rating prediction [1, 4, 19, 20]; Naïve Bayes and Random Forest for review-level classification [22].
Deep Learning Approaches	Hybrid CNN–BiLSTM models [20]; large-scale pretraining with BanglishBERT [27]; hybrid BERT-KAN combining transformer depth with Kolmogorov–Arnold Networks [7].
Challenges and Trends	Managing imbalanced rating distributions [1, 4]; addressing linguistic complexity in code-mixed/Romanized text [7, 17]; handling multimodal inputs and hidden emotions like sarcasm [33].

Chapter 3

Methodology

Our methodology, illustrated in Fig 3.1, covers the collection and preparation of the Daraz Multimodal Multilingual Dataset (DarazMM), as well as the training and evaluation of various text-only and multimodal models. We collected product information, reviews and associated images, and organized them into a structured dataset. The dataset was curated to ensure quality and appropriate validation and test splits were created. Models were then evaluated under multiple paradigms, including zero-shot, few-shot and supervised fine-tuning, with experiments designed to assess both textual and visual information for star rating prediction. Detailed procedures for data processing, augmentation, model architectures and evaluation are provided in the following sections. In particular, our text-only Daraz Multilingual (DarazM) dataset is described in detail in Section 3.5.

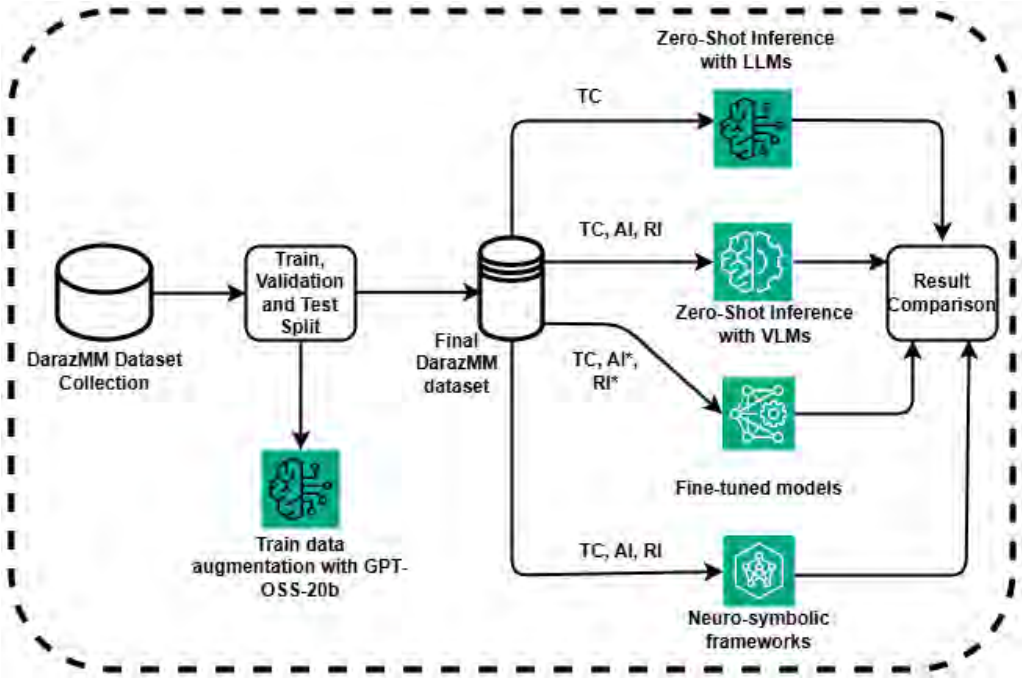


Figure 3.1: Overall workflow of the methodology. Abbreviations: TC, Text Comment; AI, Advertised Image; RI, Reviewer/Customer-uploaded Image. An asterisk (*) indicates that only a subset of models utilized the corresponding modality.

3.1 Data Collection

The dataset was collected between April 2025 and September 2025 using an automated web scraping pipeline, and covers reviews posted between 2019 and 2025 from the Daraz Bangladesh e-commerce platform [13]. We employed Selenium with a Chrome WebDriver [34] to handle dynamically loaded content and paginated product and review pages.

We first crawled product listing pages and retained products with more than ten customer reviews to ensure sufficient opinion diversity. Each retained product was assigned a unique internal Product ID, resulting in a total of 4,109 distinct products in the final corpus. For each selected product, we collected the advertised product name, product description and product images. We then iterated through all available review pages and extracted customer-written reviews along with star ratings, review text, review dates, customer identifiers and customer-uploaded review images when available. Since a single review may contain multiple user-uploaded images, we store only the first image for consistency while recording the total number of uploaded images as metadata. Each review instance in the final DarazMM dataset contains the following fields: Product ID, Product Name, Product Description, Review Comment, Rating, Review Date, Reviewer Number and the total number of uploaded images, together with the corresponding advertised product image and the first customer-uploaded review image. Reviewers refer to customers who have written the product reviews. Reviewer Numbers are anonymized internal indices and do not correspond to real user accounts or identities; they are used solely to link review text with associated images.

Both product-level information and review-level data were stored in structured CSV format. Images were downloaded and organized by Product ID to preserve alignment between reviews, advertised product images and descriptions. From this corpus, we construct two datasets: **DarazM**, a large-scale text-only review dataset containing all reviews, and **DarazMM**, a multimodal subset containing only reviews paired with customer-uploaded images. Together, DarazM and DarazMM constitute a large-scale multilingual review corpus containing Bangla, English, Romanized Bangla and code-mixed comments, supporting both text-only and multimodal learning scenarios. Table 3.1 shows several representative review instances from DarazMM.

Table 3.1: Example review instances from the DarazMM dataset.

Product ID	**Review Comment**	**Reviewer No.**	**Rating**
PROD_0016	darun hoyeche! satisfied! [It's excellent! satisfied!]	82	1
PROD_0003	খুব ভালো ছিল ধন্যবাদ [It was very good, thanks.]	7	5
PROD_0407	What I wanted & what I received!	22	1

Note: For brevity, only selected fields are presented.

Table 3.2 compares DarazM and DarazMM with existing Bangla-centric e-commerce review datasets in terms of dataset size and available features, including review images, metadata and product-level information such as product name, description and advertised images.

Table 3.3 compares the language coverage of our datasets with prior work, indicating

Table 3.2: Comparison of Bangla-centric e-commerce review datasets in terms of number of samples and features.

Dataset	**# Samples**	**Review Images**	**Review Date**	**Product Name**	**Product Desc.**	**Ad. Images**
Hossain et al. [4]	7874	No	Yes	Yes	No	No
Rashid et al. [12]	78130	No	No	Yes	No	No
Akter et al. [15]	7905	No	No	Yes	No	No
Hossain et al. [17]	7181	No	No	No	No	No
Shanto et al. [16]	1000	No	No	No	No	No
Shamael et al. [27]	1.74M	Partial	Yes	Yes	No	No
Ours (DarazM)	257817	No	Yes	Yes	Yes	No
Ours (DarazMM)	25452	Yes	Yes	Yes	Yes	Yes

Note: Ad. Images refer to official advertised images provided by sellers, distinct from customer-uploaded images. Product Desc. indicates the description of the product. "Partial" indicates that only a subset of reviews contains images.

whether reviews are available in Bangla, English, code-mixed Bangla–English or Romanized Bangla.

Table 3.3: Language coverage of Bangla-centric e-commerce review datasets.

Dataset	**Bangla**	**English**	**Code-mixed**	**Romanized**
Hossain et al. [4]	Yes	No	No	No
Rashid et al. [12]	Yes	Yes	No	No
Akter et al. [15]	Yes	No	No	No
Hossain et al. [17]	Yes	Yes	No	Yes
Shanto et al. [16]	Yes	No	No	No
Shamael et al. [27]	Yes	Yes	Yes	Yes
Ours (DarazM)	Yes	Yes	Yes	Yes
Ours (DarazMM)	Yes	Yes	Yes	Yes

Note: "Code-mixed" refers to reviews containing a mixture of Bangla and English, while "Romanized" indicates Bangla written using Latin script.

3.2 Data Splitting Strategy

The DarazMM dataset was divided into training, validation and test splits using a class-balanced strategy to ensure fair evaluation across rating categories. Since the rating distribution is highly skewed toward the 5-star class, a uniform split would cause the test set to be dominated by majority-class samples and under-represent minority ratings. Therefore, we performed class-wise splitting.

For ratings 1–4, approximately 50% of the samples from each class were reserved for testing to guarantee sufficient evaluation coverage of minority classes. The remaining samples were used for model development. From this remaining portion, 250 samples were randomly selected to form the validation set, and the rest were used for training. In contrast, rating 5 contains substantially more samples than the other classes. Retaining 50% of this class for testing would disproportionately inflate the test set and bias evaluation metrics. Hence, a smaller, proportionally suitable test subset was selected for rating 5 to maintain a balanced test distribution while preserving most samples for training.

Overall, the final test set comprises 4,105 samples across all rating categories. The resulting split sizes for each rating class are illustrated in Fig 3.2.

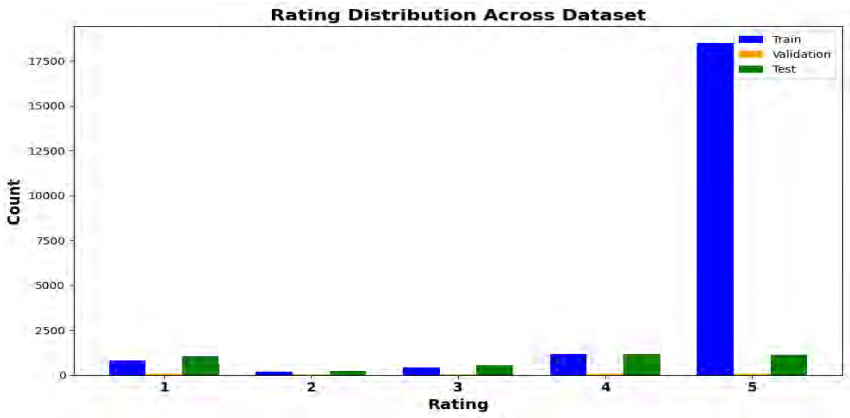


Figure 3.2: Distribution of samples across rating categories in the training, validation and test splits.

3.3 Data Augmentation

To address the class imbalance in the training split of the DarazMM dataset, we employed a controlled data augmentation strategy using GPT-OSS. The augmentation was applied exclusively to the training data and was used only for models that were fine-tuned in a supervised setting.

Specifically, the majority class (rating 5) was downsampled by retaining only a subset of its original samples, which were used directly for training without augmentation. All samples belonging to the remaining rating categories (ratings 1–4) were augmented via GPT-OSS to increase their representation and reduce skew. The augmentation process generated paraphrased, translated and semantically consistent review comments that preserve the original rating polarity and intent. The prompt template illustrated in Fig 3.3 was used with GPT-OSS to generate synthetic Romanized Bangla reviews for minority rating classes during training data

augmentation. Similar prompts were used to generate synthetic reviews in English and Bangla.

```
<|start|>system<|message|>
You are ChatGPT, a large language model trained by OpenAI.
Knowledge cutoff: 2024-06
Current date: 2025-10-06
Reasoning: medium
# Valid channels: analysis, commentary, final. Channel must be included for every message.
<|end|>

<|start|>user<|message|>
You are a helpful assistant that generates synthetic Romanized Bengali product reviews for a
sentiment classification dataset.

The star rating scale is:
1 = khub kharap obhiggota [very bad experience]
2 = kharap [bad]
3 = motamuti [okay / average]
4 = bhalo [good]
5 = khub bhalo obhiggota [very good experience]

Example of transliteration style:
"খুব বাজে" → "khub baje" [very bad]
"একদম ভালো না" → "ekdom bhalo na" [not good at all]
"ডেলিভারি দেরিতে এসেছে" → "delivery derite esheche" [delivery arrived late]

Here is one example review for a product (in Bengali):
"একেবারে ভালো না" [not good at all]

Now generate ONE new Romanized Bengali product review that expresses a similar sentiment, tone,
and style as the example above.

Rules:
- The review must be written fully in Romanized Bengali (no Bengali script).
- It should sound natural and realistic.
- Do not repeat or closely paraphrase the original.
- Do not include any translation, explanation, or label - only the review text itself.
<|end|>
```

Figure 3.3: Example prompt used for GPT-OSS-based synthetic review generation during data augmentation. English glosses shown in gray are provided for reader clarity and are *not* part of the actual prompt.

Fig 3.4 illustrates the resulting distribution of samples across rating categories after augmentation.

3.4 Exploratory Data Analysis

To better understand the characteristics of the DarazMM dataset, we conducted an exploratory data analysis using a set of quantitative visualizations. These visualizations summarize the distribution of ratings, review length statistics, product-level review density, temporal trends and lexical patterns in customer comments.

Rating Distribution. Fig 3.2 shows the distribution of star ratings across all reviews. The dataset is highly skewed toward positive feedback, with 5-star ratings being the most frequent, followed by 4-star, 1-star, 3-star and 2-star ratings (i.e., $5 > 4 > 1 > 3 > 2$).

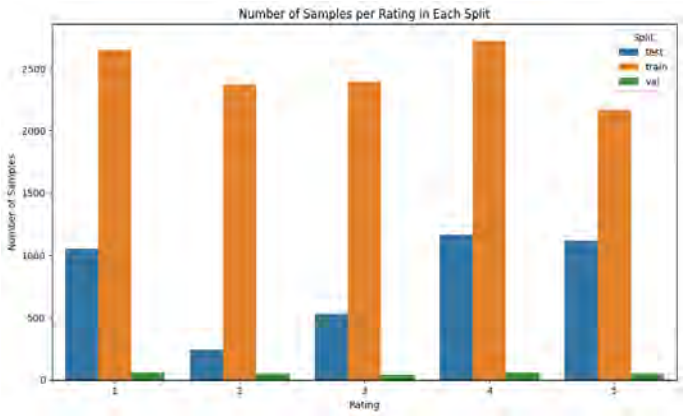


Figure 3.4: Distribution of samples across rating categories in the training, validation and test splits after data augmentation.

Comment Length Analysis. Fig 3.5 presents the density distribution of word counts per comment. A large portion of the reviews are extremely concise, with most comments containing only 2–3 words. At the same time, the dataset also includes a small number of long-form reviews, with the maximum comment length exceeding 150 words, indicating substantial variability in customer engagement.

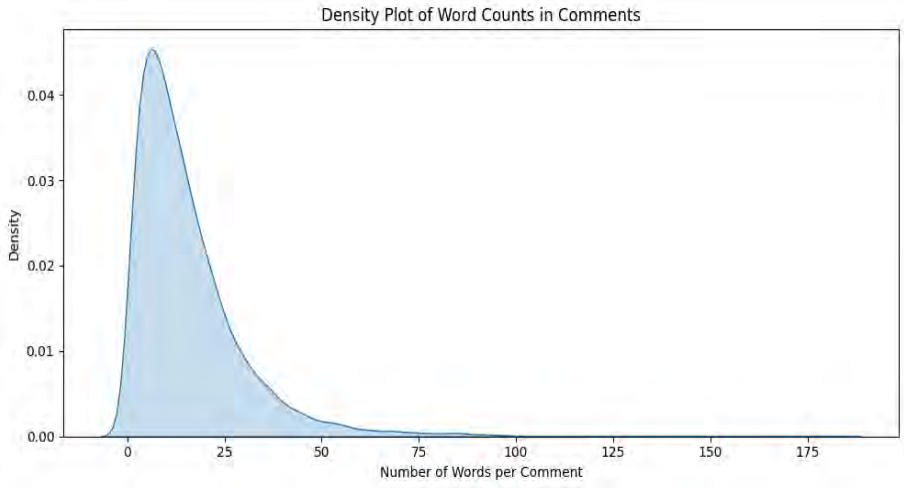


Figure 3.5: Density plot of the number of words per comment in the DarazMM dataset.

Reviews per Product. Fig 3.6 illustrates the density of the number of reviews per product. The majority of products receive between 10 and 20 reviews, which aligns with the dataset construction strategy that enforces a minimum number of reviews per product while preserving natural variation in review frequency.

Temporal Trends. The evolution of review activity over time is shown in Fig 3.7. We observe a steady increase in the number of reviews from 2019 to 2023, suggesting growing customer engagement on the platform. The peak review volume occurs in November 2023, after which the activity slightly tapers off.

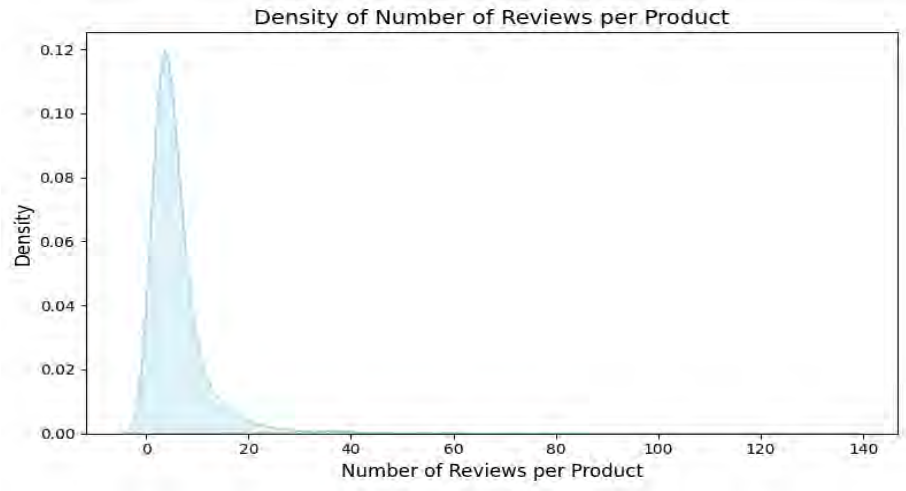


Figure 3.6: Density of the number of reviews per product in the DarazMM dataset.

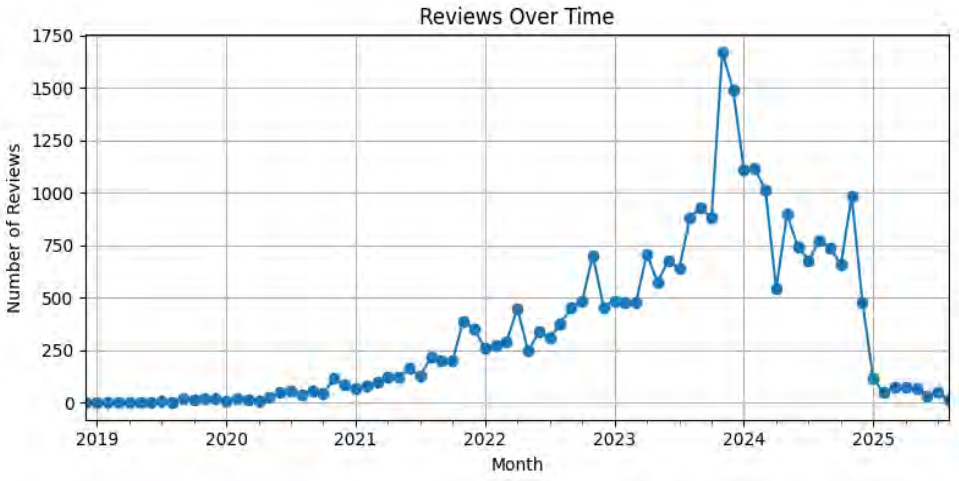


Figure 3.7: Monthly distribution of reviews over time in the DarazMM dataset.

Lexical Patterns and Sentiment. Fig 3.8 presents word clouds generated from English (including Romanized Bangla) and Bangla comments. In both cases, words expressing positive sentiment are highly prevalent. The English word cloud (left) contains frequent tokens such as "good product", "recommended" and "thank", as well as Romanized Bangla phrases like "nite paren" (can buy / recommended). Similarly, the Bangla word cloud (right) is dominated by expressions such as অনেক ভালো (very good), অনেক সুন্দর (very nice), ঠিক আছে (okay), and নিতে পারেন (can buy), further reinforcing the overall positive tone of the dataset.

3.5 Daraz Multilingual Dataset (DarazM)

DarazM is a large-scale **text-only source corpus** comprising 257,817 product reviews collected from the Daraz platform. The multimodal dataset used for experimentation, **DarazMM**, is constructed as a subset of DarazM by selecting only those reviews that contain both textual comments and reviewer-uploaded images. Consequently, DarazMM inherits the linguistic, temporal and rating characteristics

Figure 3.8: Word clouds for English/Romanized Bangla (left) and Bangla (right) comments.

of DarazM while adding visual information.

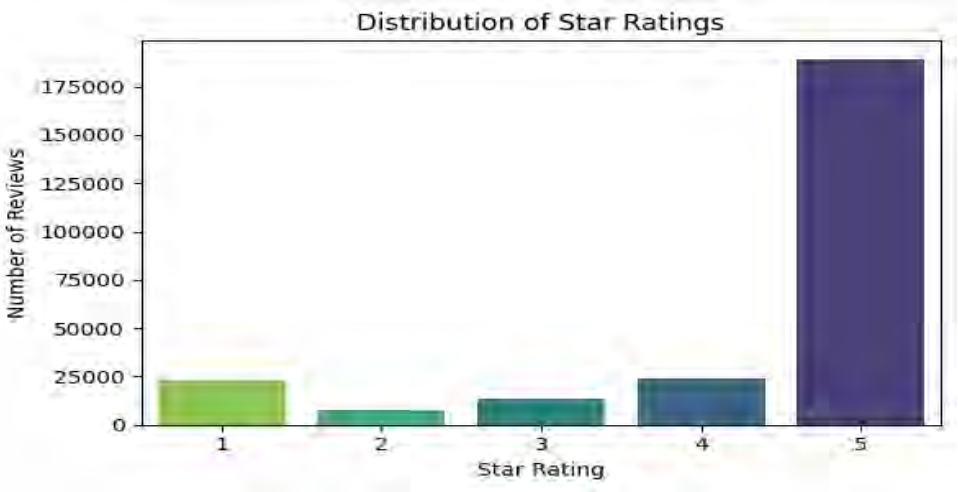


Figure 3.9: Distribution of star ratings in the DarazM (text-only parent) dataset.

Fig 3.9 shows the overall rating distribution of DarazM, which exhibits the same skew toward higher ratings observed in DarazMM. Fig 3.10 illustrates the distribution of word counts per comment. Most comments consist of between 2 and 50 words. Fig 3.11 presents the distribution of reviews per product. Fig 3.12 shows the monthly review frequency. The number of reviews peaks in late 2023, similar to the DarazMM dataset.

3.6 Problem Formulation

Let $\mathcal{D} = \{(c_i, I_{ad,i}, I_{rev,i}, y_i)\}_{i=1}^{N}$ denote the DarazMM dataset, where c_i represents the customer review text, $I_{ad,i}$ denotes the advertised product image, $I_{rev,i}$ denotes the customer-uploaded image and $y_i \in \{1, 2, 3, 4, 5\}$ is the ground-truth star rating. All instances in DarazMM contain both textual and visual information. However, different modeling paradigms utilize different subsets of the available modalities. We define the full input tuple as

$$x_i = (c_i, I_{ad,i}, I_{rev,i}).$$

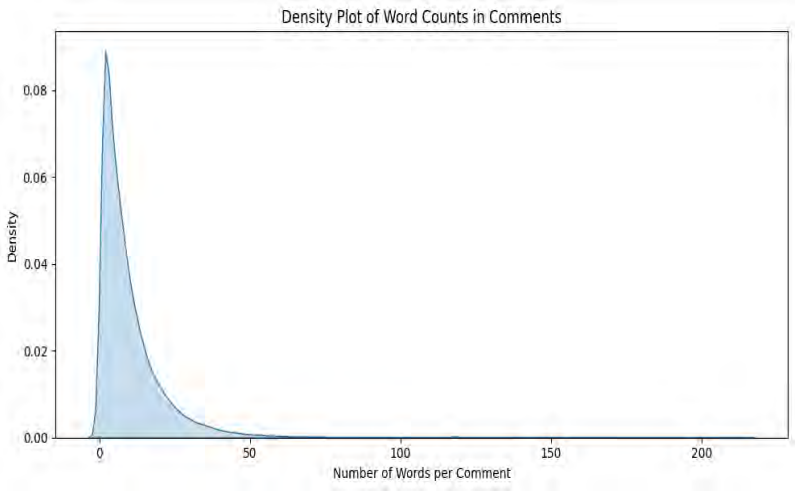


Figure 3.10: Density plot of the number of words per comment in the DarazM dataset.

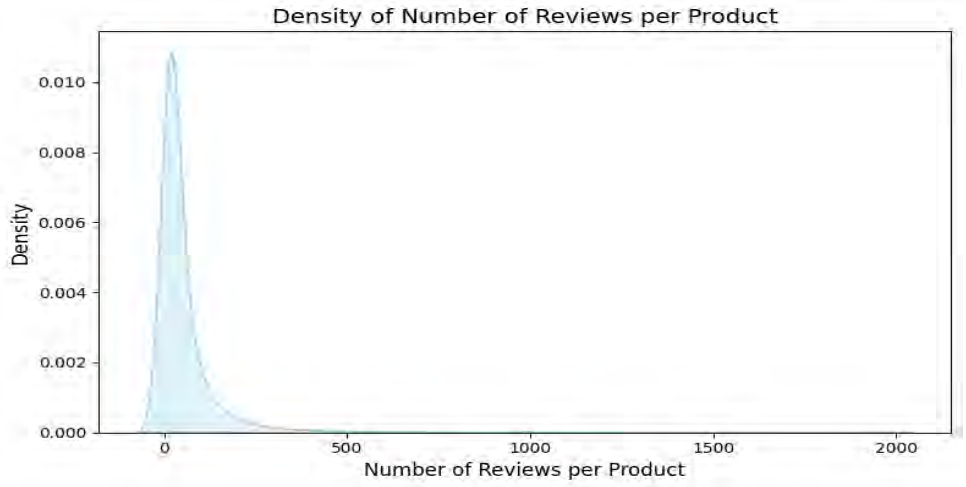


Figure 3.11: Density of the number of reviews per product in the DarazM dataset.

3.6.1 Modality Utilization Settings

Text-Only Models. Text-only models are defined on the reduced input space

$$f_\theta^{text} : \mathcal{C} \to \mathcal{Y},$$

where only the textual component c_i is provided to the model and $\mathcal{Y} = \{1, 2, 3, 4, 5\}$ denotes the set of possible ratings.
Text-only modeling can equivalently be interpreted as applying a coordinate selection operator

$$\pi_{\text{text}}(x_i) = c_i,$$

which discards the visual components before prediction. Although visual information exists in the dataset, it is not accessible to the model in this setting.

Multimodal Models. Multimodal models jointly leverage textual and visual evidence by operating on the complete input tuple:

$$f_\theta^{multi}(x_i) = f_\theta^{multi}(c_i, I_{ad,i}, I_{rev,i}).$$

Figure 3.12: Monthly distribution of reviews over time in the DarazM dataset.

These models learn representations that integrate semantic sentiment cues from text with visual consistency signals derived from advertised and customer-uploaded images.

3.6.2 Supervised Learning Objective

For fine-tuned models, star rating prediction is formulated as a 5-class classification problem. Let

$$\tilde{x}_i = \begin{cases} c_i, & \text{text-only models} \\ x_i, & \text{multimodal models} \end{cases}$$

The model outputs a probability distribution over rating classes:

$$p_\theta(y \mid \tilde{x}_i) = \text{Softmax}\big(g_\theta(\tilde{x}_i)\big), \tag{3.1}$$

where g_θ denotes the model's logit function.
Model parameters are learned by minimizing the training loss over all examples:

$$\min_\theta \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}\big(p_\theta(y \mid \tilde{x}_i), y_i\big), \tag{3.2}$$

where $\mathcal{L}$ is a task-specific loss function.

3.6.3 Prompt-Based Inference (Zero-Shot and Few-Shot)

In zero-shot and few-shot settings, no task-specific parameter updates are performed. A pretrained language model is prompted to infer the rating directly from the input. Let $\pi(\tilde{x}_i)$ denote a structured prompt. The predicted rating is:

$$\hat{y}_i = \arg \max_{k \in \{1,\ldots,5\}} p_\theta\big(k \mid \pi(\tilde{x}_i)\big), \tag{3.3}$$

where θ corresponds to fixed pretrained parameters.
In multimodal zero-shot settings, $\pi(\tilde{x}_i)$ incorporates visual embeddings:

$$\pi(\tilde{x}_i) = \pi(c_i, I_{ad,i}, I_{rev,i}) \tag{3.4}$$

Few-shot inference extends this by including a small set of labeled demonstration examples in the prompt.

3.6.4 Neuro-Symbolic Inference

In the neuro-symbolic framework, multiple predictors are combined:

$$\mathcal{M} = \{f_x, f_1, f_2, f_v\},$$

where f_x denotes the hierarchical XLM-RoBERTa classifier, f_1 and f_2 are independent LLM predictors, and f_v is the multimodal vision–language model.
Each predictor produces a rating estimate and, where applicable, a confidence score. A symbolic decision function $\Phi(\cdot)$ governs arbitration among outputs:

$$\hat{y}_i = \Phi\big(f_x(\tilde{x}_i), f_1(\tilde{x}_i), f_2(\tilde{x}_i), f_v(\tilde{x}_i)\big), \tag{3.5}$$

where Φ implements agreement rules, confidence thresholds and fallback logic (see Algorithms 1 and 2).

Objective. Across all paradigms, the goal is to learn or infer a function

$$f_\theta(x_i) \quad \text{or} \quad f_\theta(c_i)$$

that maps textual and optional visual review evidence to a discrete star rating

$$y_i \in \{1, 2, 3, 4, 5\},$$

while accounting for multilingual and multimodal variability.

3.7 Model Architectures and Learning Objectives

In this section, we describe the architectures of the models used for star rating prediction, including both text-only decoder LLMs and multimodal vision–language models. We also outline their learning objectives, highlighting how autoregressive generation and cross-modal fusion enable the extraction of rating information from textual and visual inputs.

3.7.1 Text Encoder (XLM-RoBERTa)

Textual inputs c_i are processed using XLM-RoBERTa [35], a multilingual transformer-based encoder. The encoder is built on the standard Transformer [36] architecture. Fig. 3.13 shows the original Transformer design with stacked encoder and decoder blocks [36]; the visualization is adapted from [37]. XLM-RoBERTa utilizes only the encoder stack.
Self-attention computes contextualized token representations by measuring pairwise interactions between queries, keys and values. Formally, scaled dot-product attention is defined as:

$$\text{Attention}(Q, K, V) = \text{Softmax}\Big(\frac{QK^\top}{\sqrt{d_k}}\Big)V, \tag{3.6}$$

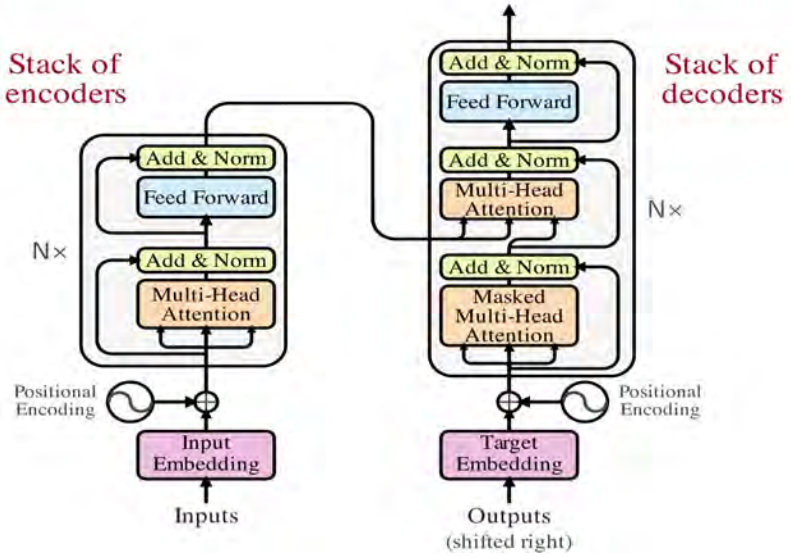


Figure 3.13: Original Transformer architecture with stacked encoder and decoder layers.

where Q, K, and V are the query, key, and value matrices derived from token embeddings, and d_k is the dimensionality of the keys. Multi-head attention allows the model to capture different contextual relations simultaneously:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \ldots, \text{head}_h)\, W^O, \tag{3.7}$$

with each $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$.
The transformer output for a review sequence c_i is

$$h_i^{text} = \text{XLM-RoBERTa}(c_i) \in \mathbb{R}^{L \times d_t}, \tag{3.8}$$

where L is the sequence length and d_t is the embedding dimension. This captures semantic, cross-lingual and contextual sentiment information.

3.7.2 Vision Encoder (ViT / CLIP)

Visual inputs $I_{ad,i}$ and $I_{rev,i}$ are encoded using either a Vision Transformer (ViT) [38] or the CLIP image encoder [39].

ViT. Each image is split into N patches of size $P \times P$, flattened and linearly projected to embeddings $x_p \in \mathbb{R}^{N \times d_v}$. The overall Vision Transformer pipeline [38], including patch embedding and transformer encoder blocks, is illustrated in Fig. 3.14. A learnable class token x_{cls} is prepended:

$$x_0 = [x_{cls}; x_1, \ldots, x_N] + E_{pos}, \tag{3.9}$$

where E_{pos} is the positional embedding. Transformer blocks with multi-head self-attention produce patch-level representations, and the class token embedding serves as the image representation:

$$h_i^{ad} = \text{ViT}(I_{ad,i}), \quad h_i^{rev} = \text{ViT}(I_{rev,i}). \tag{3.10}$$

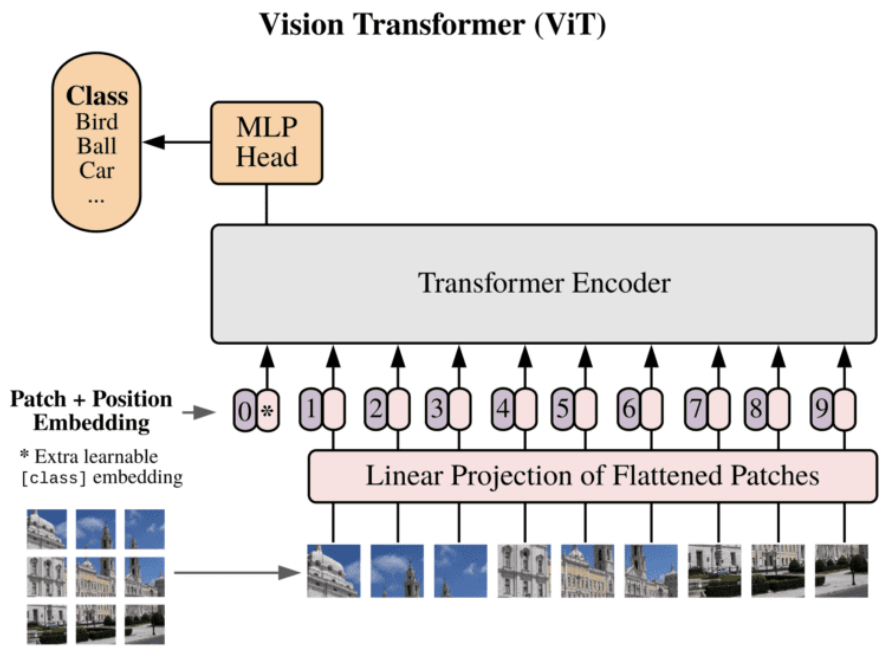


Figure 3.14: Original Vision Transformer (ViT) architecture showing image patch embedding and stacked transformer encoder layers.

CLIP. CLIP maps images and text into a shared latent space:

$$z_i^{ad}, z_i^{rev} = \text{CLIP}_{img}(I_{ad,i}, I_{rev,i}), \quad z_i^{text} = \text{CLIP}_{text}(c_i), \tag{3.11}$$

enabling multimodal alignment.

3.7.3 Decoder-Only Large Language Models

Decoder-only LLMs are autoregressive architectures that generate tokens sequentially, conditioned on prior context. These models build upon the original Transformer [36] decoder block, which uses masked self-attention to prevent future tokens from influencing the current token:

$$\text{MaskedAttention}(Q, K, V) = \text{Softmax}\Big(\frac{QK^\top}{\sqrt{d_k}} + M\Big)V \tag{3.12}$$

where M is a causal mask ensuring autoregressive generation.
The **gpt-oss** family (e.g., `gpt-oss-20B`) [40] exemplifies a high-capacity decoder-only LLM. Building upon the original Transformer architecture, it introduces several architectural refinements to improve scalability, efficiency and long-context modeling.
1. **Mixture-of-Experts (MoE):** GPT-OSS [40] introduces sparsely activated MoE blocks. Each token routes to the top-k experts, producing

$$h_{\text{MoE}} = \sum_{i \in \text{Top-4}} \text{Softmax}(\text{Router}(x))_i \cdot \text{Expert}_i(x) \tag{3.13}$$

which increases model capacity without proportionally increasing compute.

2. **Attention Variants:** Alternating banded window and dense attention patterns, combined with Grouped Query Attention (GQA), reduce memory usage while supporting long-context reasoning (up to 131K tokens with Rotary Positional Embeddings, RoPE).
3. **Normalization and Activation:** RMS normalization stabilizes deep network activations:

$$\text{RMSNorm}(x)_i = \frac{x_i}{\sqrt{\frac{1}{n}\sum_{j=1}^{n} x_j^2 + \epsilon}} g_i, \tag{3.14}$$

while SwiGLU activations improve expressivity beyond the original Transformer's GELU.
Star ratings are generated autoregressively from a structured prompt $\pi(c_i)$:

$$\hat{y}_{\text{tokens}} = \text{Generate}(\pi(c_i)) \quad \Rightarrow \quad \hat{y}_i = \text{ExtractRating}(\hat{y}_{\text{tokens}}),$$

where ExtractRating$(\cdot)$ parses the generated text to obtain the final star rating. Few-shot prompting extends $\pi(c_i)$ with labeled examples.
Although GPT-OSS demonstrates extreme scale and sparsity, its core autoregressive decoder mechanism remains aligned with the original Transformer decoder, retaining residual connections, causal attention and layer-wise normalization (here replaced by RMSNorm).

3.7.4 Vision–Language Models

Vision–language (VL) models extend the Transformer to integrate textual and visual information. Given textual embeddings h_i^{text} and visual embeddings h_i^{vis} (from $I_{ad,i}$ and $I_{rev,i}$), cross-modal attention fuses these modalities:

$$\text{Attention}(Q_{text}, K_{vis}, V_{vis}) = \text{Softmax}\Big(\frac{Q_{text}K_{vis}^{\top}}{\sqrt{d_k}}\Big)V_{vis} \tag{3.15}$$

The models considered here–**Gemma 3n**, **Llama 3.2 Vision Instruct** and **Qwen 3-VL**–differ in how they integrate visual features:
1. **Gemma 3n** uses a MobileNet-V5 vision encoder [41]. Parameter-efficient mechanisms such as Per-Layer Embedding (PLE) caching and conditional parameter loading allow selective activation of submodules. Its MatFormer architecture nests smaller Transformers within a larger network, enabling resource-adaptive multimodal inference.
2. **Llama 3.2 Vision Instruct** leverages a pre-trained ViT-H/14 vision encoder and injects multi-layer visual features into the decoder via cross-attention adapters, preserving spatial detail while maintaining a dense Transformer core.
3. **Qwen 3-VL** employs a SigLIP-2 vision encoder [42] with DeepStack integration, feeding multi-layer visual tokens into the language decoder through residual connections. Interleaved MRoPE positional encodings support temporal and spatial modeling for video and image sequences.
All three VL models share the following enhancements over the original Transformer:
- **Cross-modal Fusion:** Attention is computed between textual queries and visual keys/values, producing a joint representation:

$$h_i^{multi} = \text{Fuse}(h_i^{text}, h_i^{vis}). \tag{3.16}$$

- **Extended Context and Positional Encodings:** RoPE and MRoPE enable extremely long sequences and multi-dimensional inputs, supporting dense or sparse token streams beyond the original Transformer's capacity.
- **Parameter Efficiency and Modularization:** Conditional parameter loading and nested submodules reduce memory footprint and computation while preserving multimodal expressivity.
Star ratings are generated autoregressively from the fused representation:

$$\hat{y}_{\text{tokens}} = \text{Generate}(h_i^{multi}) \quad \Rightarrow \quad \hat{y}_i = \text{ExtractRating}(\hat{y}_{\text{tokens}}),$$

where the final rating is obtained by extracting the relevant token(s) from the generated sequence. This allows the VL models to reason jointly over text and image content in an autoregressive manner.

3.7.5 Learning Objectives

For the supervised classification setting, fine-tuned models are primarily optimized using the standard cross-entropy loss over the five rating classes:

$$\mathcal{L}_{\text{CE}}(\theta) = -\sum_{k=1}^{5} y_{i,k} \log p_\theta(y_{i,k} \mid x_i), \tag{3.17}$$

where $y_{i,k}$ denotes the one-hot encoded ground-truth label for class k. This objective serves as the main loss function for the classification task.
In addition to this standard formulation, we use a custom loss function tailored to our multimodal setting, which is described in a later subsection.
Zero-shot and few-shot models instead rely on the pretrained LLM objective without task-specific parameter updates.

3.7.6 Optimization

Parameters are optimized with AdamW [43] and learning rate scheduling. Multimodal training jointly aligns text and visual embeddings in the shared latent space.

3.8 Zero-Shot Methods (Text-Only)

In the text-only zero-shot setting, models are prompted to predict star ratings directly from customer comments without any task-specific fine-tuning. The prompts instruct the model to infer sentiment intensity and map it to a discrete 5-point rating scale. This setting evaluates the intrinsic reasoning and sentiment understanding capabilities of large language models under minimal supervision.
We evaluate Qwen-3-VL-8B [44], Gemma-3n-4B [45], LLaMA-3.2-11B-Vision-Instruct [46] and GPT-OSS-20B [40] models.
All models receive only the textual review content as input, with no access to product metadata or images.

3.9 Zero-Shot Methods (Multimodal)

To assess the contribution of visual context, we extend zero-shot inference to the multimodal setting by providing both textual reviews and associated images. The images include advertised and customer-uploaded photos, enabling the model to jointly reason over textual and visual cues.
The following vision–language models (VLMs) are evaluated: Qwen-3-VL-8B, Gemma-3n-4B and LLaMA-3.2-11B-Vision-Instruct.
Models are prompted to integrate visual evidence with textual sentiment when predicting the final star rating.

3.10 Few-Shot Methods (Text-Only)

In the few-shot setting, models are provided with a small number of labeled examples within the prompt, demonstrating how textual reviews correspond to star ratings. This setup evaluates the in-context learning ability of large language models under limited supervision.
We experiment with the same family of models used in the zero-shot setting: Qwen-3-VL-8B, Gemma-3n-4B, LLaMA-3.2-11B-Vision-Instruct, GPT-OSS-20B.

3.11 Fine-Tuned Models

Beyond prompting-based approaches, we train supervised models on DarazMM to establish strong task-specific baselines.

XLM-Roberta-Base (XLMR-base). We fine-tune XLM-Roberta-base [35] on review text to predict 5-class star ratings using a standard cross-entropy loss.

XLM-Roberta + ViT with Hybrid Loss (XLMR-hybrid). We further develop a multimodal XLM-Roberta-based model that jointly leverages textual and visual information for rating prediction. Review comments and product descriptions are encoded using XLM-Roberta-base [35], while advertised and customer-uploaded images are processed with a ViT-B/16 [38] encoder. The resulting modality-specific representations are projected into a shared embedding space for joint optimization via contrastive and consistency objectives, while rating prediction is performed from the review-comment representation.
Training is performed using a weighted hybrid objective that combines supervised classification with cross-modal alignment and visual consistency constraints:

$$\mathcal{L}_{\text{total}} = 0.8\,\mathcal{L}_{\text{CE}} + 0.1\,\mathcal{L}_{\text{InfoNCE}} + 0.1\,\mathcal{L}_{\text{MSE}}, \tag{3.18}$$

where $\mathcal{L}_{\text{CE}}$ denotes the categorical cross-entropy loss for 5-class star rating prediction (eq. 3.17), $\mathcal{L}_{\text{InfoNCE}}$ is a bidirectional temperature-scaled InfoNCE loss [47] that aligns review/comment embeddings ($\mathbf{r}_i$) and product description embeddings ($\mathbf{d}_i$) via symmetric contrastive learning, and $\mathcal{L}_{\text{MSE}}$ is a feature-space mean squared error loss [48] that minimizes the L2 distance between embeddings of advertised and customer-uploaded images to encourage visual consistency.

The InfoNCE loss is defined as:

$$\mathcal{L}_{\text{InfoNCE}} = -\frac{1}{N}\sum_{i=1}^{N}\Big[\log\frac{\exp(\text{sim}(\mathbf{r}_i,\mathbf{d}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{r}_i,\mathbf{d}_j)/\tau)} + \log\frac{\exp(\text{sim}(\mathbf{d}_i,\mathbf{r}_i)/\tau)}{\sum_{j=1}^{N}\exp(\text{sim}(\mathbf{d}_i,\mathbf{r}_j)/\tau)}\Big], \tag{3.19}$$

and the MSE consistency loss is:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N}\sum_{i=1}^{N}\|\mathbf{v}_i^{\text{ad}} - \mathbf{v}_i^{\text{rev}}\|_2^2, \tag{3.20}$$

where $\mathbf{v}_i^{\text{ad}}$ and $\mathbf{v}_i^{\text{rev}}$ are the projected embeddings of advertised and customer-uploaded images for sample i, $\mathbf{r}_i$ is the review/comment embedding, $\mathbf{d}_i$ is the product description embedding, $\text{sim}(\cdot,\cdot)$ denotes cosine similarity, τ is the temperature hyperparameter, and N is the batch size.

XLM-Roberta + CLIP-ViT Feature Fusion (XLMR + CLIP). To incorporate visual information, we concatenate textual features from XLM-Roberta-base with image embeddings extracted using CLIP ViT-B/16 [39]. Text and image features are fused via standard concatenation before classification. For training, we use the same hybrid loss as defined in eq. 3.18.

Two-Stage XLM-Roberta (XLMR-two-stage). We implement a hierarchical classification strategy to reduce confusion between adjacent rating classes. In the first stage, reviews are classified into three coarse categories: *positive* (ratings 4–5), *neutral* (rating 3) and *negative* (ratings 1–2). The first-stage sentiment classifier is trained using cross-entropy loss combined with ordinal smoothing. Reviews predicted as positive are routed to a binary 4-vs-5 classifier, while negative reviews are routed to a 1-vs-2 classifier. Both of these second-stage classifiers were trained using standard cross-entropy loss. Neutral reviews require no further processing.
Hyperparameters for all fine-tuned models are summarized in Table 3.4.

Table 3.4: Training hyperparameters for the fine-tuned models on the DarazMM dataset.

Model	**Batch Size**	**Optimizer**	**Loss Function**	**Learning Rate**
XLMR-base	32	AdamW	Cross-Entropy	2×10^{-5}
XLMR-hybrid	32	AdamW	Hybrid (Eq. 3.18)	1×10^{-5}
XLMR + CLIP	32	AdamW	Hybrid (Eq. 3.18)	1×10^{-5}
XLMR-two-stage	32	AdamW	CE + Ordinal Smoothing	2×10^{-5}

Note: Hybrid loss refers to the weighted combination of cross-entropy, contrastive and perceptual losses defined in Equation 3.18. CE refers to standard cross-entropy loss.

3.12 Prompts Used for Zero-Shot and Few-Shot Inference

In our experiments, we employed carefully designed prompts to guide model behavior across different tasks, including zero-shot inference and few-shot inference. These prompts are tailored to each model to ensure accurate understanding of review content and to generate reliable ratings. The following sections summarize the specific prompts used for each purpose.

3.12.1 Prompt Used for Zero-Shot Inference (GPT-OSS)

Fig 3.15 shows the prompt template used with GPT-OSS for zero-shot prediction of star ratings from customer reviews.

```
<|start|>system<|message|>
You are ChatGPT, a large language model trained by OpenAI.
Knowledge cutoff: 2024-06
Current date: 2025-10-06
Reasoning: medium
# Valid channels: analysis, commentary, final. Channel must be included for every message.
<|end|>

<|start|>user<|message|>
You are a helpful assistant that predicts product review star ratings based on customer comments.
The star rating scale is:
1 = খুব খারাপ অভিজ্ঞতা [very bad experience]
2 = খারাপ [bad]
3 = মোটামুটি [okay / average]
4 = ভালো [good]
5 = খুব ভালো অভিজ্ঞতা [very good experience]

Comment: "products is very good. as like picture. This time delivery was fast. keep it up.
Thanks to seller. Thanks to Daraz for fast delivery."

Give only the star rating (1-5) that best matches the sentiment.
<|end|>
```

Figure 3.15: Example GPT-OSS prompt used for star rating prediction from customer reviews. English glosses shown in gray are provided for reader clarity and are *not* part of the actual prompt.

3.12.2 Prompt Used for Zero-Shot Multimodal Inference (Gemma-3n)

The prompt in Fig 3.16 was used for zero-shot multimodal inference with the Gemma-3n model. The model jointly considers textual inputs (comments and product description) and visual inputs (advertised image and reviewer uploaded image).

<|start|>user<|message|>
You are a helpful assistant that predicts product review star ratings by analyzing:
1. The reviewer's comment (highest priority)
2. The advertised product image (Image 1)
3. The official product description
4. The reviewer uploaded image (Image 2, if available)

Image 1: Advertised product

Product description:
"Brand Name : Curren Movement : QUARTZ Clasp Type : Bracelet Clasp Origin : CN(Origin) Case Material : Alloy Water Resistance Depth : 3Bar Style : Fashion ;Casual Band Width : 12mm Case Shape : ROUND Case Thickness : 8mm Feature : None Dial Window Material Type : Hardlex Model Number : 9072 Band Length : 22cm Band Material Type : STAINLESS STEEL Dial Diameter : 31mm Watches Women CURREN Top Brand Fashion Thin Quartz Wristwatch with Charming Stainless Steel Bracelet"

Image 2: Reviewer uploaded image

Reviewer comment: "My Order colour rose good but Delivery Blu One Refund my money"

Compare and evaluate the following:
- How well the reviewer's comment reflects the product experience.
- Similarity between Image 1 and the product description.
- Similarity or dissimilarity between Image 1 and Image 2 (reviewer's image).
- Similarity or dissimilarity between the reviewer's comment and the product description.

Based on these evaluations, give ONLY the star rating (1-5):
1 = খুব খারাপ অভিজ্ঞতা [very bad experience]
2 = খারাপ [bad]
3 = মোটামুটি [okay / average]
4 = ভালো [good]
5 = খুব ভালো অভিজ্ঞতা [very good experience]

Do NOT provide any explanation, reasoning, or extra text.
Only output a single digit (1-5).
<|end|>

Figure 3.16: Example zero-shot multimodal prompt used with the Gemma-3n model for text–image based star rating prediction. English glosses shown in gray are provided for reader clarity and are *not* part of the actual prompt.

3.12.3 Prompt Used for Few-Shot Inference (Gemma-3n)

The prompt used for few-shot star rating prediction with the Gemma-3n model has been illustrated in Fig 3.17. The model was provided with five labeled examples per class before being asked to predict the star rating of a new, unseen review.

```
<|start|>user<|message|>
You are a classifier that predicts product review star ratings (1-5) based on customer
comments.

The star rating scale is:
1 = খুব খারাপ অভিজ্ঞতা [very bad experience]
2 = খারাপ [bad]
3 = মোটামুটি [okay / average]
4 = ভালো [good]
5 = খুব ভালো অভিজ্ঞতা [very good experience]

Here are some examples:

1 → "khub baje product experience korlam"[Experienced a really bad product]
1 → "vaiya fizlami korlen? dekhan akta disen arek ta"[Brother, are you joking? You showed
one, and gave another.]
1 → "ছবিতে কি দেখলাম আর কি পাইলাম।" [What I saw in the picture vs. what I received]
1 → "ফালতু পুরাই। "দারাজ' এর কাছে এমন প্রতারণা আশা করিনি। এক কোয়ালিটি দেখিয়ে অন্য পণ্য দেয়া মারাত্মক প্রতারণা।"
[Completely useless; severe deception from Daraz by showing one quality and delivering
another]
1 → "Doesn't match the picture plus no reply from the seller!"

2 → "বাইরে থেকে দেখতে খুব একটা খারাপ না কিন্তু ভেতরের তিনটা পার্ট ফালতু লাগে এবং এইটা দেওয়ার কথা ছিল না।দাম হিসেবে বাজে
একটা প্রোডাক্ট ।" [From the outside, it doesn't look too bad, but the three internal parts seem
unnecessary, and this should not have been included. It's a poor product for the price.]
2 → "Poor product quality. পেনড্রাইভ লাগাইতে গেলে কভার খুলে আসে।" [The cover comes off when
inserting a pen drive]
2 → "খুব তাড়াতাড়ি পেয়েছি, কিন্তু ছবির সাথে মিল নেই, পন্যের মানটাও খারাপ। শুধু পায়ে দেয়ার পরে এই অবস্থা। জুতার ফিতার
হুকগুলো ছিড়ে যাবার উপক্রম। ডেলিভারি ম্যান টা খুব ভালো।" [Delivered fast, but doesn't match the picture and
quality is bad; laces are breaking; delivery person was good]
2 → "valo na chaisi akta paisi arakta, daraz seller Ra na dite parle dibe na, but onno ta
kno dey!"[It's not good. I received one, but not the other. Daraz sellers either don't give
it or won't give it, yet someone else gives the other one!]
2 → "picture not matching product not good. full plastic belt"

3 → "মোটামুটি" [Average]
3 → "Giving review after 20 days of use. Mixed Experience. The left foot shoe sole was
short. Very uncomfortable to wear. Also the sole gum was bad. Need super glue to fix it.
Still a little bit uncomfortable. I think my one was a quality fail product. Also the seller
response was bad. Buy 3 pairs only my one is faulty but wearable. Fast delivery. The other
thing was good."
3 → "jototo kharap mone korcilam toto taw kharap na abar khub besi j valo taw na but dam
hisebe thiki ace valoi chile kew nite paren"[It's not as bad as I had thought. It's not very
good either, but for the price, it's reasonable. Overall, it's fine and anyone can buy it.]
3 → "Not bad, but a few scratches, tho expected a bit more quality at this price."
3 → "মোটামুটি ঠিক আছে তবে দাম একটু বেশি।" [It's generally okay, but the price is a bit high.]

4 → "দাম অনুপাতে ব্যাগটি সুন্দর কিন্তু, আরও একটি পকেট ও জিপার দিলে আরও ভালো হতো।" [Nice for the price, but
another pocket and zipper would be better]
4 → "Sneakers fresh ache jodio old product, dham hisebe vloi, size ja diyecilm tai asce
perfect hoyece. Original North star real price 999৳ 50% discount e peyeci."[The sneakers are
in good condition despite being an older product. Considering the price, they are good. The
size I ordered arrived perfectly. The original North Star, priced at 999৳, I got it at a 50%
discount.]
4 → "আলহামদুলিল্লাহ ভালো ছিল, আপনারাও নিতে পারেন কম খরচে ঠকবেন না" [Thankfully it was good; you can buy
it cheaply without being cheated]
4 → "Product is good enough as price. But delivery was so late."
4 → "প্রোডাক্টটা মোটামোটি ভালো।" [Product is fairly good.]

5 → "Purchased it for my mom as a gift. She liked it very much. Thank you daraz and seller
for this amazing product. The packaging was good, product was good, hoping that color does
not fade way."
5 → "এক কথায়, অসাধারন।" [simply amazing]
5 → "product quality khub vlo, Apnara Sobai nita paran. Seller vai khub vlo."[The product
quality is very good. Everyone should consider buying it. The seller is very good.]
5 → "খুব ভালো যেমন চাইছি ঠিক তেমন পাইছি। ধন্যবাদ দারাজকে" [Very good; received exactly what I wanted;
thanks to Daraz]
5 → "Same as picture. excellent product. খুব কিউট চাবির রিং গিফট পেয়েছি। বিক্রেতা কে অসংখ্য ধন্যবাদ।"
[Received a very cute key ring gift; many thanks to the seller]

Now classify ONLY the following reviewer comment (not the examples above):

Comment: "My Order colour rose good but Delivery Blu One Refund my money"

Give only the star rating (1-5) that best matches the sentiment.
Do not write any explanation or anything else. <|end|>
```

Figure 3.17: **Example few-shot prompt used with the Gemma-3n model for star rating prediction from customer reviews.** English glosses shown in gray are provided for reader clarity and are *not* part of the actual prompt.

3.12.4 Neuro-symbolic Frameworks

We experiment with two neuro-symbolic frameworks for product review rating inference. The primary framework, illustrated in Fig 3.18, serves as the main experimental pipeline and is formalized in Algorithm 1. We additionally introduce an alternative framework with modified decision logic.

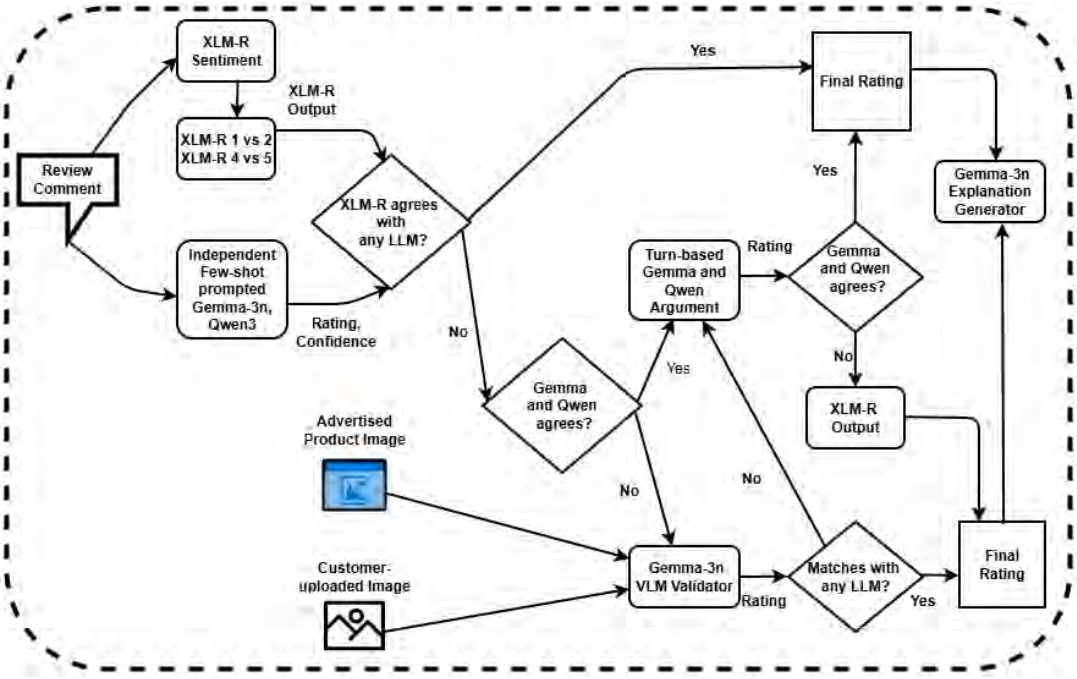


Figure 3.18: Overview of the proposed Neuro-symbolic Framework 1.

The framework integrates vision–language models (VLMs; e.g., Qwen-3-VL-8B and Gemma-3n) operated in text-only mode for rating prediction, a pre-trained Two-Stage XLM-Roberta model for hierarchical sentiment classification, an additional multimodal VLM for joint text–image reasoning and a symbolic decision controller. For brevity, VLMs used exclusively with textual inputs are referred to as *LLMs* in the remainder of the paper. Neural components provide probabilistic predictions, while symbolic rules govern model interaction through explicit decision logic.

Text-based Rating Prediction (LLMs). Two independent LLM predictors, implemented using Gemma-3n and Qwen-3-VL-8B, are used to infer star ratings (1–5) from customer comments. Each LLM outputs a discrete rating along with a self-estimated confidence score. The models operate independently to encourage diversity in reasoning and reduce single-model bias. These confidence scores are later used for arbitration and consensus decisions.

Multimodal Rating Inference (VLM). The Gemma-3n VLM is employed to infer ratings by jointly reasoning over textual and visual inputs, including the customer comment, advertised product image and customer-uploaded image when available. Visual comparison enables detection of discrepancies between advertised and received products. The VLM outputs a single integer rating and is used primarily as a validator rather than a standalone predictor.

Cross-Model Argumentation (LLM–LLM). When the two LLMs (Gemma-3n and Qwen-3-VL-8B in text-only mode) disagree, a structured argumentation phase is triggered. Each LLM is exposed to the other model's prediction and asked to either maintain or revise its own rating, producing an updated confidence score. A rating is accepted only when both models converge with sufficient confidence.

Symbolic Decision Controller. A symbolic controller orchestrates the pipeline using deterministic rules. It prioritizes agreement with the multilingual text classifier, namely the Two-Stage XLM-Roberta model, invokes multimodal validation under disagreement and applies LLM–LLM argumentation as a final reconciliation step. This design ensures interpretability, bounded reasoning depth and stable decision-making.

Explanation Generation (VLM). After a final rating is selected, the Gemma-3n VLM generates a concise natural language explanation grounded in the customer comment and visual evidence. This stage is decoupled from rating inference to avoid feedback leakage while improving transparency.

Overall Inference Procedure. Algorithm 1 summarizes the interaction between all components in the primary neuro-symbolic framework used in our experiments.

Algorithm 1 Neuro-Symbolic Rating Inference Framework 1

Require: Review text c, images I_{ad}, I_{rev}
Ensure: Final rating r^*
Obtain r_x from XLMR-two-stage
Obtain $(r_1, conf_1)$ from Gemma LLM using c
Obtain $(r_2, conf_2)$ from Qwen LLM using c
if $r_x = r_1$ **or** $r_x = r_2$ **then**
    $r^* \leftarrow r_x$
    return r^*
end if
if $r_1 \neq r_2$ **then**
    Obtain r_v from VLM using (c, I_{ad}, I_{rev})
    if $r_v = r_1$ **or** $r_v = r_2$ **then**
        $r^* \leftarrow r_v$
        return r^*
    end if
end if
$r_a \leftarrow$ LLM–LLM argumentation using $(r_1, conf_1), (r_2, conf_2)$ with threshold τ
if $r_a \neq$ None **then** ▷ Mutual agreement and sufficient confidence
    $r^* \leftarrow r_a$
    return r^*
end if
$r^* \leftarrow r_x$
return r^*

3.12.5 Neuro-Symbolic Framework 2

Algorithm 2 presents the second rating inference framework that leverages LLM ratings along with a text classifier. Unlike Framework 1, it first trusts the text classifier if it agrees with any VLM, performs confidence-aware VLM argumentation if all predictions disagree, uses the text classifier as a tiebreaker and selectively invokes visual reasoning for highly ambiguous cases. An explanation module generates reasoning for the final rating.

Algorithm 2 Neuro-Symbolic Rating Inference Framework 2

Require: Review text c, images I_{ad}, I_{rev}
Ensure: Final rating r^*
Obtain r_x from XLMR-two-stage
Obtain $(r_1, conf_1)$ from Gemma LLM using c
Obtain $(r_2, conf_2)$ from Qwen LLM using c
if $r_x = r_1$ **or** $r_x = r_2$ **then**
    $r^* \leftarrow r_x$ ▷ Trust text classifier if it matches any VLM
    return r^*
end if
$r_a \leftarrow$ LLM–LLM argumentation using $(r_1, conf_1), (r_2, conf_2)$ with threshold τ
if $r_a \neq$ None **then** ▷ Mutual agreement and sufficient confidence
    $r^* \leftarrow r_a$
    return r^*
end if
$r^* \leftarrow r_x$ ▷ Fallback to XLMR-two-stage classifier as tiebreaker
if $|r_1 - r_2| \geq 2$ **then**
    Obtain r^* from VLM using (c, I_{ad}, I_{rev}) ▷ Highly ambiguous cases
end if
return r^*

3.13 Experimental Setup

All experiments were conducted using Kaggle T4 GPUs with standard Python libraries, including PyTorch.

Chapter 4

Result Analysis

We evaluate zero-shot, few-shot, fine-tuned and neuro-symbolic framework approaches on the DarazMM dataset. Accuracy, Macro F1 and Weighted F1 are used as evaluation metrics. The results are presented and analyzed in the following sections.

4.1 Zero-shot evaluation results

Qwen3-VL-8B-Instruct (referred to as **Qwen3**), Gemma-3n-E4B-it (**Gemma-3n**), Llama-3.2-11B-Vision-Instruct (**LLaMA-3.2**) and GPT-OSS-20b (**GPT-OSS**) models are evaluated in zero-shot approach. Here, (t) denotes text-only input and (t, i) denotes joint text–image input.
Zero-shot evaluation results for both text-only and multimodal (text-image) settings are presented in Table 4.1, highlighting clear trends in model behavior under minimal supervision. Across the board, text-only models generally outperform their multimodal counterparts in zero-shot settings, suggesting that visual information may introduce noise or distract from the textual cues that drive initial predictions. Specifically, Qwen3 exhibits a notable drop in accuracy when images are incorporated (from 0.66 to 0.58), indicating that the addition of visual inputs does not necessarily complement its learned text representations in a zero-shot context. In contrast, Gemma-3n demonstrates remarkable stability across both settings, with only a minor decrease in macro F1 and weighted F1 when visual information is included, implying a more robust integration of multimodal signals even without task-specific fine-tuning. LLaMA-3.2 shows the most drastic deterioration when images are added, with accuracy dropping from 0.56 to 0.29, suggesting a possible misalignment between its textual and visual embeddings in the absence of supervision. Among text-only models, GPT-OSS achieves the highest zero-shot accuracy (0.71) and maintains competitive weighted F1 (0.72), underscoring its strong ability to generalize solely from textual inputs. These results collectively indicate that while zero-shot multimodal performance is highly sensitive to model architecture and pretraining alignment, text-only models currently provide the most reliable predictions under this evaluation paradigm.

Table 4.1: Zero-shot evaluation results for text-only and multimodal settings.

Model	Accuracy	Macro F1	Weighted F1
Qwen3 (t)	0.66	0.62	0.68
Qwen3 (t,i)	0.58	0.56	0.61
Gemma-3n (t)	0.69	**0.66**	**0.72**
Gemma-3n (t,i)	0.69	0.63	0.70
LLaMA-3.2 (t)	0.56	0.50	0.57
LLaMA-3.2 (t,i)	0.29	0.24	0.28
GPT-OSS (t)	**0.71**	0.65	**0.72**

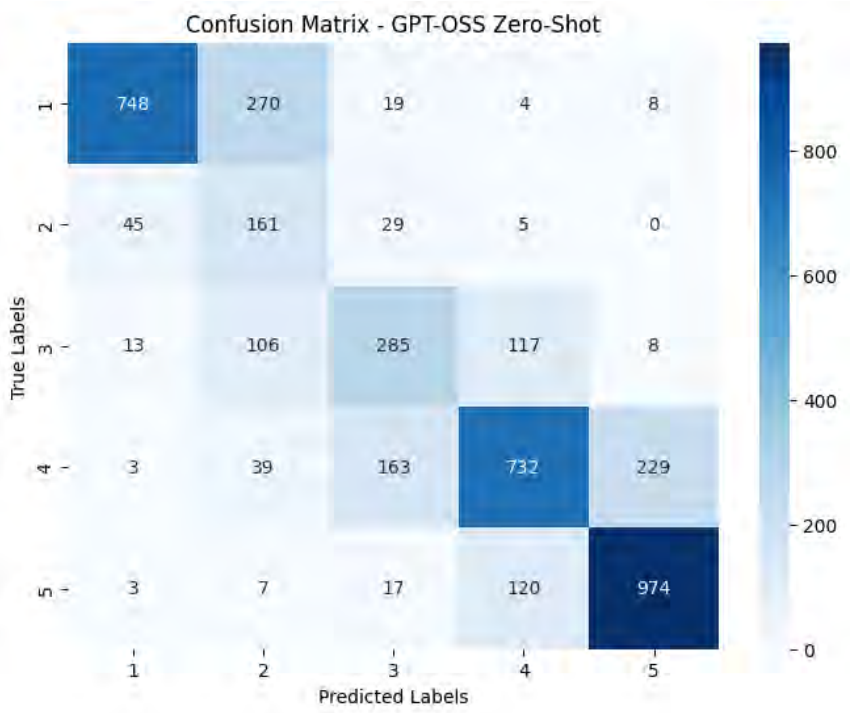


Figure 4.1: Confusion matrix of GPT-OSS (Zero-Shot) on the DarazMM test set.

Beyond aggregate metrics, class-wise behavior in zero-shot settings reveals systematic patterns. As shown in Fig 4.1, the GPT-OSS zero-shot model performs strongly on extreme ratings, particularly class 5, and reasonably on class 1. It struggles to distinguish mid-range ratings, causing substantial confusion among classes 2, 3 and 4. Neutral reviews (class 3) are often misclassified as class 2 or 4, and class 4 reviews are frequently predicted as class 3 or 5. This pattern reflects difficulty in modeling subtle sentiment intensity and fine-grained rating boundaries without task-specific calibration.

4.2 Few-shot evaluation results

We further evaluate the same set of models under a few-shot setting to assess their ability to leverage limited task-specific examples. As shown in Table 4.2, few-shot prompting generally leads to noticeable improvements in performance for Qwen3 and Gemma-3n, demonstrating that even a small number of examples can significantly enhance model predictions. In particular, Gemma-3n achieves the highest

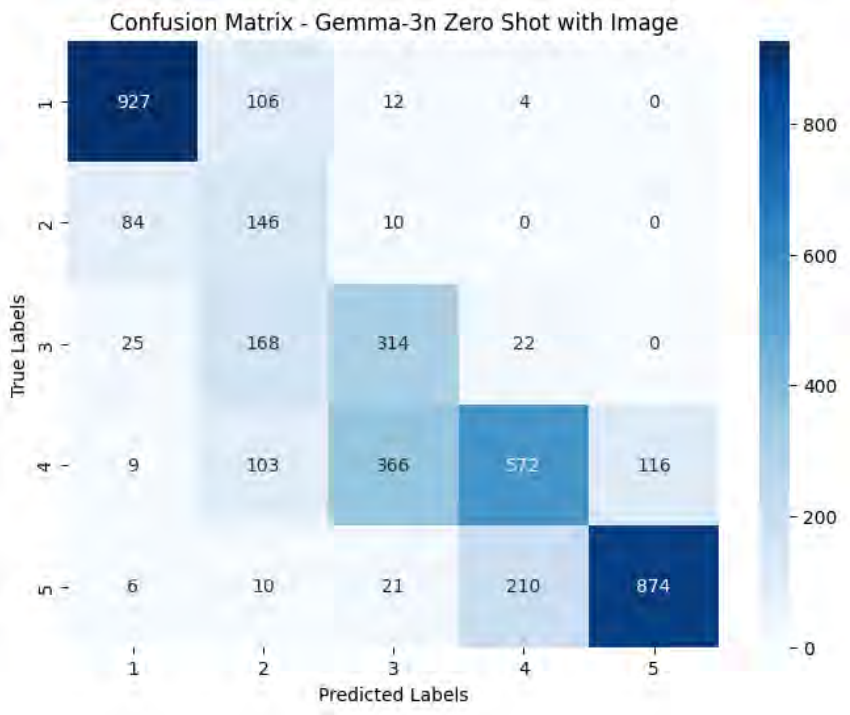


Figure 4.2: Confusion matrix of Gemma-3n (text and image) on the DarazMM test set.

accuracy (0.79), along with the strongest macro F1 (0.74) and weighted F1 (0.80), indicating that it not only predicts the most labels correctly but also maintains balanced performance across different classes. Qwen3 also benefits from few-shot examples, with its accuracy rising to 0.74, reflecting effective adaptation to the task despite its zero-shot limitations.

In contrast, LLaMA-3.2 and GPT-OSS show poor performance in the few-shot scenario, highlighting the varying sensitivity of models to prompt-based learning. LLaMA-3.2's moderate performance (accuracy 0.50) suggests that its capacity or pretraining alignment may not sufficiently support few-shot adaptation, while GPT-OSS struggles the most, with accuracy dropping dramatically to 0.22. This emphasizes that model size, pretraining objectives, and alignment with the target task play critical roles in determining few-shot learning success. Overall, the results underline that while few-shot prompting can substantially boost performance for certain architectures, others may require additional fine-tuning or supervision to achieve competitive results, pointing to the limitations of relying solely on prompt-based adaptation.

Fig 4.2 shows the behavior of Gemma-3n in the zero-shot multimodal setting, where incorporating images improves separation for extreme ratings (1 and 5), while mid-range ratings remain ambiguous. Gemma-3n frequently maps ratings 2 and 3 downward to 1 or 2, and class 4 ratings are often predicted as 3 or 5, suggesting that visual cues partially moderate textual signals when explicit dissatisfaction is not visually evident.

The impact of few-shot learning on class-wise predictions is illustrated in Fig 4.3. Few-shot prompting with Gemma-3n substantially reduces confusion among mid-range ratings (2, 3 and 4), while maintaining strong separation for extreme classes.

Table 4.2: Few-shot evaluation results using text-only input.

Model	Accuracy	Macro F1	Weighted F1
Qwen3 (t)	0.74	0.70	0.75
Gemma-3n (t)	**0.79**	**0.74**	**0.80**
LLaMA-3.2 (t)	0.50	0.48	0.53
GPT-OSS (t)	0.22	0.19	0.22

The in-context examples help Gemma-3n better align linguistic cues such as delivery issues or partial satisfaction with appropriate intermediate ratings rather than collapsing them toward neutral or positive classes.

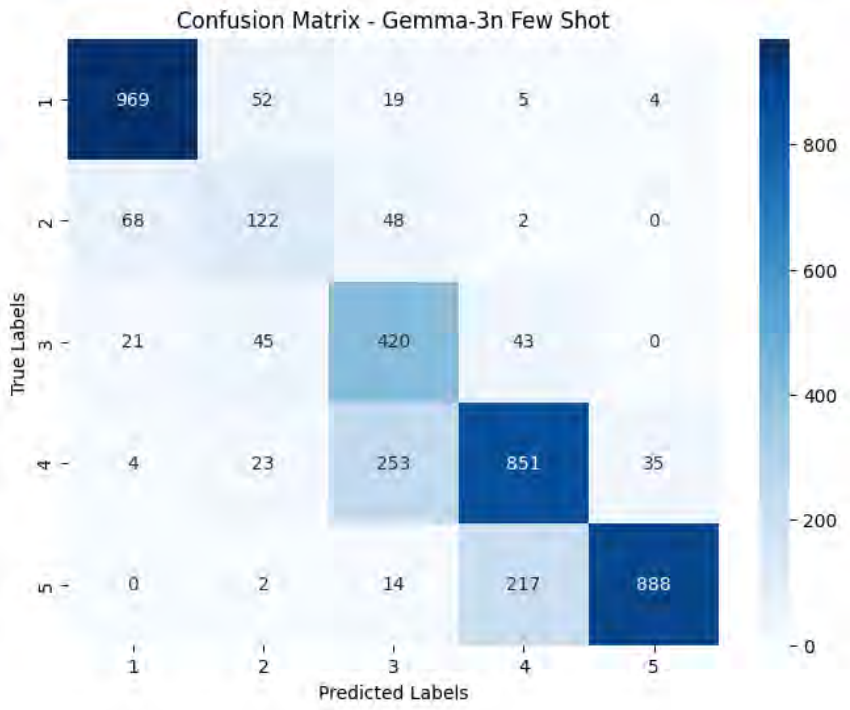


Figure 4.3: Confusion matrix of Gemma-3n (Few-Shot) on the DarazMM test set.

4.3 Fine-Tuned Models and Neuro-symbolic Frameworks

Following the evaluation of zero-shot and few-shot approaches, we next examine supervised fine-tuned models alongside the proposed neuro-symbolic frameworks to better understand the impact of task-specific training and structured reasoning. As reported in Table 4.3, supervised fine-tuning consistently provides substantial gains over both zero-shot and few-shot methods, highlighting the benefits of leveraging labeled data for domain adaptation. The baseline XLM-R model achieves moderate performance (accuracy 0.55), demonstrating that while pre-trained language representations provide a strong foundation, text-only supervision is insufficient for fully capturing multimodal nuances in the dataset.

Incorporating visual features through a CLIP-based fusion strategy significantly improves model accuracy to 0.71, indicating that carefully integrated visual cues can

enhance prediction robustness by complementing textual signals. Further improvements are achieved through hybrid loss optimization, which balances classification and representation alignment objectives, pushing accuracy to 0.76. The two-stage XLM-R training approach yields even stronger performance (accuracy 0.80), suggesting that progressive fine-tuning with intermediate representation learning enables the model to better capture complex patterns across both textual and visual modalities.

The proposed neuro-symbolic frameworks deliver the most substantial gains, surpassing all baseline models by a wide margin. Framework 2 achieves an accuracy of 0.86, while Framework 1 attains the highest overall performance with 0.87 accuracy and 0.83 macro F1. These results underscore the value of combining pre-trained large language models (LLMs) and vision-language models (VLMs) with a two-stage XLM-R backbone and explicit symbolic reasoning, which together facilitate more robust, interpretable and consistent predictions. The neuro-symbolic design allows the model to leverage structured decision logic, effectively guiding the final output and reducing errors arising from purely statistical inference. Collectively, these findings demonstrate that task-specific supervision, multimodal integration and symbolic reasoning synergistically enhance model reliability, providing a clear trajectory for improving performance in complex multimodal tasks.

Table 4.3: Performance of fine-tuned XLM-R models and proposed neuro-symbolic frameworks.

Model	**Accuracy**	**Macro F1**	**Weighted F1**
XLMR-base	0.55	0.46	0.54
XLMR + CLIP	0.71	0.58	0.69
XLMR-hybrid	0.76	0.78	0.76
XLMR-two-stage	0.80	0.72	0.79
Neuro-symbolic Framework 1	**0.87**	**0.83**	**0.87**
Neuro-symbolic Framework 2	0.86	0.81	0.86

Fig 4.4 illustrates the confusion matrix of the two-stage XLM-Roberta model, which shows consistent behavior across most classes with limited confusion between adjacent mid-range ratings. The hierarchical design enables XLM-R to first separate coarse sentiment polarity, reducing severe misrouting of negative reviews into positive classes before fine-grained rating decisions.

As shown in Fig 4.5, Neuro-Symbolic Framework 2 exhibits balanced class-wise behavior, with reduced confusion across both extreme and mid-range ratings. By combining XLM-R, multiple LLMs and symbolic arbitration, the framework effectively corrects borderline cases such as neutral-versus-slightly-positive reviews and avoids systematic bias toward any single rating level.

We further visualize the class-wise prediction patterns of Framework 1 using a confusion matrix, as shown in Fig 4.6. The model demonstrates very high accuracy for extreme ratings (1 and 5) and performs well for rating 4, while some confusion occurs between mid-range ratings (2 and 3).

Table 4.4 further analyzes the contribution of each decision module by reporting their usage frequency and accuracy across both frameworks. In both settings, the XLMR-priority stage dominates the final predictions, accounting for the majority of cases (3041 instances) while achieving the highest accuracy (0.95), indicating that

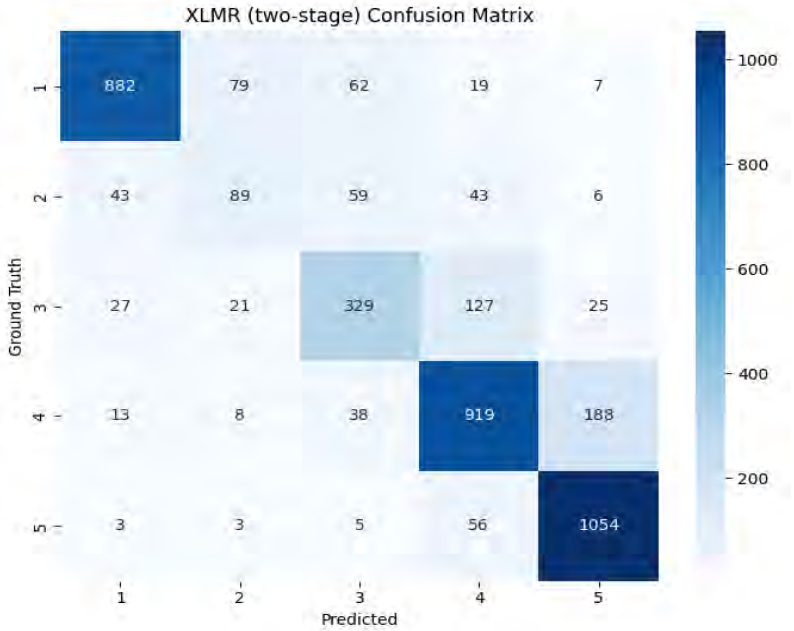


Figure 4.4: Confusion matrix of XLMR (two-stage) on the DarazMM test set.

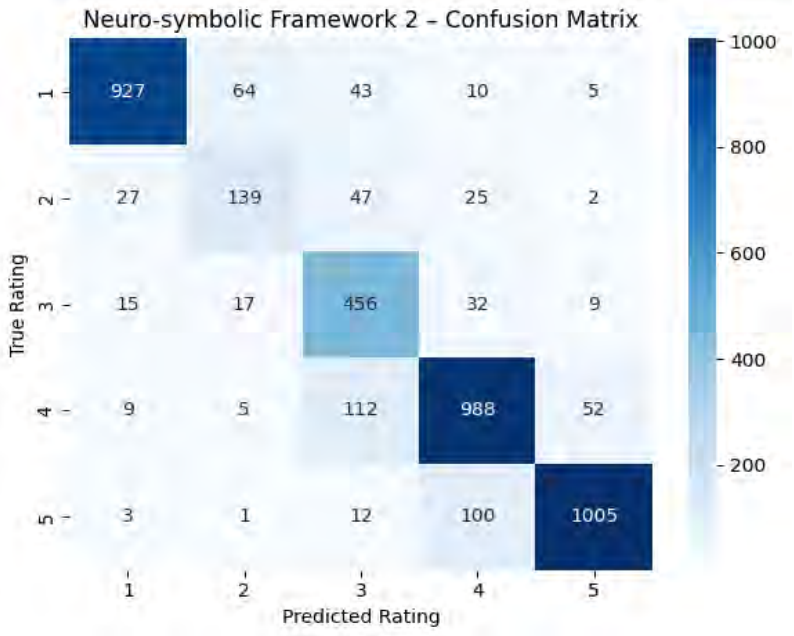


Figure 4.5: Confusion matrix of Neuro-symbolic Framework 2 on the DarazMM test set.

the XLM-RoBERTa two-stage model provides a strong and reliable primary signal. The LLM-agreement module is invoked moderately often and maintains competitive performance (0.65–0.68 accuracy), demonstrating the benefit of collaborative reasoning between language models when initial consensus is absent. In contrast, the VLM and XLMR-tiebreaker stages are triggered less frequently and exhibit comparatively lower accuracy, as they primarily operate in harder or ambiguous cases where earlier stages fail to agree. Notably, Framework 2 reduces reliance on the VLM while slightly increasing fallback decisions, reflecting different routing dynam-

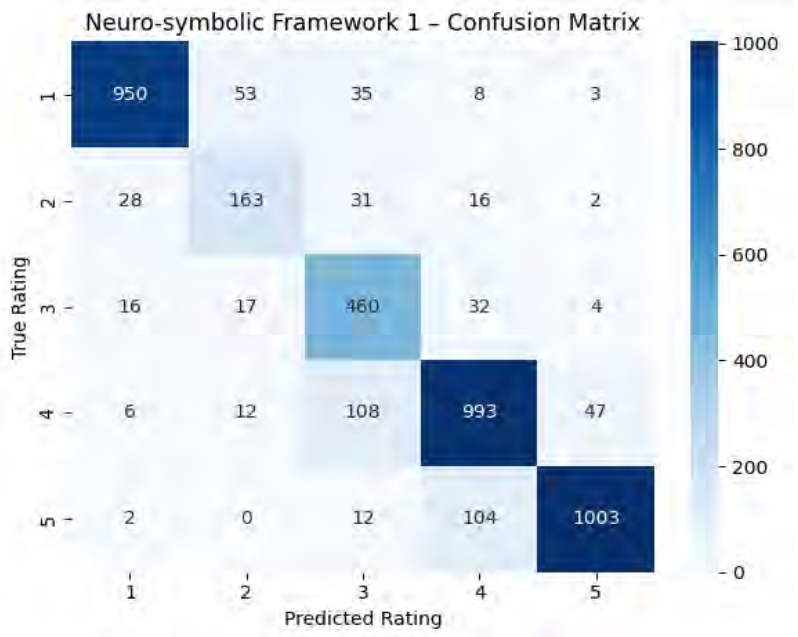


Figure 4.6: Confusion matrix of Neuro-symbolic Framework 1 on the DarazMM test set.

Table 4.4: Final decision module usage and accuracy comparison for the two frameworks.

Module	**Framework 1**			**Framework 2**		
	Used	**Correct**	**Acc**	**Used**	**Correct**	**Acc**
VLM	130	82	0.63	29	11	0.38
XLMR_priority	3041	2889	0.95	3041	2889	0.95
LLM_agreement	648	439	0.68	695	452	0.65
XLMR_tiebreaker	286	159	0.56	340	163	0.48

Note: XLMR_priority and XLMR_tiebreaker both correspond to decisions made by the XLMR-two-stage model. XLMR_priority denotes the initial consensus where XLMR agrees with at least one LLM prediction, while XLMR_tiebreaker represents the final fallback decision when no agreement is achieved by VLM or LLM–LLM argumentation.

ics but consistent trends in module effectiveness. Overall, these results confirm that early consensus from the text-based backbone drives most correct predictions, while later modules act as complementary safeguards to improve robustness.

Across all experimental settings, performance improves progressively from zero-shot to few-shot and finally to supervised and neuro-symbolic approaches. While text-only models generally outperform multimodal models in zero-shot scenarios, carefully designed fusion strategies and symbolic control mechanisms enable effective utilization of visual information under supervision. The strong performance of the proposed frameworks demonstrates that structured reasoning and modular integration are crucial for reliable multimodal sentiment and rating prediction in real-world e-commerce settings.

4.4 Discussion

The results confirm that the proposed objectives enable a more realistic evaluation of star rating prediction in Bangla-centric e-commerce settings. The constructed multilingual and multimodal dataset captures linguistic variability, informal writing and heterogeneous inputs that are often missing from existing benchmarks. This broader coverage ensures that models are assessed under practical deployment conditions rather than idealized scenarios.

Across all experimental settings, performance improves progressively from zero-shot to few-shot and finally to supervised and neuro-symbolic approaches. Text-only models generally perform better in zero-shot cases, whereas multimodal models benefit from supervision and carefully designed fusion strategies. These findings suggest that visual information is most effective when integrated through structured and guided learning rather than naive feature concatenation.

Furthermore, the neuro-symbolic frameworks show that explicitly modeling rating ordinality and cross-modal interactions leads to more reliable predictions. By combining structured reasoning with learned representations, the proposed approach achieves stable and consistent improvements. Overall, the outcomes validate the research objectives and highlight the importance of modular design and principled multimodal integration for real-world sentiment and rating prediction.

4.5 Sample Inference Using the Proposed Neuro-Symbolic Framework

In this section, we present several illustrative inference examples generated using the proposed neuro-symbolic framework. Among the different architectural variants explored in this study, Neuro-Symbolic Framework 1 is selected for demonstration, as it achieved the best overall performance across evaluation metrics. The inference pipeline integrates multimodal inputs, including the original text comment, the customer-uploaded product image and the advertised product image. Although translated versions of the textual comments are provided to facilitate interpretation and readability, they are not used as inputs to the predictive models.

The multimodal inputs are processed through a combination of specialized components within the framework. Textual representations are modeled using XLMR-two-stage, while large language models (LLMs), such as Gemma-3n and Qwen-3, are employed to perform reasoning and rating prediction. In addition, the visual modality is analyzed using a vision-language model (VLM), specifically Gemma-3n, to capture discrepancies or alignments between the advertised and received products. The framework incorporates a structured reasoning mechanism in which model outputs may interact through turn-based argumentation to refine intermediate predictions before arriving at a consolidated decision.

Finally, the explanation module synthesizes the predicted rating together with evidence extracted from both textual and visual inputs to generate a coherent justification for the final decision. This module enhances the interpretability of the framework by explicitly linking multimodal evidence to the predicted outcome, thereby providing transparency and supporting trust in the model's reasoning process.

Fig 4.7 illustrates a complete inference trace of the proposed neuro-symbolic frame-

Sample Inference Trace of the Proposed Framework

Input Review

Original Text Comment:
দেখাইলেন কি? দিলেন কি?
দেখাইলেন MX4, দিলেন HM501
Product পেয়েছি, তবে যেটা চেয়েছিলাম সেটা পাইনি, তবে দেখে মনে হচ্ছে Thermal Paste ভালো হবে।
Translated Text Comment:
What did you advertise? What did you provide?
You advertised MX4, but provided HM501.
I received a product, but not the one I requested.
However, judging by its appearance, the thermal paste seems to be good.
Ground truth rating: 3

Ratings from Models

XLMR-two-stage: 3
Gemma: 1 (high confidence)
Qwen: 2 (medium confidence)

Turn based LLM argumentation
Gemma decision: 3 confidence: medium (after receiving Qwen: 2)
Gemma reason: The reviewer indicates the product is okay but not ideal, specifically mentioning thermal paste, suggesting a neutral to slightly positive experience.

Qwen decision: 3 confidence: medium (after receiving Gemma: 1)
Qwen reason: The reviewer indicates the product isn't what they wanted, but the thermal paste seems good, suggesting a neutral to slightly positive experience.

Visual Reasoning with Gemma-3n

Advertised product

Reviewer-uploaded image

Gemma-3n VLM Output
Rating: 3

Final Decision
Predicted rating: 3
Ground truth: 3

Explanation Module
The reviewer expresses disappointment, stating "তবে যেটা চেয়েছিলাম সেটা পাইনি" which translates to "but I didn't get what I wanted." While they received the MX4 and HM501 products, their desired item was not included. The reviewer also suggests the thermal paste "will be good," implying a positive aspect despite the overall disappointment.
The images show the received products, MX4 and HM501, which confirms the reviewer received something, but doesn't address the core issue of not receiving the desired item. Therefore, the final rating is 3, reflecting a mixed experience.

Figure 4.7: Inference example illustrating the integration of text classification, LLM argumentation, visual reasoning and explanation, resulting in a final rating of 3.

work (Framework 1), demonstrating how textual, visual and argumentative reasoning are combined to produce a final rating. Initially, the two-stage XLM-R model predicts a neutral rating of 3, while the LLMs provide lower preliminary estimates (Gemma: 1, Qwen: 2), reflecting differences in how each model interprets the review content. According to Algorithm 1, the framework first checks for agreement between XLM-R and either LLM; since no direct match occurs, the system proceeds to reconcile the LLM disagreement. Through turn-based LLM argumentation, the two models iteratively exchange confidence-weighted opinions, eventually converging on a mutually agreed rating of 3. Simultaneously, visual reasoning is employed using

Sample Inference Trace of the Proposed Framework

Input Review

Original Text Comment:
প্রোডাক্টটা খুবই ভালো ছিল। কিন্তু ওয়ালে লাগানোর পরপরই আঠাটা ছুটে গেল। আঠাটা আর কাজ করছে না।
আমি অনেক দুঃখিত, প্রোডাক্টটি পেয়ে ভালো লাগেনি।
আঠা লাগানোর কোন উপায় জানা থাকলে জানাবেন প্লিজ।

Translated Text Comment:
The product was very good. But immediately after attaching it to the wall, the glue came off.
The adhesive no longer works.
I am very disappointed; I did not like receiving this product.
Please let me know if there is any way to apply the adhesive properly.

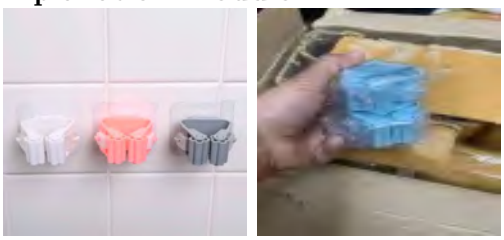
Ground truth rating: 2

Ratings from Models

XLMR-two-stage: 2
Gemma: 2 (high confidence)
Qwen: 2 (high confidence)

Final Decision
Predicted rating: 2
Ground truth: 2

Explanation Module



Advertised product Reviewer-uploaded image

The final rating is 2 because the reviewer states "কিন্তু ওয়ালে লাগানোর পরপরই আঠাটা ছুটে গেল। আঠাটা আর কাজ করছে না।" which translates to "But the glue came off immediately after being attached to the wall. The glue no longer works." This indicates a significant defect in the product's adhesive, leading to dissatisfaction. The reviewer explicitly expresses disappointment ("পেয়ে ভালো লাগেনি"). The image does not provide additional information regarding the adhesive failure. Therefore, the faulty adhesive is the core reason for the low rating.

Figure 4.8: Inference example where XLM-R agrees with both LLMs, leading to a direct prediction of rating 2. The explanation module highlights textual evidence for the low rating.

the submitted images (I_{ad}, I_{rev}) to validate the textual interpretation, confirming that the reviewer's experience aligns with the neutral assessment. The explanation module then provides interpretable reasoning: although the reviewer did not receive the exact desired product, they acknowledge a positive aspect in the thermal paste, resulting in a rating that captures this mixed sentiment. This trace exemplifies the algorithm's hierarchical decision process—starting from XLM-R predictions, moving through multimodal verification, and resolving conflicts via LLM argumentation—culminating in a robust, human-interpretable final rating. The example highlights the effectiveness of integrating symbolic decision logic with pre-trained language and vision-language models to produce consistent, transparent and nuanced inference outcomes.

Fig 4.8 presents an inference example where the text classifier XLM-R fully agrees with both LLMs, Gemma and Qwen, with all models predicting the correct low rating of 2. In this case, the proposed framework identifies the unanimous consensus at the initial stage, allowing it to directly adopt the XLM-R prediction as

Sample Inference Trace of the Proposed Framework

Input Review

Original Text Comment:
Nice shirts. Not an export quality but high quality made for local product. Fair price, thank you!
Ground truth rating: 4

Ratings from Models

XLMR-two-stage: 5
Gemma: 4 (high confidence)
Qwen: 4 (high confidence)

Turn based LLM argumentation
Gemma decision: 4 confidence: high (after receiving Qwen: 4)
Gemma reason: The review indicates a positive experience with a fair price for a product intended for local use, aligning with the initial rating.

Qwen decision: 4 confidence: high (after receiving Gemma: 4)
Qwen reason: The review indicates a positive experience with a fair price for a locally made, high-quality product, aligning with the initial rating.

Final Decision
Predicted rating: 4
Ground truth: 4

Explanation Module


Advertised product


Reviewer-uploaded image

The final rating is 4 because the reviewer states "high quality made for local product" and "Fair price," indicating a positive experience with the product's quality and value for its intended purpose. While not export quality, it meets expectations for a local product. The reviewer also expresses gratitude, suggesting overall satisfaction.

Figure 4.9: Inference example where XLM-R initially disagrees with both LLMs, triggering argumentation. Both LLMs converge on rating 4, which becomes the final prediction, supported by textual evidence.

the final rating without engaging additional LLM argumentation or invoking visual reasoning. This illustrates the efficiency of the algorithm when model predictions are aligned, as it minimizes unnecessary computational steps while maintaining accuracy. The explanation module further provides interpretability by highlighting the critical textual evidence underlying the low rating, specifically emphasizing that the faulty adhesive led to user dissatisfaction. This example underscores how the framework effectively leverages agreement among models to produce a reliable and transparent output, demonstrating that when textual cues are clear and consistent, the neuro-symbolic reasoning pipeline can confidently resolve the rating using minimal processing, preserving both efficiency and interpretability.
Fig 4.9 illustrates an inference example in which the text classifier XLM-R initially predicts a higher rating of 5, while both LLMs, Gemma and Qwen, suggest a slightly lower rating of 4. This disagreement between the text classifier and the LLMs ac-

Sample Inference Trace of the Proposed Framework

Input Review

Original Text Comment:
My Order colour rose good but Delivery Blu One Refund my money
Ground truth rating: 1

Ratings from Models

XLMR-two-stage: 1
Gemma: 2 (high confidence)
Qwen: 2 (medium confidence)

Turn based LLM argumentation
Gemma decision: 3 confidence: 0.6 (after receiving Qwen: 2)
Gemma reason: The review indicates a mixed experience with the product and delivery, suggesting a neutral to slightly positive overall sentiment despite the issues.

Qwen decision: 3 confidence: 0.6 (after receiving Gemma: 2)
Qwen reason: The review expresses mixed sentiment regarding the product (color good) but also mentions a delivery issue and refund, suggesting a moderate overall experience.

Final Decision
Predicted rating: 3
Ground truth: 1

Explanation Module

Advertised product

Reviewer-uploaded image

The final rating is 3 because the reviewer states "Order colour rose good but Delivery Blu One Refund my money." This indicates a positive experience with the product's color ("rose good") but a negative experience with the delivery and refund process. The coexistence of positive and negative aspects leads to an overall average assessment, despite the ground truth reflecting dissatisfaction.

Figure 4.10: Inference example where the framework integrates conflicting signals from the text classifier and LLMs, resulting in a final rating of 3.

tivates the framework's conflict-resolution mechanism, triggering LLM–LLM argumentation. During this iterative exchange, the two LLMs evaluate their respective confidence scores and textual evidence, ultimately converging on a consensus rating of 4. The framework then adopts this agreed-upon value as the final predicted rating, demonstrating its ability to reconcile differing model opinions through structured argumentation. The explanation module further provides interpretability by highlighting the textual evidence that justifies the decision, emphasizing that the product quality is generally positive but falls short of perfection. This example highlights the effectiveness of the neuro-symbolic framework in mediating disagreements between text-based and generative models, ensuring that the final prediction reflects a balanced assessment while maintaining transparency and grounding in the review content.
Fig 4.10 presents an inference example in which the text classifier XLM-R initially predicts a low rating of 1, whereas both LLMs, Gemma and Qwen, provide slightly

higher preliminary ratings of 2. This discrepancy triggers the LLM–LLM argumentation process, during which the models iteratively exchange and evaluate their confidence scores and textual evidence. Through this structured dialogue, both LLMs adjust their predictions upward, ultimately converging on a rating of 3. The framework then selects this consensus value as the final predicted rating, demonstrating its ability to mediate conflicts and produce a balanced decision that accounts for multiple perspectives. The explanation module further contextualizes the rating by highlighting that while the product's color received positive feedback, there were issues with delivery that negatively affected the user experience. This combination of textual analysis and argumentation results in a moderate overall assessment, illustrating how the neuro-symbolic framework effectively integrates model predictions and interpretable reasoning to provide nuanced, transparent and context-aware ratings.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

We presented DarazMM, a Bangla-centric multilingual and multimodal dataset for five-class star rating prediction, designed to capture the linguistic diversity and visual complexity of real-world e-commerce reviews. By integrating review text with customer-uploaded images, advertised product images and product descriptions, DarazMM enables product-aware multimodal reasoning that is not supported by existing benchmarks. Extensive experiments across zero-shot, few-shot, supervised and neuro-symbolic settings demonstrate that explicitly modeling rating ordinality and cross-modal interactions leads to substantial performance gains over strong baselines.

Despite these contributions, DarazMM has several limitations. The dataset may contain a non-trivial amount of noisy or weak labels due to the large-scale and semi-automated data collection process. Fully annotating or verifying the entire dataset with domain experts was not feasible because of time and resource constraints. Additionally, the dataset exhibits inherent class imbalance, as reviews accompanied by images are disproportionately associated with highly positive ratings. This imbalance may bias models toward majority classes and limit performance on underrepresented rating categories.

5.2 Future Work

Future work includes reducing label noise through large-scale expert verification or advanced automated filtering methods, as well as exploring data balancing strategies to mitigate class skew. DarazMM may also be expanded with additional annotated data to improve coverage and representation of underrepresented rating categories. This additional data will support the development and evaluation of more effective ordinal and multimodal learning approaches for star rating prediction in low-resource Bangla settings. We hope DarazMM will serve as a valuable benchmark for advancing multilingual and multimodal sentiment analysis and star rating prediction research.

Bibliography

[1] F. U. Zaman, M. Z. Hossain, and M. K. Bhuiyan, "An empirical study on the impact of smote on imbalanced text features for review rating prediction in bengali," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, IEEE, 2023.

[2] G. Ganu, N. Elhadad, and A. Marian, "Beyond the stars: Improving rating predictions using review text content.," in *WebDB*, vol. 9, pp. 1–6, 2009.

[3] B. H. Ahmed and A. S. Ghabayen, "Review rating prediction framework using deep learning," *Journal of Ambient Intelligence and Humanized Computing*, vol. 13, no. 7, pp. 3423–3432, 2022.

[4] M. I. Hossain, M. Rahman, M. T. Ahmed, M. S. Rahman, and A. T. Islam, "Rating prediction of product reviews of bangla language using machine learning algorithms," in *2021 International Conference on Artificial Intelligence and Mechatronics Systems (AIMS)*, pp. 1–6, IEEE, 2021.

[5] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," *ACM Transactions on Intelligent Systems and Technology*, vol. 16, no. 5, pp. 1–72, 2025.

[6] K. I. Roumeliotis, N. D. Tselikas, and D. K. Nasiopoulos, "Llms in e-commerce: A comparative analysis of gpt and llama models in product review evaluation," *Natural Language Processing Journal*, vol. 6, p. 100056, 2024.

[7] M. R. A. Rashid, A. Das, K. F. Hasan, M. R. Hasan, M. Sultana, M. Hasan, R. U. Islam, R. A. Tuhin, and M. S. H. Khan, "Bert-kan: Enhancing bilingual sentiment analysis in bangladeshi e-commerce through fine-tuned large language models," *Natural Language Processing Journal*, p. 100190, 2025.

[8] L. Zhao, J. Zhou, J. Cao, and W. Zhu, "Emsa: Explainable multilingual sentiment analysis models providing sentiment analysis across multiple languages," *PLoS One*, vol. 20, no. 11, p. e0333508, 2025.

[9] M. Mohiuddin, "Overview the e-commerce in bangladesh," *IOSR Journal of Business and Management*, vol. 16, no. 7, pp. 01–06, 2014.

[10] S. Ahmed, M. M. Samia, M. H. Sayma, M. M. Kabir, and M. Mridha, "trf-bert: A transformative approach to aspect-based sentiment analysis in the bengali language," *PloS one*, vol. 19, no. 9, p. e0308050, 2024.

[11] A. Bhattacharjee, T. Hasan, W. Ahmad, K. S. Mubasshir, M. S. Islam, A. Iqbal, M. S. Rahman, and R. Shahriyar, "Banglabert: Language model pretraining and benchmarks for low-resource language understanding evaluation in bangla," in *Findings of the Association for Computational Linguistics: NAACL 2022*, pp. 1318–1327, 2022.

[12] M. R. A. Rashid, K. F. Hasan, R. Hasan, A. Das, M. Sultana, and M. Hasan, "A comprehensive dataset for sentiment and emotion classification from bangladesh e-commerce reviews," *Data in Brief*, vol. 53, p. 110052, 2024.

[13] "Daraz bangladesh." https://www.daraz.com.bd. [Accessed 04 February 2026].

[14] "pickaboo." https://www.pickaboo.com. [Accessed 04 February 2026].

[15] M. T. Akter, M. Begum, and R. Mustafa, "Bengali sentiment analysis of e-commerce product reviews using k-nearest neighbors," in *2021 International conference on information and communication technology for sustainable development (ICICT4SD)*, pp. 40–44, IEEE, 2021.

[16] S. S. Shanto, Z. Ahmed, and A. I. Jony, "Mining user opinions: A balanced bangla sentiment analysis dataset for e-commerce," *Malaysian Journal of Science and Advanced Technology*, pp. 272–279, 2023.

[17] M. J. Hossain, D. D. Joy, S. Das, and R. Mustafa, "Sentiment analysis on reviews of e-commerce sites using machine learning algorithms," in *2022 International Conference on Innovations in Science, Engineering and Technology (ICISET)*, pp. 522–527, IEEE, 2022.

[18] D. Q. Nguyen, T. Vu, S. B. Pham, *et al.*, "Sentiment classification on polarity reviews: an empirical study using rating-based features," in *Proceedings of the 5th workshop on computational approaches to subjectivity, sentiment and social media analysis*, pp. 128–135, 2014.

[19] M. A. Shafin, M. M. Hasan, M. R. Alam, M. A. Mithu, A. U. Nur, and M. O. Faruk, "Product review sentiment analysis by using nlp and machine learning in bangla language," in *2020 23rd International Conference on Computer and Information Technology (ICCIT)*, pp. 1–5, IEEE, 2020.

[20] R. Chowdhury, F. U. Zaman, A. Sharker, M. Rahman, and F. M. Shah, "Rate insight: A comparative study on different machine learning and deep learning approaches for product review rating prediction in bengali language," in *2022 25th International Conference on Computer and Information Technology (ICCIT)*, pp. 406–411, IEEE, 2022.

[21] L. Zeng, "Leveraging large language models for code-mixed data augmentation in sentiment analysis," in *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pp. 85–101, 2024.

[22] X. Fang and J. Zhan, "Sentiment analysis using product review data," *Journal of Big data*, vol. 2, no. 1, p. 5, 2015.

[23] L. Yang, Y. Li, J. Wang, and R. S. Sherratt, "Sentiment analysis for e-commerce product reviews in chinese based on sentiment lexicon and deep learning," *IEEE access*, vol. 8, pp. 23522–23530, 2020.

[24] H. He, G. Zhou, and S. Zhao, "Exploring e-commerce product experience based on fusion sentiment analysis method," *Ieee Access*, vol. 10, pp. 110248–110260, 2022.

[25] M. Demircan, A. Seller, F. Abut, and M. F. Akay, "Developing turkish sentiment analysis models using machine learning and e-commerce data," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 202–207, 2021.

[26] F. Barbieri, L. E. Anke, and J. Camacho-Collados, "Xlm-t: Multilingual language models in twitter for sentiment analysis and beyond," in *Proceedings of the thirteenth language resources and evaluation conference*, pp. 258–266, 2022.

[27] M. N. Shamael, S. Nawshin, S. Shatabda, and S. Islam, "Banglishrev: A large-scale bangla-english and code-mixed dataset of product reviews in e-commerce," *arXiv preprint arXiv:2412.13161*, 2024.

[28] Y. Ling, J. Yu, and R. Xia, "Vision-language pre-training for multimodal aspect-based sentiment analysis," in *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: long papers)*, pp. 2149–2159, 2022.

[29] H. Zhang, Y. Wang, G. Yin, K. Liu, Y. Liu, and T. Yu, "Learning language-guided adaptive hyper-modality representation for multimodal sentiment analysis," in *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 756–767, 2023.

[30] F. Koto, T. Beck, Z. Talat, I. Gurevych, and T. Baldwin, "Zero-shot sentiment analysis in low-resource languages using a multilingual sentiment lexicon," in *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 298–320, 2024.

[31] P. Quoc-Hung, D. Van-Dan, L. Huu-Loi, L. Thi-Viet-Huong, N. Thu-Ha, P. Xuan-Hieu, N. Minh-Tien, and P. Ngoc-Hung, "Aspect-based sentiment analysis of clothing reviews in vietnamese e-commerce," in *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pp. 231–238, 2024.

[32] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," in *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3881–3906, 2024.

[33] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: A survey," *Displays*, vol. 80, p. 102563, 2023.

[34] "Selenium." https://www.selenium.dev/. [Accessed 04 February 2026].

[35] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th annual meeting of the association for computational linguistics*, pp. 8440–8451, 2020.

[36] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[37] V. Mirjalili, "Postln, preln and residual transformers — eliminating the need for warm-up stage for training transformers." https://pyml.substack.com/p/postln-preln-and-residual-transformers. PyML Studio, [Accessed 04 February 2026].

[38] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[39] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*, pp. 8748–8763, PmLR, 2021.

[40] S. Agarwal, L. Ahmad, J. Ai, S. Altman, A. Applebaum, E. Arbus, R. K. Arora, Y. Bai, B. Baker, H. Bao, *et al.*, "gpt-oss-120b and gpt-oss-20b model card," *arXiv preprint arXiv:2508.10925*, 2025.

[41] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[42] M. Tschannen, A. Gritsenko, X. Wang, M. F. Naeem, I. Alabdulmohsin, N. Parthasarathy, T. Evans, L. Beyer, Y. Xia, B. Mustafa, *et al.*, "Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features," *arXiv preprint arXiv:2502.14786*, 2025.

[43] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[44] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, Z. Cheng, L. Deng, W. Ding, C. Gao, C. Ge, W. Ge, Z. Guo, Q. Huang, J. Huang, F. Huang, B. Hui, S. Jiang, Z. Li, M. Li, M. Li, K. Li, Z. Lin, J. Lin, X. Liu, J. Liu, C. Liu, Y. Liu, D. Liu, S. Liu, D. Lu, R. Luo, C. Lv, R. Men, L. Meng, X. Ren, X. Ren, S. Song, Y. Sun, J. Tang, J. Tu, J. Wan, P. Wang, P. Wang, Q. Wang, Y. Wang, T. Xie, Y. Xu, H. Xu, J. Xu, Z. Yang, M. Yang, J. Yang, A. Yang, B. Yu, F. Zhang, H. Zhang, X. Zhang, B. Zheng, H. Zhong, J. Zhou, F. Zhou, J. Zhou, Y. Zhu, and K. Zhu, "Qwen3-vl technical report," 2025.

[45] G. Team, "Gemma 3n," 2025.

[46] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, *et al.*, "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[47] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[48] C. M. Bishop and N. M. Nasrabadi, *Pattern recognition and machine learning*, vol. 4. Springer, 2006.