Leveraging Large Language Models for Marketing Analytics

Tommaso Mazzarini
Politecnico di Torino
Turin, Italy

Alessio Gioè
Politecnico di Torino
Turin, Italy

Vincenzo Catalano
Politecnico di Torino
Turin, Italy

Abstract

Analyzing social media comments and reviews is vital for companies aiming to evaluate the true impact of product launches. This paper introduces an AI-driven pipeline designed to convert online feedback about Lavazza products into actionable marketing insights using a multi-level classification framework powered by Large Language Models (LLMs). The proposed architecture processes feedback through four key stages: product relevance classification, multi-topic sentiment detection, overall sentiment analysis, and star rating prediction. Extensive experiments were conducted to evaluate the performance of general-purpose LLMs (Gemma-2, Llama-3.1, Mistral) under zero-shot and few-shot configurations, as well as specialized BERT-based models. The results demonstrated high performance across tasks, achieving 82.6% accuracy in product relevance classification, F1 scores of 89.9% and 74.5% for positive and negative multi-topic sentiment detection, 87.7% overall sentiment accuracy, and 69.2% star rating accuracy, with a 95.5% accuracy rate when allowing a one-star margin of error. While general-purpose LLMs delivered impressive accuracy, they incurred significantly higher computational costs compared to specialized models. These findings underscore the potential of LLM-based pipelines for marketing analytics, while highlighting the need to balance performance with computational efficiency for practical deployment.

1 Introduction

In today's digital era, understanding consumer responses to product launches is essential for business success. Social media platforms and review websites generate a wealth of user-generated feedback, offering valuable insights that can shape marketing strategies. However, the high volume and unstructured nature of this data make manual analysis both time-consuming and impractical.

This research addresses the challenge by developing an AI-driven pipeline that leverages Large Language Models (LLMs) to analyze consumer feedbacks. The pipeline employs a multi-level classification framework to process social media comments and product reviews systematically. It begins with product relevance classification to filter out unrelated comments. Next, it performs topic extraction with sentiment tagging, followed by overall sentiment analysis and star rating prediction. The outputs of these classification tasks are aggregated into Key Performance Indicators (KPIs), which are designed to provide actionable insights for marketing teams.

This structured approach transforms raw text data into clear, interpretable metrics that enable businesses to evaluate product performance, track consumer reception, and guide future strategic decisions.

Our methodology combines zero-shot and few-shot prompting techniques to test the capabilities of various LLM architectures in marketing analytics tasks. Specifically, we evaluate general-purpose models such as Gemma-2, Llama-3.1, and Mistral, alongside specialized BERT-based models, to identify the optimal configuration for performance and efficiency.

The remainder of this paper is organized as follows:

- **Section 2: Related Work** discusses the evolution of Large Language Models, along with an overview of zero-shot and few-shot prompting techniques.
- **Section 3: Methodology** outlines the problem statement, describes the data collection and preprocessing steps, details the multi-level classification architecture, and explains the KPI extraction process.
- **Section 4: Experiments and Evaluation** presents the datasets, experimental setup, evaluation metrics, and performance analysis.
- **Section 5: Conclusion** summarizes the study's findings on LLMs for review analysis and proposes future enhancements.

2 Related Work

This section provides an overview of key research in natural language processing (NLP), with a focus on Large Language Models (LLMs) and specialized models. We discuss their capabilities, applications across various tasks, and their relevance for extracting insights from consumer-generated content.

2.1 Large Language Models (LLMs)

Large Language Models (LLMs) have become essential in natural language processing (NLP). Pre-trained on extensive text datasets, they are versatile tools capable of supporting a wide range of applications, making them particularly well-suited for analyzing product feedbacks.

2.1.1 ***General-Purpose Models.*** General-purpose LLMs handle a variety of NLP tasks. These models can be adapted to new tasks easily, making them ideal for scalable applications.

Key general-purpose models include:

- **Gemma2**: Developed by Google, this model excels in NLP tasks and supports multiple languages, making it ideal for global applications [3].
- **Llama3.1**: Meta's Llama model is excellent for handling large datasets and complex tasks, such as analyzing massive text volumes [5].
- **Mistral**: A compact model optimized for text analysis, Mistral is perfect for fast, efficient processing of user-generated content [1].

2.1.2 ***Specialized Models.*** Specialized LLMs are fine-tuned for specific tasks, often outperforming general models in those spefic areas. These models are ideal for tasks like sentiment analysis or rating prediction, offering higher accuracy by understanding domain-specific language.

Examples of specialized models include:

- *DistilBERT-base-multilingual-cased-sentiments*: A lightweight BERT model fine-tuned for multilingual sentiment analysis [4].
- *Twitter-XML-RoBERTa-base-sentiment-finetuned* and *Twitter-RoBERTa-base-sentiment-latest*: Optimized for analyzing sentiment in social media content [2].
- *Bert-base-multilingual-uncased-sentiment*: A BERT model fine-tuned for multilingual star rating prediction [6].
- *Multilingual-sentiment-analysis*: Fine-tuned for star rating prediction across multiple languages [8].

Specialized models provide higher accuracy for specific tasks but are limited in their application.

Both general-purpose and specialized models are valuable for analyzing consumer-generated content. General-purpose models like Gemma2 and Llama3.1 offer a flexible foundation for tasks related to NLP. Specialized models, on the other hand, provide deep domain expertise for specific tasks like sentiment classification and rating prediction. By integrating both, we can ensure efficient, accurate analysis of large volumes of content, helping businesses understand customer sentiment and preferences.

2.2 Prompt Engineering

Prompt engineering is a key technique for adapting general purpose Large Language Models (LLMs) to specific tasks. It involves crafting input prompts that effectively guide models, using approaches such as zero-shot and few-shot prompting. These methods allow LLMs to perform diverse tasks without requiring extensive task-specific training, enhancing their flexibility and efficiency in natural language processing applications.

Key studies [7] have highlighted two main strategies:

- **Zero-shot prompting**: In this approach, the model is provided with detailed task instructions but no examples of the expected output. It relies on the model's ability to generalize and infer the task requirements from the instructions alone. Zero-shot prompting is particularly useful for rapid prototyping or when annotated data is unavailable.
- **Few-shot prompting**: Here, task instructions are supplemented with a small number of curated examples that illustrate the desired input-output relationship. These examples help the model better understand nuanced or domain-specific tasks, significantly improving its performance. Few-shot prompting is especially effective in cases where minimal but high-quality examples are available.

In our study, we designed prompts tailored to each classification task in the pipeline. The prompts followed a structured format to maximize their effectiveness:

(1) **Task Definition**: The prompt begins by clearly defining the task, including explicit instructions. For example, the model is instructed to classify feedback as product-relevant or not, assign sentiment to specific topics, or predict a star rating.
(2) **Input Data**: The text to be analyzed, such as consumer feedback, is included as the main input for the model to process.
(3) **Examples (Few-shot only)**: In the few-shot approach, the prompt includes eight carefully chosen examples for each task. These examples illustrate the correct input-output mapping and help the model understand the task context. For zero-shot prompting, this step is omitted.
(4) **Expected Output Format**: The prompt specifies the format of the model's response, ensuring consistency and adherence to requirements. For instance, the output might be a single label (e.g., 1 or 0 for relevant comment or not), a sentiment value (e.g., "Positive," "Negative," or "Neutral"), or a star rating (ranging from 1 to 5).

By following this structured prompt design, we ensured that the model clearly understood the task, the input to analyze, and the expected output format. This approach allowed us to systematically evaluate the effectiveness of zero-shot and few-shot prompting across multiple LLMs, optimizing their performance for the tasks in our pipeline.

3 Method

This section provides a detailed overview of the methodological framework adopted for this study. We outline the research problem and associated questions which we want to find an answer for. We then describe the data collection and preprocessing techniques employed. Next, we present the multi-level classification architecture at the heart of our analysis, providing a detailed overview of both the KPI extraction pipeline and the classification block. In conclusion, we discuss about the models that we are going to test.

3.1 Problem Statement

The primary research question guiding this study is: Can Large Language Models (LLMs) be utilized to extract actionable insights from consumer comments and reviews to enhance marketing strategies? Specifically, we investigate:

- How effectively LLMs identify product-relevant comments.
- The extent to which LLMs extract product aspects from reviews and comments, and assign sentiments to them.
- How successfully LLMs assign sentiment to comments and translate them into rating stars for reviews.

3.2 Data Collection and Preprocessing

To address the limitations of existing datasets, we began with an initial dataset of 5,217 reviews provided by Lavazza. This dataset was originally collected for a previous market analysis conducted by Workflow, a long-time partner of Lavazza. Additionally, a second dataset provided by Lavazza, created by their partner Digimind, contained 476 promotional posts and tweets, which were not actual consumer feedback and were thus excluded from our analysis. To enrich the dataset, we then collected 577 comments from social media platforms such as Instagram (84 comments) and YouTube (493 comments), utilizing Apify and Google APIs, respectively.

After the data collection phase, we proceeded to the preprocessing stage, which included the following steps:

- **Data Cleaning:** Duplicate entries and reviews with null values were removed, reducing the number of usable reviews from 5,217 to 2,930.
- **Text Normalization:** The text was standardized by removing special characters, extra spaces, and ensuring consistent formatting across all entries.
- **Topic Identification:** In collaboration with Lavazza's marketing team, we categorized 103 unique product aspects from the reviews into 9 meaningful categories (*topicClass*):
  - *Coffee Quality*
  - *Machine Performance*
  - *Capsules and Compatibility*
  - *Design and Materials*
  - *User Experience*
  - *Maintenance and Durability*
  - *Services and Logistics*
  - *Environmental Impact*
  - *Price and Worth*

  This process ensured that each aspect was mapped to its relevant category, providing a structured foundation for the following analysis.
- **Translation:** Non-English reviews and comments were translated into English, followed by additional cleaning to maintain consistency across languages.

3.3 Multi-Level Classification Architecture

The core of our methodology is a multi-level classification architecture (Figure 1) designed to process preprocessed reviews and comments through a series of hierarchical classification tasks, each focused on extracting meaningful insights from the textual data.

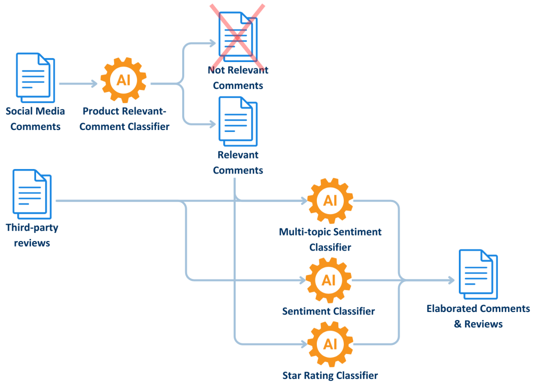

Figure 1: Multi-Level Classification Architecture.

The architecture consists of the following steps:

- **Product-Relevant Comment Classification:** This is a binary classification task that serves as the first stage in our pipeline. The goal is to filter out irrelevant content by identifying which comments are specifically related to the product.
- **Multi-Topic Sentiment Classification:** This is a multi-label classification task with non-mutually exclusive labels. The system identifies various aspects or features of the product mentioned in the comments, such as "Coffee Quality," "Machine Performance," "User Experience," or "Design and Materials." For each detected aspect, the model assigns a positive or negative sentiment label. Additionally, it identifies the primary positive and negative aspects, if present, to enable more precise analysis.
- **Overall Sentiment Classification:** This is a multi-label classification task with mutually exclusive labels. Following the aspect-based sentiment analysis, this classifier determines the overall sentiment of a comment. The system assesses whether the general tone of the feedback is positive, negative, or neutral.
- **Star Rating Prediction:** This is another multi-label classification task with mutually exclusive labels, similar to the previous one. In this final step, the system predicts a star rating (ranging from 1 to 5 stars) for each review. The rating is derived from the sentiment and tone expressed in the review, providing a quantitative measure that reflects the overall perception of the product.

Each of these tasks contributes to a comprehensive and layered understanding of customer feedback, enabling more targeted insights into product performance, customer satisfaction, and areas for improvement.

3.4 KPI Pipeline

The final stage of our methodology is the KPI extraction pipeline (Figure 2). This pipeline utilizes predefined functions to derive key performance indicators (KPIs) that are critical for evaluating the product launch. It processes classified data (elaborated comments and reviews) to provide actionable metrics, supporting future strategic decision-making.

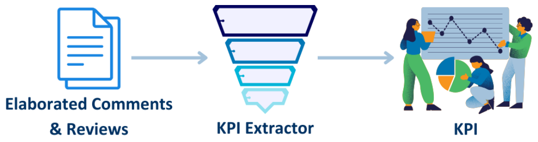

Figure 2: Pipeline for KPI Extraction.

3.5 Classification Block

The classification block is a fundamental component of the multi-level classification architecture, and its structure is adapted to the model used for each specific task, as illustrated in Figure 3.

- **General-Purpose Models:** For general-purpose models, the classification block consists of the model along with a task-specific prompt template. Depending on the task to be performed, we employ different prompting strategies, such as few-shot or zero-shot learning, to assist the model in generating accurate predictions.
- **Specialized Models:** For task-specific models, the classification block contains only the model itself. These models are fine-tuned for particular tasks, meaning they do not require

additional prompting, as they are already optimized for the target task.

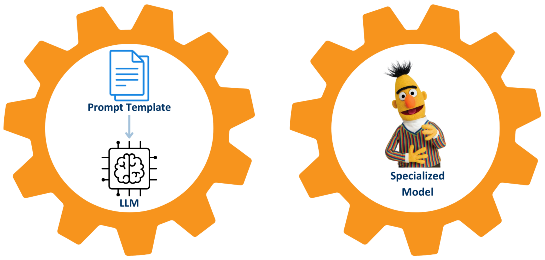

Figure 3: Illustration of the classification blocks: the block on the left corresponds to general-purpose models, while the block on the right represents specialized models.

3.6 Models

For this study, we employed two categories of models to address the various classification tasks:

- **General-Purpose Models:** This category includes distilled versions of the models introduced in Section 2.1. Specifically, we utilized *Gemma2* (9B parameters), *Llama3.1* (9B parameters), and *Mistral* (7B parameters). These models are designed for general-purpose natural language processing (NLP) tasks, offering a balance between performance and computational efficiency.
- **Specialized Models:** For more specific tasks such as sentiment classification and star rating prediction, we used fine-tuned versions of BERT-based models available on Hugging Face. These models are optimized for domain-specific applications, as outlined below:
  - *distilbert-base-multilingual-cased-sentiments (DB-sentimet)*: This multilingual model was fine-tuned for sentiment analysis across various languages, enabling it to effectively analyze the sentiment of customer reviews and comments in different linguistic contexts.
  - *twitter-xlm-roberta-base-sentiment-finetuned (TR-sentimet)*: This model is specifically optimized for sentiment analysis in social media content, making it well-suited for tasks involving customer feedback on social platforms, where informal language and short text are common.
  - *twitter-roberta-base-sentiment-latest (TRL-sentimet)*: An updated version of the Twitter-RoBERTa model, which was fine-tuned to perform sentiment analysis with high accuracy, especially for detecting positive, negative, and neutral sentiments in reviews and comments.
  - *bert-base-multilingual-uncased-sentiment (BBM-rating)*: A fine-tuned version of the BERT model, particularly effective for predicting star ratings in reviews based on sentiment and other contextual clues in the text.
  - *multilingual-sentiment-analysis (MLA-rating)*: This multilingual model was fine-tuned to predict star ratings (from 1 to 5 stars) based on the sentiment expressed in customer reviews.

4 Experiment

In this section, we will discuss the experiments conducted to evaluate the performance of the adopted method. Specifically, we will present the datasets used for the experiments, the experimental setup, and finally, we will showcase the evaluation metrics and discuss the results.

4.1 Dataset

The datasets used in this study are categorized into **reviews**, **comments**, and **synthetic data**.

After the preprocessing stage, the reviews dataset contains a total of 2,930 records. However, to reduce the execution time of the experiments, we selected a sample of 500 reviews.

For the purposes of these experiments, we focused on the following features, which were critical for analyzing the review data:

- **feedbackRating**: This feature represents the review score, ranging from 1 to 5, and is present in 100% of the reviews. The distribution of **feedbackRating** across the dataset is illustrated in Figure 4:

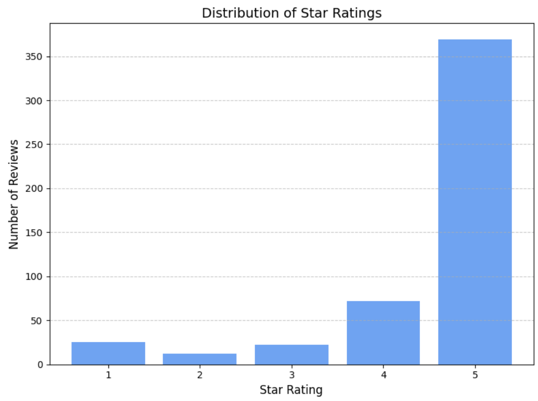

Figure 4: Distribution of Feedback Ratings

  As shown in the figure, the distribution is heavily skewed toward higher ratings. This trend is typical in customer feedback datasets, reflecting a natural tendency for users to provide more positive feedback.
- **positiveTopicClass** and **negativeTopicClass**: These features highlight the positive and negative aspects, respectively, that emerge from the review. In particular, 95% of the reviews present at least one aspect as a positive or negative topic. However, this percentage can be misleading, as upon inspecting some of the feedbacks, we noticed that the labels were incomplete and did not capture all the positive and negative aspects emerging from the reviews. For this reason, we decided to enrich the labels synthetically.

In addition to the reviews, we also collected 577 comments from Instagram and YouTube. Unfortunately, these comments are unlabeled, which means they cannot be used directly for model evaluation. To address this, we generated synthetic data for the comments.

To tackle the labeling challenges in both the reviews and the comments, we leveraged ChatGPT-4, a powerful text generation model.

For the reviews, we generated labels for the **positiveTopicClass** and **negativeTopicClass** based on the 500 reviews we analyzed. For the comments, we generated 236 new labeled comments, including **relevantLabel**, **positiveTopicClass**, **negativeTopicClass**, and **sentiment**.

Below we can see the distributions of the **relevantLabel** and **sentiment** for the comments (Figure 5 and Figure 6), as well as the distributions of **positiveTopicClass** and **negativeTopicClass** for both comments and reviews (Figure 7).

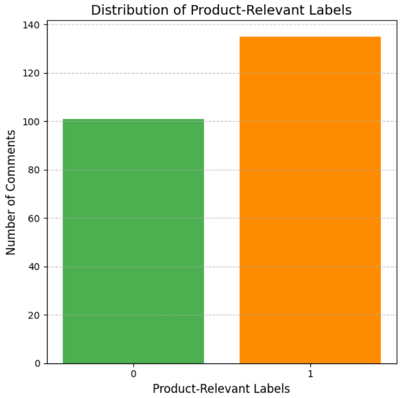

Figure 5: Distribution of Relevant Labels for Comments

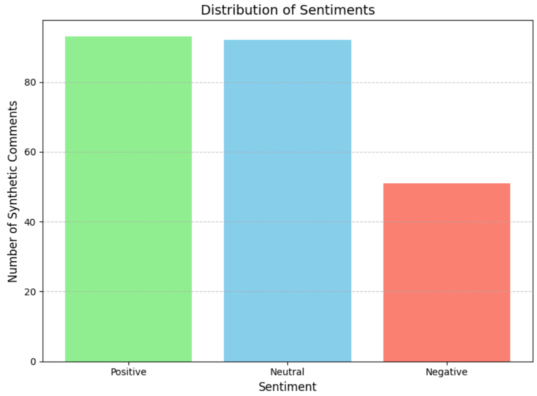

Figure 6: Distribution of Sentiment Labels for Comments

As observed in the first two graphs (Figures 5 and 6), the distribution of **relevantLabel** and **sentiment** in the dataset is relatively balanced. However, when examining the distribution of **positiveTopicClass** and **negativeTopicClass** across both comments and reviews (Figure 7), we can see that the predominant class is **NONE**.

4.2 Experiment Configuration

The experiments were conducted using a Kaggle notebook equipped with two NVIDIA Tesla T4 GPUs.

In total, four experiments were performed, each corresponding to one of the following tasks: **product-relevant comment classification**, **multi-topic sentiment classification**, **sentiment classification,** and **star rating prediction**. For each task, we carefully selected a batch size that balanced computational constraints with execution time. Both general-purpose models and task-specific models were evaluated. For the general-purpose models, we tested both zero-shot and few-shot prompting configurations.

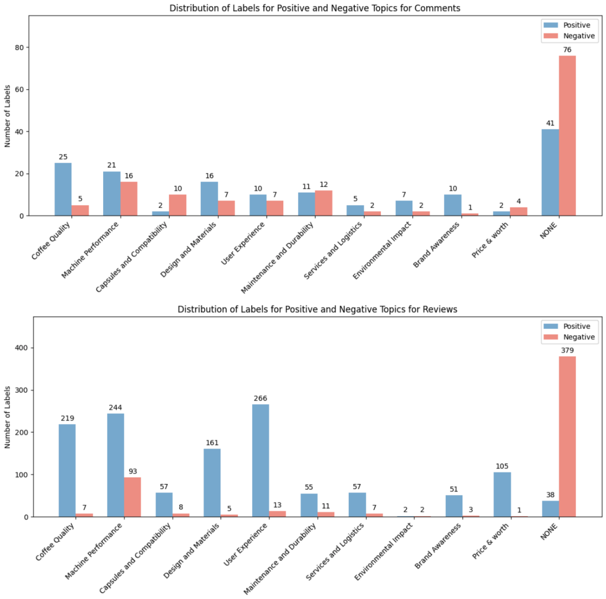

Figure 7: Distribution of Positive and Negative Topic Classes for Comments and Reviews

The specific configurations for each task are outlined below:

- **Product-Relevant Comment Classification**: The batch size was set to 15. The models tested included Gemma2, Llama3.1, and Mistral, using both zero-shot and few-shot prompting configurations.
- **Multitopic Sentiment Classification**: Given the higher complexity of this task, the batch size was reduced to 5. The models tested were Gemma2, Llama3.1, and Mistral, with both zero-shot and few-shot prompting configurations.
- **Sentiment Classification**: For general-purpose models, the batch size was set to 20, while for task-specific models, it was increased to 50. General-purpose models included Gemma2, Llama3.1, and Mistral, with both prompting configurations.
- **Star Rating Classification**: Similar to sentiment classification, the batch size for general-purpose models was set to 20, and for task-specific models, it was 50. The general-purpose models tested included Gemma2, Llama3.1, and Mistral, using both zero-shot and few-shot prompting configurations.

4.3 Evaluation

To evaluate the models on the various tasks, we focused on specific evaluation metrics tailored to each task's characteristics.

For **binary classification tasks**, such as product-relevant comment classification, we used accuracy, recall, precision, and F1-score as the key metrics:

- **Accuracy:** Measures the proportion of correctly classified instances.

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

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives.

- **Precision:** The proportion of positive predictions that are actually correct.

$$\text{Precision} = \frac{TP}{TP + FP}$$

- **Recall:** The proportion of actual positives that are correctly identified.

$$\text{Recall} = \frac{TP}{TP + FN}$$

- **F1-Score:** The harmonic mean of precision and recall, providing a balance between the two.

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

For **multi-label classification tasks**, such as multi-topic sentiment classification, in addition to accuracy, we employed weighted precision, weighted recall, and weighted F1-score:

- **Weighted Precision:** The average precision across all labels, weighted by their frequency in the dataset.

$$\text{Weighted Precision} = \frac{\sum_i w_i \times \text{Precision}_i}{\sum_i w_i}$$

where w_i is the weight of label i, and Precision_i is the precision for that label.

- **Weighted Recall:** The average recall across all labels, weighted by their frequency.

$$\text{Weighted Recall} = \frac{\sum_i w_i \times \text{Recall}_i}{\sum_i w_i}$$

- **Weighted F1-Score:** The average F1-score across all labels, weighted by their frequency.

$$\text{Weighted F1-Score} = \frac{\sum_i w_i \times \text{F1-Score}_i}{\sum_i w_i}$$

For **Star Rating Prediction**, we also consider the **"off-by-one" accuracy**, which measures how often the predicted star rating is within one point of the true rating.

$$\text{Off-by-one Accuracy} = \frac{\sum_i \mathbb{I}(|\hat{y}_i - y_i| \leq 1)}{N}$$

where $\hat{y}_i$ is the predicted star rating, y_i is the true star rating, and $\mathbb{I}(\cdot)$ is the indicator function, which returns 1 if the condition is true and 0 otherwise. N represents the total number of reviews.

Additionally, another important metric considered was **inference time**, which reflects the models' efficiency in processing data. This metric measures the time taken by the model to make predictions on the dataset, providing insight into the computational cost of the models.

The results of the experiments are summarized in the following tables (Tables 1, 2, 3, and 4), categorized by task.

Model	Acc.	Pre.	Rec.	F1	Time (s)
Gemma 0-shot	0.758	0.738	0.896	0.809	3.40
Gemma Few-shot	0.814	0.783	0.933	0.851	3.63
Llama 0-shot	0.763	0.737	0.911	0.815	**2.77**
Llama Few-shot	**0.826**	**0.794**	**0.941**	**0.861**	2.93
Mistral 0-shot	0.758	0.747	0.874	0.805	3.25
Mistral Few-shot	0.797	0.764	0.933	0.840	3.53

Table 1: Metrics and Execution Times for Models for Product Relevant-Comment Classification

4.4 Observations

From the results in the tables (Tables 1, 2, 3, and 4), several key observations can be made for each task:

- **Product-Relevant Comment Classification**: In this task, the few-shot configuration consistently outperforms the zero-shot configuration across all models. Notably, Llama with few-shot prompting achieves the highest accuracy of 0.826, while maintaining a relatively low inference time (2.93 seconds) as seen in Table 1. This demonstrates that, while few-shot models require more computational resources, they provide significantly better performance compared to zero-shot models, especially in terms of accuracy and recall.
- **Multi-topic Sentiment Classification**: All three models (Gemma, Llama, Mistral) perform well in both configurations, with Gemma Few-shot emerging as the best overall performer, as shown in Table 2. Gemma Few-shot shows superior performance, particularly in terms of F1 score (Positive-Weighted) and Precision (Positive-Weighted). The higher inference time (12.82 seconds) is a trade-off for its exceptional classification of both positive and negative topics. The models' ability to classify both positive and negative aspects, as demonstrated in the weighted metrics, indicates that Gemma Few-shot is highly capable of handling multi-topic sentiment classification.
- **Sentiment Classification**: The results for sentiment classification (Table 3) show a notable advantage for the general-purpose models, particularly Mistral and Gemma, in terms of weighted metrics. Mistral 0-shot performs well in all areas, though its inference time is slightly longer compared to the other models (4.75 seconds). Models with task-based architectures, such as DB-sentimet, provide significantly lower inference times but have much lower accuracy and F1 scores, indicating a trade-off between speed and accuracy. The best compromise between accuracy and inference speed is TRL-sentimet, which has an F1 score roughly 4% lower than Mistral 0-shot but significantly reduced inference time ($8.41{\times}10^{-3}$ seconds).
- **Star Rating Classification**: In this task, task-based models not only exhibit significantly faster inference times compared to general-purpose models but, in the case of BBM-rating, also outperform them across almost all metrics. Specifically, BBM-rating achieves the highest accuracy (0.692),

Model	F1 (Neg.)	F1 (Pos.)	Prec. (Neg.)	Prec. (Pos.)	Rec. (Neg.)	Rec. (Pos.)	Time (s)
Gemma 0-shot	0.895	0.732	0.901	0.834	0.900	0.705	12.10
Gemma Few-shot	**0.899**	0.732	**0.905**	**0.844**	**0.904**	0.694	12.82
Llama 0-shot	0.880	**0.745**	0.901	0.795	0.879	0.718	**8.27**
Llama Few-shot	0.878	0.730	0.895	0.811	0.874	0.694	9.55
Mistral 0-shot	0.873	0.738	0.885	0.761	0.873	**0.735**	11.12
Mistral Few-shot	0.880	0.715	**0.905**	0.789	0.873	0.679	20.96

Table 2: Weighted Metrics and Execution Times for Models for Multi-topic Sentiment Classification

Model	Acc.	Prec.	Rec.	F1	Time (s)
Gemma 0-shot	0.869	0.878	0.869	0.866	3.75
Gemma Few-shot	0.852	0.856	0.852	0.849	3.95
Llama 0-shot	0.818	0.818	0.818	0.815	2.76
Llama Few-shot	0.814	0.825	0.814	0.812	2.96
Mistral 0-shot	**0.877**	**0.880**	**0.877**	**0.876**	4.75
Mistral Few-shot	0.860	0.860	0.860	0.859	3.71
DB-sentimet	0.496	0.436	0.496	0.388	$\mathbf{6.52} \times 10^{-3}$
TR-sentimet	0.682	0.751	0.682	0.680	6.77×10^{-3}
TRL-sentimet	0.831	0.837	0.831	0.829	8.41×10^{-3}

Table 3: Weighted Metrics and Execution Times for Models for Sentiment Classification

Model	Acc.	Prec.	Rec.	F1	Off-1 Acc.	Time (s)
Gemma 0-shot	0.596	0.763	0.596	0.650	0.908	8.07
Gemma Few-shot	0.592	0.765	0.592	0.647	0.908	8.17
Llama 0-shot	0.682	0.764	0.682	0.713	0.948	5.73
Llama Few-shot	0.682	**0.777**	0.682	0.715	0.942	5.92
Mistral 0-shot	0.498	0.741	0.498	0.567	0.912	6.12
Mistral Few-shot	0.434	0.732	0.434	0.510	0.902	6.45
BBM-rating	**0.692**	0.757	**0.692**	**0.718**	**0.955**	0.158
MLA-rating	0.552	0.706	0.552	0.606	0.845	$\mathbf{10.45} \times 10^{-3}$

Table 4: Weighted Metrics, Accuracy, Off-by-One Accuracy, and Execution Times for Models for Star Rating Classification

weighted F1 score (0.718), weighted recall (0.692) and off-by-one accuracy (0.955), as shown in Table 4. The only exception is the weighted precision, where BBM-rating lags behind Llama Few-shot by approximately 2 percentage points. Despite this, the overall performance of BBM-rating, combined with its fast inference time of 0.158 seconds, makes it a highly competitive choice for star rating classification.

After evaluating the models across different tasks, we decided to use *Gemma Few-shot* for product-relevant comment classification and multi-topic sentiment classification, due to its excellent overall performance in both tasks. For sentiment and star rating classification, we opted for task-based models, such as *Twitter RoBERTa latest* for sentiment classification and *BERT base multilingual* for star rating prediction, which offer very high performance with significantly reduced inference time.

Conclusion

This study investigated how Large Language Models (LLMs) can extract valuable insights from consumer reviews to optimize marketing strategies, with a focus on the launch of Lavazza products.

An AI-driven pipeline was designed, incorporating a multi-level classification approach for tasks including product relevance detection, topic extraction, sentiment analysis, and star rating prediction. The evaluation covered both general-purpose LLMs (*Gemma-2*, *Llama-3.1*, *Mistral*) and specialized models (*DistilBert*, *Twitter-RoBERTa*) across zero-shot and few-shot configurations.

Key findings revealed that *Gemma Few-shot* delivered superior performance in product relevance and multi-topic sentiment analysis. Meanwhile, specialized models like *Twitter-RoBERTa latest* and *BERT base multilingual* achieved faster inference times for sentiment and star rating predictions, striking a balance between accuracy and efficiency.

Future work could explore advanced prompting methods, particularly chain-of-thought reasoning, to enhance model accuracy by enabling more complex and structured analysis of customer feedback. These improvements aim to refine the analysis process and support more impactful, data-driven marketing strategies.

References

[1] Mistral AI. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[2] cardiffnlp. 2025. Twitter xlm roberta base sentiment latest. https://huggingface.co/cardiffnlp/twitter-roberta-base-sentiment-latest

[3] Google DeepMind. 2024. Gemma 2: Improving Open Language Models at a Practical Size. arXiv:2408.00118 [cs.CL] https://arxiv.org/abs/2408.00118

[4] lxyuan. 2024. Distilbert base multilingual cased sentiments student. https://huggingface.co/lxyuan/distilbert-base-multilingual-cased-sentiments-student

[5] Meta. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[6] nlptown. 2024. Distilbert base multilingual uncased sentiments. https://huggingface.co/nlptown/bert-base-multilingual-uncased-sentiment

[7] OpenAI. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL] https://arxiv.org/abs/2005.14165

[8] tabularisai. 2023. Multilingual sentiment analysis. https://huggingface.co/tabularisai/multilingual-sentiment-analysis