Lavazza Project

Table of Contents

Checkpoint 3

- Value Proposition
- Project Goal
- Research Questions
- Method
- Experiments
- Conclusion

Project Value Proposition

For **marketers** seeking deeper **insights** on product launches, our **analytics software** transforms raw online comments and reviews into **actionable metrics**.

Project Goal

Developing an **application software** in order to understand the **real impact of a product launch**, in terms of **consumers sentiment** crossing social media **comments** and third-party **reviews**.

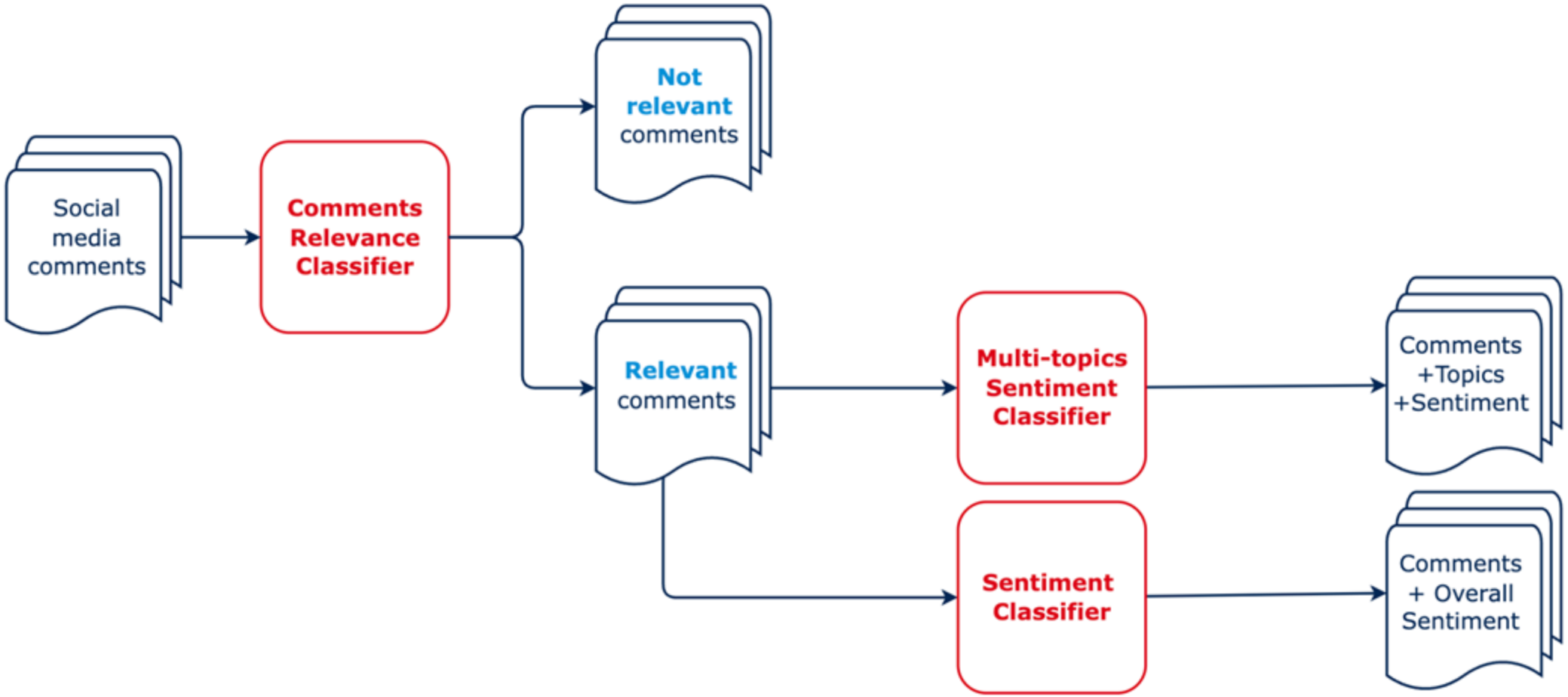
Research questions

Can **Large Language Models** be used to extract useful insights from consumers comments and reviews to improve marketing strategies?

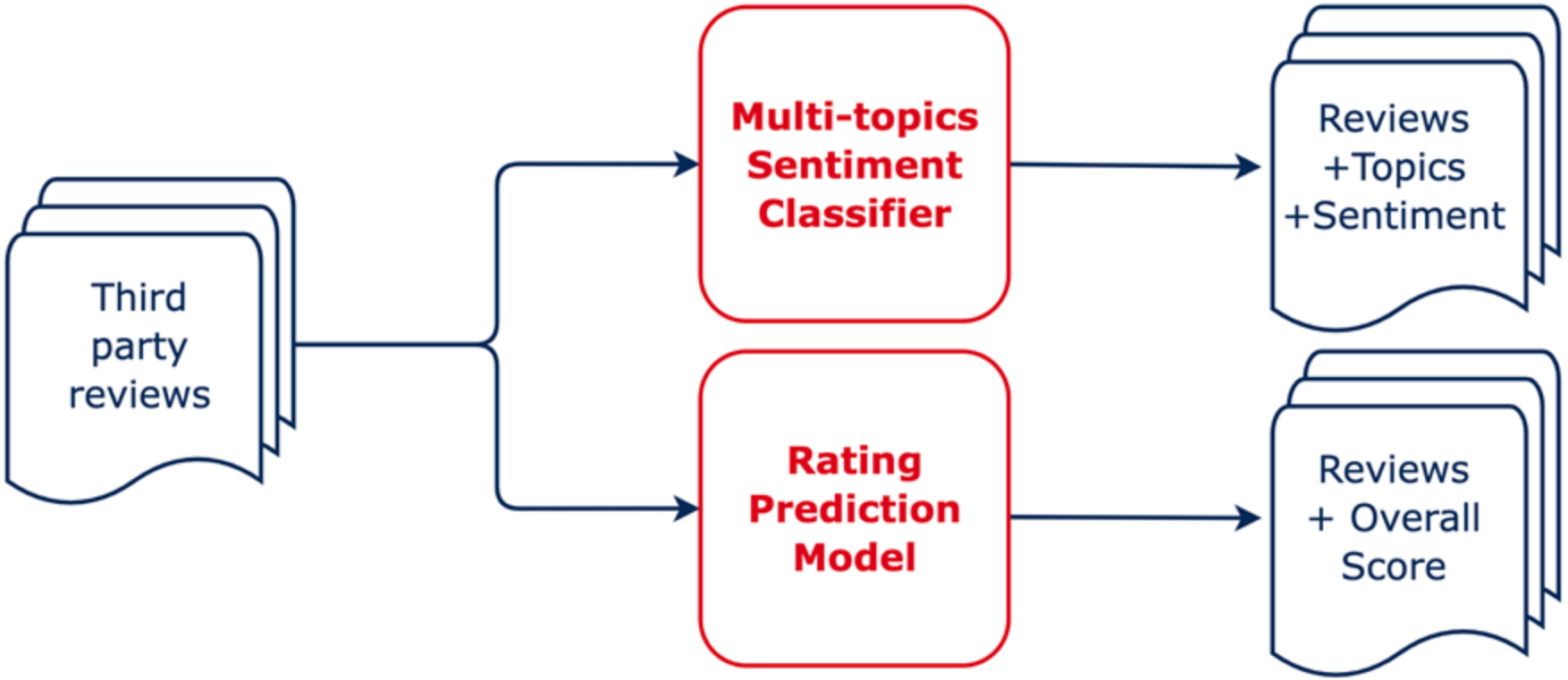
In particular:

1. How effectively can **LLMs** identify **product-relevant comments**?
2. To what extent **LLMs** are able to **extract product aspects** from reviews and comments and **assign sentiments** to them?
3. How successfully are **LLMs** able to **assign** a **sentiment** to comments and **rating stars** to reviews?

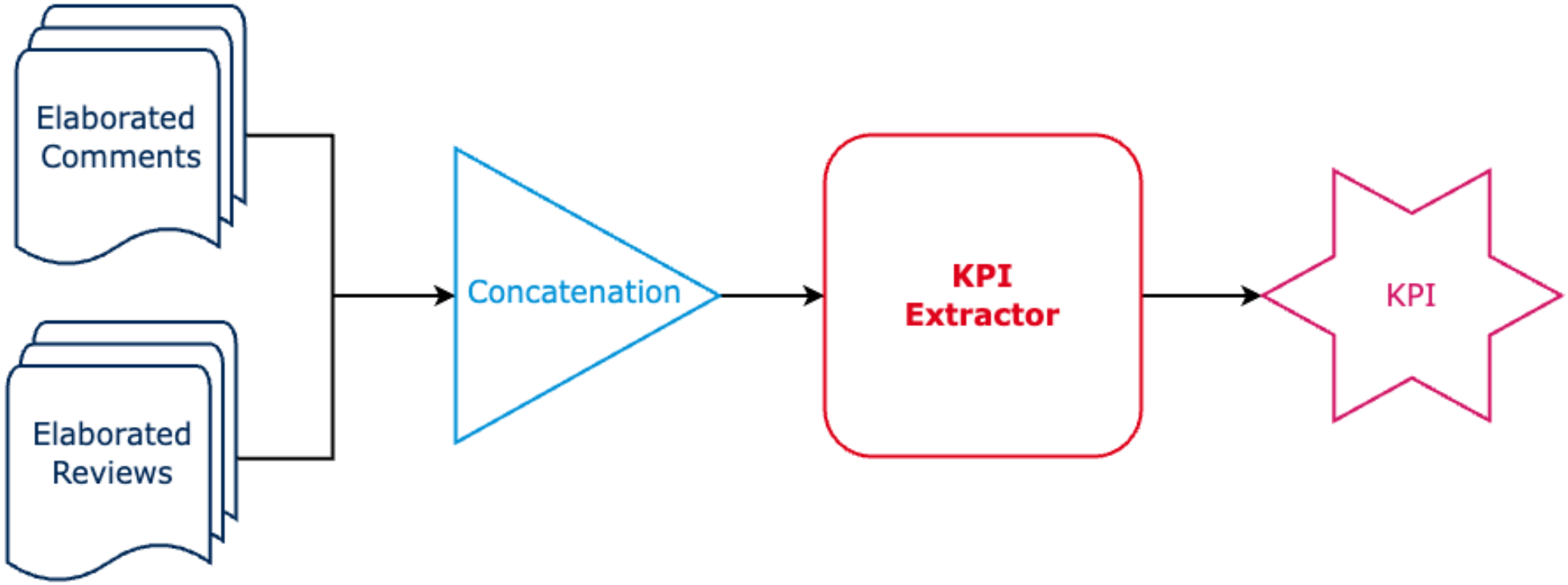
Method - Comments Pipeline



Method - Reviews Pipeline

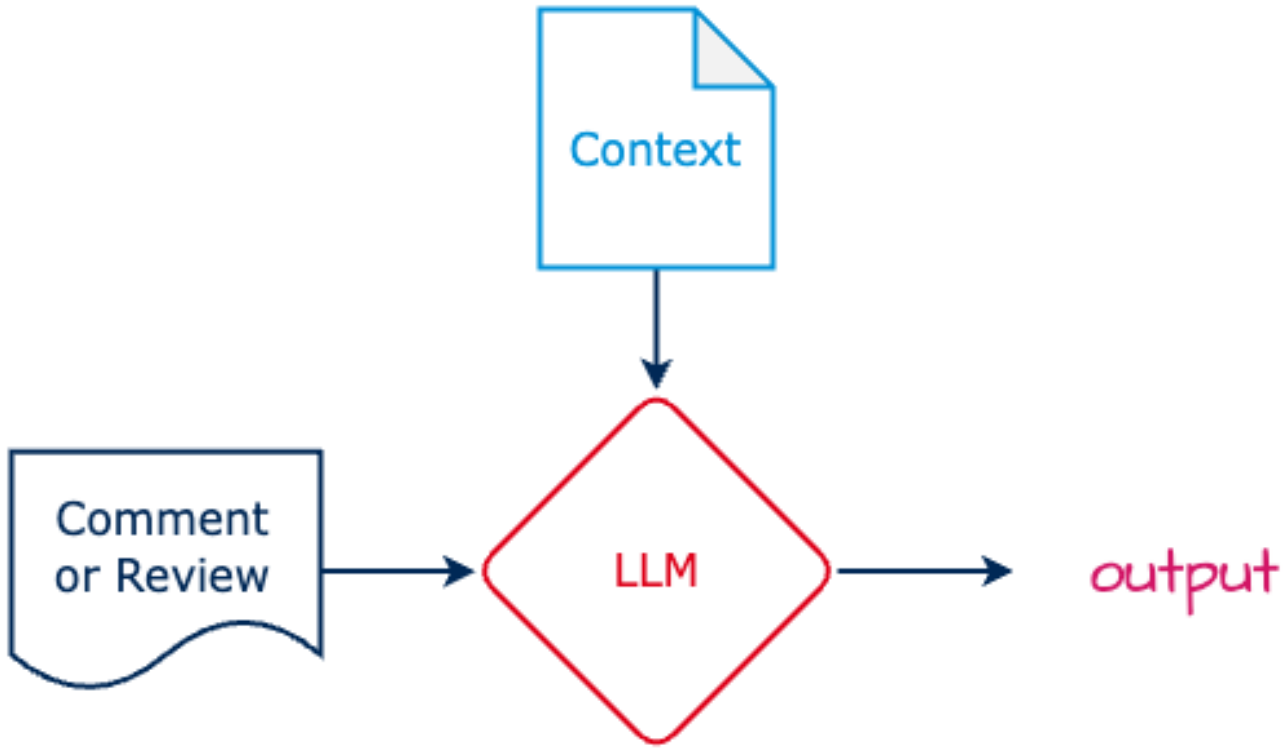


Method - KPI Extraction Pipeline



KPI Extractor: A **set** of **predefined functions** designed to **draw conclusions** or **perform analyses** on final data.

Method - Classification Block



- **Context**: A prompt containing **task instructions**, either with examples for a **few-shot** approach or without examples for a **zero-shot** approach.
- **Output**: Classification labels i.e.
  - relevance label
  - topic label and sentiment
  - overall sentiment or star rating

Method - Models

We have tested various models in order to define which is the best for our goals.

Two different **type** of models:

- **General purpose, chat like**
- **Specialized**

Method - General Purpose Models

Google DeepMind

Gemma2: large language model from Google, different versions: 2B, 9B and 27B

Meta

Llama3: large language model from Meta, 8B, 70B and 405B

MISTRAL AI_

Mistral: large language model from Mistral AI

Method - Specialized Models

DistilBert and Twitter-RoBerta:
fine-tuned versions of Bert

Experiments - Data

Politecnico di Torino

Data Sources:

Wonderflow

Digiming

YouTube

Instagram

Generative Models

Datasets:

Reviews

Comments

Synthetic Comments

LAVAZZA
TORINO, ITALIA, 1895

Experiments - Datasets

Reviews:

- Records: **2930**
- Reviews **labelled** for **Sentiment Classification**: **96,5%** of the reviews have at least one label between positive, negative and neutral topics.
- Reviews **labelled** for **Star Rating**: **100%**
- **Tiny Eco** Reviews**: 79**

Comments:

- Records: **573**
- **Labelled** Comments: **0%**
- **Tiny Eco** Reviews**: 76**

Synthetic Comments:

- Records: **236**
- **Labelled** Comments: **100%**

Experiments - Tasks

1. Relevance classification

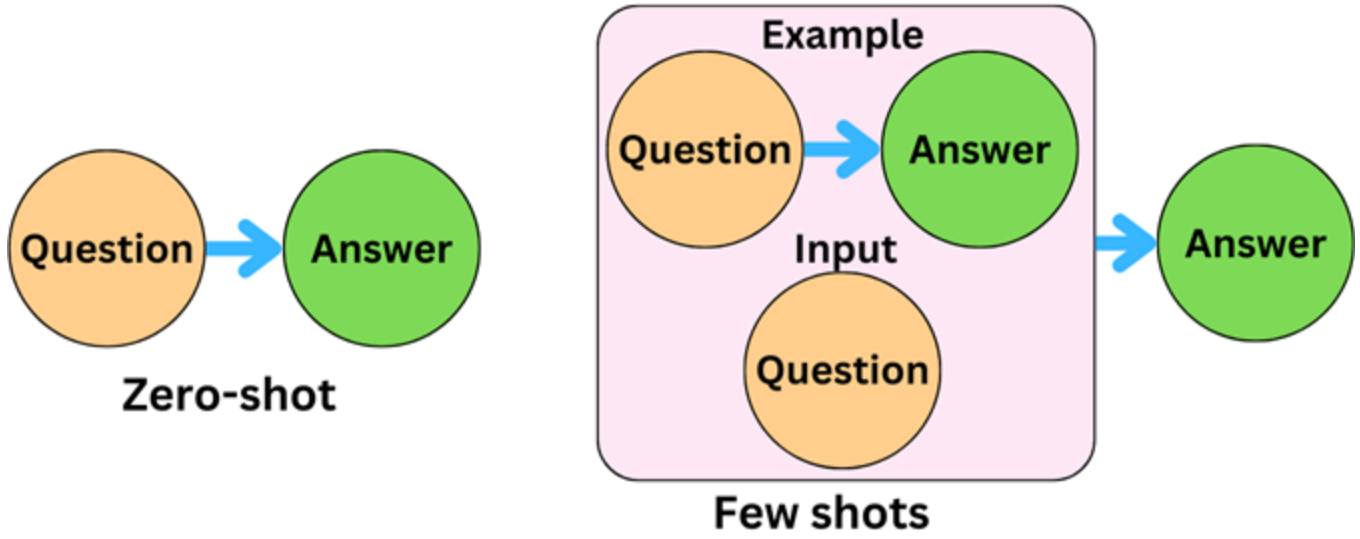
1. Multi-topic sentiment classification
   - Positive Topics
   - Negative Topics

3. Sentiment classification
   - Positive
   - Negative
   - Neutral

3. Star rating: 

Experiments - Configurations

Run Configuration: Few Shot and Zero Shot prompting



Hardware Configuration: Experiments on a Kaggle notebook, using two NVIDIA Tesla T4 GPUs

Experiments - Relevance Classification

- **Data:** 236 synthetic comments
- **Batch-size:** 15 comments

Model	Accuracy (0-Shot)	Accuracy (Few-Shot)
Gemma-2-9B-it	79.24%	80.51%
Llama-3.1-8B-it	**81.78%**	**83.9%**
Mistral-7b-it-v0.3	77.54%	79.24%

Experiments - Multi-topic Sentiment Classification

- **Data:** 500 labeled reviews + 236 synthetic comments
- **Batch-size:** 1 comment/review

Model	Accuracy (0-Shot)	Accuracy (Few-Shot)
Gemma-2-9B-it	**96.83%**	**99.07%**
Llama-3.1-8B-it	91.53%	96.19%
Mistral-7b-it-v0.3	85.94%	89.74%

Experiments - Sentiment Classification

- **Data:** 236 synthetic comments
- **Batch-size:** 15 comments

Model	Accuracy (0-Shot)	Accuracy (Few-Shot)
Gemma-2-9B-it	**89.41%**	**90.68%**
Llama-3.1-8B-it	85.17%	83.47%
Mistral-7B-it-v0.3	83.47%	82.63%
RoBerta-base-sentiment	91.59%	91.59%

Experiments - Star Rating

- **Data:** 500 labeled reviews
- **Batch-size:** 1 comment/review

Model	Accuracy	Accuracy off by one
Gemma-2-9B-it	63.85%	88.46%
Llama-3.1-8B-it	31.07%	82.52%
Mistral-7B-it-v0.3	**64.69%**	**93.08%**
Distilbert-base-uncased	69.99%	95.48%

Experiments - Models Inference Time

- The **time**, in seconds, taken by the model **to generate a response**.
- This time was **taken** during the **multi-topics sentiment classification process**.

Model	**Time (0-Shot)**	**Time (Few-Shot)**
Gemma-2-9B-it	7.06s	9.13s
Llama-3.1-8B-it	**2.93s**	**3.51s**
Mistral-7b-it-v0.3	6.09s	6.43s
Distilbert-base-uncased	0.01s	0.01s

Conclusion

We can conclude by saying that:

- **LLMs** are able to identify **relevant product comments**, with an accuracy of up to **83.9%.**
- **LLMs** are able to extract **product aspects** from reviews and comments and **assign sentiment** to them, with an accuracy of up to **99.07%.**
- **LLMs** are able to **assign sentiment** to comments and **rating stars** to reviews, with an accuracy of up to **90.68%** and **93.08% respectively.**

In particular, we have found that the model that **generally performs best** is **Gemma 2** with the **few shot prompting configuration.**

Thank you

Alessio Gioè - Catalano Vincenzo - Tommaso Mazzarini