DS372 – Project Report

Milestone 1: Data Integration
Milestone 2: Data Transformation

Prepared by:
Team Members: [Omar Al-Najjar , Laith Al-Daoud , Gaith Diabat , Retaj Melhem]

Milestone 1/Milestone 2– Data Integration, Exploration, Cleaning, and Transformation

1. Introduction / Background

Analyzing unstructured data is critical for businesses to understand user satisfaction, yet it presents significant challenges due to linguistic diversity (e.g., Arabic dialects vs. English) and inconsistent data formats. Our approach is to ingest over 900,000 records from Kaggle and Yelp datasets. We implemented a unified schema mapping strategy to standardize heterogeneous label formats (e.g., 5-star ratings and textual labels) into a single integer-based target variable, followed by extensive text normalization and tokenization to prepare the data for fine-tuning a multilingual BERT model.

2. Project Objective

The primary objective of this project is to engineer a robust end-to-end data pipeline that integrates, cleans, and analyzes customer reviews from four disparate sources. Specifically, we aim to develop a machine learning model capable of classifying sentiment into three distinct categories (Positive, Neutral, Negative) across both Arabic and English texts.

3. Datasets Overview

Dataset Name	**Source Link**	**Language**	**Size**
Arabic 100k Reviews	`https://www.kaggle.com/datasets/abedkhooli/arabic-100k-reviews`	Arabic	54.36 MB (39.5k samples)
330K Arabic Sentiment Reviews Dataset	`https://www.kaggle.com/datasets/abdallaellaithy/330k-arabic-sentiment-reviews`	Arabic	212.23 MB (330k samples)
Arabic Company Reviews ()	`https://www.kaggle.com/datasets/fahdseddik/arabic-company-reviews`	Arabic	4.41 MB (40k samples)
Yelp reviews dataset	`https://www.kaggle.com/datasets/yelp-dataset/yelp-dataset`	English	5.34 GB (7m samples)

4. ETL Pipeline

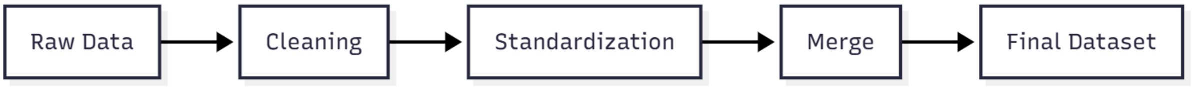


Figure 1: ETL Pipeline Diagram

5. Data Schema / ERD diagram

Column	Type	Description
review_content	String	Raw review (Arabic/English)
label	Integer	-1 Negative, 0 Neutral, 1 Positive

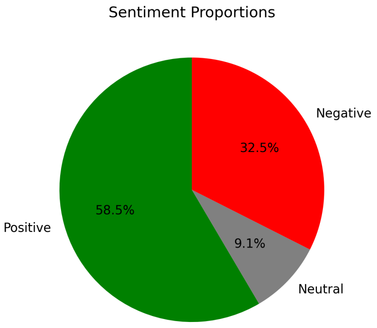
6. Cleaning Steps

List of operations:

- Removed irrelevant columns
- Mapped sentiment labels
- Standardized schema
- Removed missing/invalid entries and duplicates

7. Exploratory Data Analysis

- **Data imbalances**



(a) Sentiment distribution

- **Review Length Distribution by Sentiment**

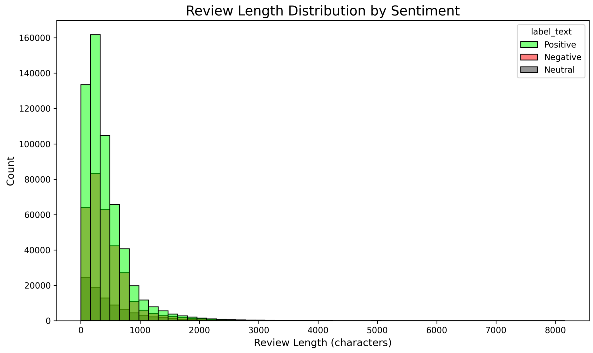


Figure 3: (b Review length distribution by sentiment)

8. Text Preprocessing / Transformation

Preprocessing Steps

- Tokenized Arabic and English text using a multilingual BERT tokenizer to prepare the data for fine-tuning.
- Converted Positive, Neutral, and Negative labels to integers for model training.
- Transformed Yelp reviews from star ratings to three categories: 1–2 stars to -1 (Negative), 3 stars to 0 (Neutral), and 4–5 stars to 1 (Positive).

9. Metadata Summary

Attribute	**Value**
Total rows after cleaning	1 million
Languages included	Arabic, English
Final columns	review_content, cleaned_text, label, tokens