Scope of Work and Preliminary EDA

Group Name: Sentivision
Members: Rabah Rouissa, Chamseddine Boukadoum, Mouna Douniazed Ayachi

Project Title

Sentiment Analysis on Yelp Reviews — Impact of Data Preprocessing on Model Performance

Project Statement

The goal of this project is to explore how data preprocessing techniques influence the performance of sentiment classification models applied to user-generated text data. Using the Yelp Reviews Dataset, the project focuses on analyzing customer feedback and predicting sentiment categories (positive, neutral, or negative).

Our main research question is:
How do different preprocessing steps — such as text cleaning, feature engineering, and normalization — impact model accuracy and overall sentiment classification performance?

We will implement, visualize, and evaluate multiple preprocessing strategies to determine their effect on data quality and predictive accuracy.

Preliminary Exploratory Data Analysis (EDA)

Before model training, a sample of ~50,000 Yelp reviews was analyzed to understand data characteristics and structure.

Key Observations:
1. Class Imbalance — Most reviews are positive (4–5 stars), with fewer negative and neutral reviews.
2. Text Characteristics — Review lengths vary widely; longer reviews tend to have higher star ratings.
3. Temporal Patterns — Average star ratings show minor year-to-year fluctuations.
4. Vocabulary Insights — Positive reviews include words like 'delicious', 'friendly', and 'amazing', while negative ones include 'slow', 'cold', 'rude'.

Planned Workflow

1. **Data Collection:** Use Yelp Reviews dataset (CSV format) — focusing on text, stars, and date columns.
2. **Data Preprocessing & Cleaning:** Remove missing, duplicate, and short reviews; normalize text (lowercase, punctuation removal).
3. **Exploration & Visualization:** Distributions (ratings, review lengths, sentiments), word clouds, and temporal plots.
4. **Model Training & Evaluation:** Train Naive Bayes (text only) and Logistic Regression (text + numeric features); compare accuracy and F1-score.

Expected Outcome

The project will demonstrate how systematic preprocessing and feature engineering significantly improve text-based model performance. We expect Logistic Regression with engineered features to outperform Naive Bayes, showing the tangible impact of well-designed preprocessing pipelines.

Tools & Libraries

Python Libraries: NumPy, Pandas, Polars, Scikit-learn, Matplotlib, Seaborn, WordCloud
Environment: Jupyter Notebook
Report Tools: LaTeX for final documentation; Google Sites for project portfolio.