Lab Requirements

Your lab report should contain two sections: the EDA section and the modelling section.

The EDA Part:

The EDA section should include:

1. Dataset Overview

- Report the number of rows and columns.
- List the column names and their data types.
- Check for missing values.

***Expected output**:* a short table or summary from pandas.info() and a comment such as “no missing values found.”

2. Target Variable (score)

- Calculate and plot the distribution of positive vs. negative classes.
- Comment on whether the dataset is balanced.

***Expected output**:* a bar chart (countplot) and one or two sentences of interpretation.

3. Review Length Analysis

- Calculate the length of reviews (in characters and words).
- Report summary statistics (mean, median, min, max).
- Plot the distribution of word counts.

Expected output:

- A histogram of word counts.
- A short note such as “Most reviews are short (10–40 words), but a few outliers are much longer.”

4. Word Frequency Analysis

- List the most common words in the dataset.
- Identify stopwords, HTML artifacts (e.g., br), mentions, or hashtags.

Expected output:

- A frequency table or bar chart of the top 10–20 words.
- A comment such as “Frequent words include many stopwords; we also see HTML artifacts (br).”

5. Positive vs. Negative Vocabulary

- Compare the most frequent words in positive vs. negative reviews.
- Highlight words that may indicate sentiment (e.g., “great”, “terrible”).

Expected output:

- A side-by-side table or two bar charts (positive vs. negative).
- A short comparison comment.

6. Noise and Preprocessing Needs

- Show examples of noisy text (mentions, hashtags, emojis, HTML tags).
- Propose preprocessing steps to handle these issues.

Expected output:

- 2–3 sample text snippets from the dataset showing noise.

- A bullet list of preprocessing actions (e.g., remove stopwords, strip HTML, lowercase text).

7. Reflection

- Summarize the main challenges of the dataset.

- Explain how your EDA results will influence your later preprocessing and model design.

***Expected output**:* 1–2 short reflective paragraphs connecting EDA findings to modeling.

Tip: Your EDA section should contain both **visuals** (plots/tables) and **interpretations** (1–3 sentences each).

Data Cleaning, Modeling, and Evaluation Part

After your EDA, you will prepare the dataset, train models, and evaluate their performance. Follow these steps carefully and document them clearly in your lab report.

1. Data Cleaning

Prepare the raw text for analysis.

Required cleaning steps:

- Lowercase all text.

- Remove HTML tags (e.g., <br>).

- Remove mentions (@username) and hashtags (#topic).

- Remove punctuation, numbers, and extra spaces.

- Remove stopwords (e.g., *the, and, to, of*).
- Optional: apply stemming (*loved → love*) or lemmatization (*better → good*).

***Expected output** in report:*

- 2–3 examples of raw vs. cleaned text.
- A short bullet list of cleaning steps you applied.

2. Feature Extraction

Convert text into numerical features.

Options:

- **Bag-of-Words (BoW)** – counts of words.
- **TF-IDF** – weighted counts that reduce the influence of very common words.

***Expected output** in report:*

- Justify your choice (why BoW, why TF-IDF, or both).
- Report the vocabulary size.
- Note any feature selection (e.g., remove very rare words).

3. Model Development

You must build and compare at least two classifiers:

1. **Naïve Bayes Classifier (NBC)**, e.g., MultinomialNB in scikit-learn.
2. **Support Vector Machine (SVM)**, e.g., LinearSVC in scikit-learn.

Training and Validation

- Use the validation set to:
  - Tune hyperparameters (e.g., SVM regularization C).
  - Compare models fairly.
- After deciding on the final settings, retrain on the **entire training data**.

Expected output in report:

- Describe how you used both the training data and validation data.
- Report any experiments with different parameters and what you observed.

4. Evaluation

After finalizing your models:

- Evaluate on the provided **validation dataset**.
- Then apply your trained models to the **evaluation dataset (5,000 Yelp reviews)**.

Metrics

You are free to choose the evaluation metrics, but you must:

- Report at least two metrics (e.g., Accuracy, F1-score, Precision/Recall, ROC-AUC).
- Justify your choice: explain why these metrics are appropriate, and why you did not choose others.
- If you only report Accuracy, explain its limitations.

***Expected output** in report:*

- A results table comparing Naïve Bayes and SVM.

- At least one visualization (e.g., a confusion matrix).
- A short discussion comparing the two models.

5. Reflection

Write a short reflection (less than 1 page):

- What did you learn about text preprocessing, model choice, and evaluation?
- Which model performed better and why?
- Did the model trained on social media data generalize well to Yelp reviews?
- What improvements could be made with more advanced techniques (e.g., word embeddings, deep learning)?

Lab Report Guidelines

Your lab report should be a clear, structured document (4–6 pages).

Structure:

1. **Title page**
    - Title of lab, your name, group number, names of group members.
2. **Introduction**
    - State the purpose of the lab (what you are trying to achieve).
3. **Exploratory Data Analysis (EDA)**

    See above.

4. **Data Cleaning & Feature Extraction**
    - Explain your cleaning steps with examples.
    - Describe the feature representation (BoW or TF-IDF).
    - Justify your choices.
5. **Model Development**
    - Describe the models used (Naïve Bayes, SVM).
    - Explain how you used a validation set.
    - Report experiments and parameter tuning.
6. **Evaluation**
    - Present chosen metrics and justify them.
    - Provide a results table comparing models.
    - Include at least one visualization (confusion matrix, etc.).
    - Discuss differences between models and performance on Yelp data.
7. **Reflection & Conclusion**
    - Summarize key findings.
    - Limitations and possible improvements.
    - What you learned from the lab.
8. **References**
    - Cite any sources (including ChatGPT if used).

Tip: Write the report as if you are explaining your work to another student who has not done the lab. Be concise, use visuals, and explain your reasoning, not just the results.