NLP Project Guidelines

M1 MIASHS

Contents

November, 2025

General Information

This project requires two deliverables:

- A complete and well-documented **Jupyter Notebook** (`.ipynb`) with clear comments explaining your methodology;
- A **10-15 minute oral presentation** with slide support (slides must be in PDF or PowerPoint format) followed by a 5-minute Q&A session.

Both the notebook and presentation may be completed in French or English. Students will work in groups of **3–4 members**. The objective is to apply NLP methods to a real-world dataset, presenting a clear analysis of your model choices and a thoughtful interpretation of the results.

1 Project Objectives

Each group should design and conduct an NLP study that demonstrates both rigorous **data understanding** and sound **modeling practices**. Your project must showcase:

(1) A thorough understanding of the dataset and the NLP problem at hand;

(2) The ability to preprocess, clean, and analyze textual data appropriately;

(3) The implementation and evaluation of one or several NLP models;

(4) Critical reflection on model performance, including limitations and potential improvements.

2 Suggested Project Structure

2.1 Dataset Presentation

Begin by describing the source and nature of your dataset, including the number of documents, language(s), average text length, and label structure (if applicable). Provide comprehensive descriptive statistics such as

$$\begin{gathered}\text{vocabulary size, average text length, class distribution,} \\ \text{standard statistical measures } n_{\text{obs}}, \text{mean}, \text{std}, \min, Q_1, \text{median}, Q_3, \max \\ \text{etc.}\end{gathered} \tag{1}$$

Visual representations such as histograms should accompany these statistics. Additionally, discuss your preprocessing strategy, addressing considerations such as text cleaning, tokenization, stopword removal, and overall dataset quality.

2.2 Task Definition

Clearly define the goal of your NLP task. Your project may focus on **text classification** (*e.g.*, sentiment analysis, spam detection), **topic modeling**, **text generation or translation**, or another NLP application with clear practical utility. The task should be well-motivated and its relevance clearly articulated.

2.3 Modeling and Algorithms

Describe and justify your choice of models. You may consider traditional embedding methods such as **Word2Vec** or **GloVe**, modern contextual embeddings like **BERT**, **CamemBERT**, or **DistilBERT**, or topic modeling approaches including **LDA** and **SIF embeddings**.

For each model employed, provide a clear explanation of its architecture, the **loss function** and its interpretation, and the training configuration (number of epochs, hyperparameters, optimizer selection). These technical choices should be justified in the context of your specific task and dataset characteristics.

2.4 Evaluation and Discussion

Evaluate your models using appropriate metrics such as accuracy, precision, recall, F1-score, or perplexity, depending on your task. Include visualizations of embeddings or clusters using dimensionality reduction techniques (PCA, t-SNE, UMAP) where relevant. For supervised learning, ***a proper train-test split must be implemented.***

Your discussion should interpret the meaning of your results in context, acknowledge the limitations of your approach, and propose concrete improvements or extensions for future work. This critical analysis is essential to demonstrate your understanding beyond mere implementation.

3 Suggested Project Ideas

Projects may focus on **regression tasks** such as:

- ☐ Predicting continuous ratings from textual reviews represented through a sentence embedding.

Or **classification problems** including:

- ☐ **Sentiment Analysis:** Classify reviews into sentiment categories.
- ☐ **Toxic Comment Detection:** Identify harassment, hate speech, or offensive language in online text.
- ☐ **Fake News Detection:** Distinguish real from fake news articles using textual embeddings or transformer encoders.
- ☐ **Product Review Analysis:** Combine sentiment classification with aspect extraction to identify satisfaction drivers.

Or **unsupervised methods** such as:

- ☐ **Topic Modeling:** Discover latent themes using LDA, NMF, or BERTopic and interpret resulting clusters.
- ☐ **Text Similarity and Clustering:** Cluster semantically similar documents using sentence embeddings and visualize via dimensionality reduction (*e.g.*, PCA, t-SNE).
- ☐ **Embedding Visualization:** Project word or sentence embeddings into 2D space to explore semantic relationships.

Students are encouraged to select tasks that align with their interests while ensuring sufficient complexity for meaningful analysis.

4 Evaluation Rubric /20

The following table presents the indicative grading scale for this project.

Component	Criteria	Points
Notebook	**Structure & Organization (2 pts):** Logical flow, well-defined sections, clear progression from data exploration to modeling and evaluation. **Documentation Quality (2 pts):** Comprehensive markdown explanations, meaningful comments, clear justification of methodological choices. **Dataset Selection (2 pts):** Appropriate difficulty level, clear relevance to the NLP task, well-justified choice with consideration of dataset characteristics and project objectives. **Code Quality & Reproducibility (1 pts):** Readable, efficient, and correct implementation; proper use of libraries and functions. **Methodological Coherence (3 pts):** Consistency between preprocessing steps, model selection, and evaluation strategy; appropriate techniques for the chosen task. **Descriptive Statistics (1 pt):** Comprehensive descriptive statistics as expected in (1).	**/11**
Presentation	**Oral Delivery (1 pts):** Clear articulation, appropriate pacing, engagement with audience, confidence in explaining technical concepts. **Results Presentation (3 pts):** High-quality figures and tables, effective visualization of model performance, clear presentation of metrics. **Critical Analysis (3 pts):** Thoughtful interpretation of results, honest discussion of limitations, well-reasoned suggestions for improvements. **Time Management (2 pts):** Adherence to 10-15 minute presentation time, balanced coverage of all sections, effective handling of Q&A session.	**/9**

Final Note

Be creative! Choose a dataset or task that motivates you, and focus on a clear, rigorous analysis rather than the size of the model. Original ideas and critical discussion are highly valued.

Recommended Datasets (Open Access)

Dataset	Description	Source
IMDb Reviews	Movie reviews labeled for sentiment analysis (positive / negative).	Stanford IMDb Dataset
60k Stack Overflow Questions Quality Rating	Dataset of 60,000 Stack Overflow posts labeled by question quality (*High Quality*, *Low Quality–Edit*, *Low Quality–Close*); useful for supervised text classification or quality assessment tasks.	From Kaggle
Consumer Sentiments and Ratings	Customer product reviews including text, sentiment, and numeric ratings — suitable for sentiment classification, regression rating prediction, and embedding analysis through PCA or t-SNE.	Consumer Ratings From Kaggle
Amazon Musical Instruments Reviews	Subset of Amazon product reviews focusing on musical instruments; includes text, star ratings, and metadata — ideal for sentiment analysis, rating prediction, or embedding visualization.	Musical Instruments Reviews From Kaggle
20 Newsgroups	Classical dataset for topic classification across 20 discussion groups.	scikit-learn: 20 Newsgroups
AG News	News topic classification dataset (World, Sports, Business, Sci/Tech).	AG News From Kaggle
Kaggle Toxic Comment	Text classification for harassment and hate speech detection.	Jigsaw Challenge (trainset is enough)

*You may choose **any dataset you prefer**, provided it enables a **meaningful NLP task** – classification, regression, or unsupervised exploration such as embedding visualization or topic modeling.*