Final Project Instructions (NLP – Text Classification)

Objective

In this project, you must choose a text classification task and work on it.

Project Steps

1. Data Selection

You can download data from any open-source platform (examples: Kaggle, Hugging Face Datasets, UCI, GitHub). Your dataset should have at least 5,000 samples.

2. Preprocessing

Clean the text (remove noise, tokenization, lowercasing, etc.). Balance the dataset if necessary.

3. Modeling

- Traditional ML Models: Use at least three models (e.g., Logistic Regression, SVM, Random Forest). For features, use statistical approaches such as TF-IDF.
- Deep Learning Models: Use at least two models (e.g., LSTM, CNN, Transformer-based). For embeddings, try different methods such as Word2Vec, GloVe, FastText, BERT embeddings.

4. Evaluation and Comparison

Evaluation metrics: Accuracy, Precision, Recall, F1-score. Provide comparison tables and graphs.

5. Report Writing

Your report should follow the structure of a scientific paper and include the following sections:

- Title & Abstract
- Introduction (problem definition and importance, with references)
- Related Work (brief review of similar works, with references)
- Methodology (dataset, preprocessing, models, features)
- Experiments & Results (numerical results, tables, charts)
- Discussion (analysis of results, strengths and weaknesses)
- References (APA or IEEE format)

6. Submission

Submit a PDF report + source code (zip or GitHub link) via email before the next class. Additionally, prepare a 15-minute presentation to present your project in class.

Important Notes

- All references must be valid (scientific articles or books).
- Results must be comparative: all models tested on the same dataset.
- Figures and graphs are mandatory in the report.