CMP4507 –Text Mining

Course Project

Spring 2025

Due Date: May 5, 2025, 23:59

Introduction

In this project, you will develop a complete text mining pipeline using a dataset you create and annotate yourself. The objective is to define and solve a real-world problem using techniques discussed in class, such as text representation, classification, and neural models.

Submission and Requirements

- Submit your **project report (PDF)**, **source code**, and **text dataset** as a .rar file on itslearning.
- **Late submissions** will not be accepted under any circumstances.
- You **must collect, clean, and annotate your own dataset** (i.e., do not use an already labeled public dataset).
- Your dataset should:
    - Contain at least **300 samples (documents, posts, sharing, comments, etc.)**.
    - Be applicable to **text classification**, **sentiment analysis**, **topic modeling**, or **similar NLP tasks**.
- Your project should explore **multiple techniques** (e.g., both traditional and neural models).
- Perform **comparative experiments**, such as:
    - Bag-of-Words vs. Word Embeddings
    - Naive Bayes vs. LSTM
    - Transformer-based vs. Classical ML models
    - Evaluation with different metrics (Accuracy, F1-score, etc.)

Evaluation Criteria

- **Project report quality**
- **Number of algorithms and techniques applied**
- **Diversity and rigor in evaluation metrics**
- **Creativity and complexity of the dataset**
- **Importance and relevance of the problem addressed**

Submission Contents (.rar archive)

- **PDF Report** (min. 6 pages, Times New Roman, 10pt) with the following structure:
    1. **Introduction**
    2. **Dataset Description** (collection method, annotation process, characteristics)
    3. **Methods** (preprocessing, vectorization, models used)
    4. **Experimental Results** (comparative results, visualizations)
    5. **Conclusions**
- **Source Code** (Python or R only, notebooks or scripts)
- **Dataset** (raw + preprocessed text, and labels)

Project Timeline

- **Project release**: Week 9
- **In-class presentations**: Weeks 13–14 (May 7 and May 14)
- **Deadline**: May 5, 2025

Group Policy

- This is a **group project**: teams of **2 students only** (individual submissions are also ok).
- **Submissions as a group of 3+** will not be accepted.

- Only one submission per group.
- File name format: {STUDENT_NUMBER1}_{STUDENT_NUMBER2}_TextMiningProject.rar

---

Presentation

Each group will present their project during class in the last two weeks.
No presentation = No credit.

Academic Integrity

All work must be original. Any plagiarism will result in **zero credit** for the project and further disciplinary actions. Note that the system uses **plagiarism detection software** that compares both web content and other student submissions.