TM&S PROJECT INSTRUCTIONS

Prof. Marco Viviani,
Joseph Muddle

marco.viviani@unimib.it
joseph.muddle@unimib.it

Introduction

- The project concerns the performance of some **tasks** related to **Text Mining**.
- The project aims to assess the understanding of what was presented during teaching from both **theoretical** and **practical** perspectives.
- The project will be carried out in **groups of 2 or maximum 3 people**, so as to stimulate collaboration as well.

TASKS

Tasks to be accomplished (1)

- **Text pre-processing (only if necessary!)**
  (text-representation-dependent, task-dependent):
  - Tokenization;
  - Normalization;
  - Stop-words removal;
  - Stemming/lemmatization;
  - For textual representation models that do not require all pre-processing operations, explain why.

- **Text representation**
  - Choose suitable representation(s) and explain the rationale behind this choice.
    - BoW (binary, TF, TF-IDF)
    - Word Embeddings (word2vec, Glove, …)
    - Contextualized Word Embeddings (BERT, ELMo, …)

Tasks to be accomplished (2)

- **“Core” tasks** (please select TWO at your choice):
  - Text classification (e.g., with respect to different topics or with respect to another aspect);
  - Text clustering;
  - Topic modeling;
  - Text summarization.

- The above-mentioned tasks must be performed on **suitable datasets**.
  - The same dataset can be used by AT MOST two groups.

Possible datasets for Text Classification

- **Different possibilities**:
  - Text Classification Dataset Repositories
  - Review Datasets
  - Online Content Evaluation Datasets
  - Sentiment Analysis Datasets

- You can have **access** to SOME of the above-mentioned datasets at the **following links**:
  - https://annotationbox.com/text-classification-datasets-for-machine-learning/
  - DO NOT use the 20 Newsgroups Dataset!

Possible datasets for Text Clustering

- Datasets employed for Text Classification can be also employed for **Text Clustering**.

- Other useful Datasets for Text Clustering:
  - https://archive.ics.uci.edu/datasets
  - https://www.kaggle.com/snap/amazon-fine-food-reviews

Possible datasets for Topic Modeling

- Datasets employed for Text Classification and Text Clustering can also be used for **Topic Modeling**.

- Other useful Datasets for Topic Modeling:
  - https://github.com/nytimes/covid-19-data
  - https://catalog.ldc.upenn.edu/LDC2008T19
  - https://www.yelp.com/dataset/

Possible datasets for Text Summarization

- **CNN/Daily Mail**
  - The dataset contains online news articles paired with multi-sentence summaries
  - https://github.com/abisee/cnn-dailymail
- **Gigaworld**
  - The dataset represents a sentence summarization/headline generation task with very short input documents and summaries
  - https://www.tensorflow.org/datasets/catalog/gigaword
- **X-Sum**
  - Data is collected by harvesting online articles from the BBC. The idea of this dataset is to create a short, one sentence news summary. More suitable for abstractive summarization.
  - https://github.com/EdinburghNLP/XSum

Other datasets at your choice

- **Dataset described in scientific papers** used or generated specifically to solve text mining tasks.

- **Any other dataset** that may be of interest to you but has particular characteristics:
  - Constituted by textual documents.
  - Characterized by an adequate number of documents.
  - Possibility of preprocessing text.
    - Datasets that already provide the representation of the text after the preprocessing phases are not adequate.
  - Adequacy with respect to the text mining task to be addressed.
    - Independently from the considered task, it is necessary to have available or be able to easily generate a "ground truth" with respect to the task addressed to provide suitable evaluations.

Tasks to be accomplished (3)

- **Evaluation**
  - Provide suitable evaluation metrics, depending on the considered task.
  - Evaluations must be COMPARATIVE. Compared to different textual representation models / compared to different algorithms for performing the task / etc.
  - The results of these evaluations must be critically discussed.

- **Important**: the proposed datasets contain textual content that refers to **different contexts**. This has to be taken into account in the development of the project.
  - Sub-sets of the data within each dataset can be considered (e.g., text referring to a specific topic), by motivating this choice.

Other instructions (1)

- **Requirements**:
  - All must be written in ENGLISH.
  - Delivery of all the material (packages, libraries, etc.) necessary to run the developed project.
  - A README.txt document of the how-to install and run the project.
  - Source code.
  - A report describing the project, the implemented solutions, the evaluations.
  - A PowerPoint presentation of the project. There will be an oral presentation and a discussion.

- The programming languages to be used for the development of the project are **R** or **Python**.

Other instructions (2)

- All the material must be shared with both Prof. Marco Viviani and Joseph Muddle at least **7 days before** the date of the written exam → **Google Drive folder**.

- The written examination and the project must be conducted in the **same examination session**.
  - If you do not pass the written examination, or if you intend to decline the grade, the mark taken in the project will be kept valid for the entire academic year.

Evaluation dimensions

- The project will be **evaluated** against:
  - Clarity in:
    - the presentation of the problem;
    - the adequate choice and treatment of the dataset(s).
  - Correctness and completeness in:
    - the pre-processing and representation of the text (use of several techniques);
    - dealing with the considered text mining task(s);
    - the carried-out evaluations.
  - Adequacy of:
    - the report;
    - all material sent.

Evaluation score

- The project will make it possible to obtain **from 0 to 4 points**. 4 points will be assigned only to particularly original projects.

- **Projects that will be better evaluated** in terms of scoring will be those that:
  - Propose non-discounted datasets and models;
  - Compare their models with any available models trained on the same dataset;
  - Will implement models described in scientific articles, but which do not have an implementation available on GitHub.

- These points will be **added** to the evaluation obtained in the written (theoretical) exam.
  - E.g., written exam: 25, project: 3 → Final score: 28/30.
  - Honors (*lode*) are acquired with a total grade equal to or greater than 32/30 → 30 e lode.

EXTRA PROJECT

Extra Project (1)

- **EVALITA** is a **periodic evaluation campaign** of Natural Language Processing (NLP) and speech tools for the Italian language.
  - The general objective of EVALITA is to promote the development of language and speech technologies for the Italian language, providing a shared framework where different systems and approaches can be evaluated in a consistent manner.

- **MultiPRIDE – EVALITA 2026 Task**
  - Multilingual Automatic Detection of Reclamation of Slurs in the LGBTQ+ Context
  - https://multipride-evalita.github.io/

- **November 21, 2025: Call for Interest deadline**

Extra Project (2)

- A **binary classification task**, in which systems must classify whether a term related to LGBTQ+ context in a sentence is used with a **reclamatory intent or not**.
  - A "reclamatory intent" in the LGBTQ+ community refers to the intentional act of taking a derogatory term, or slur, and giving it a new, positive, or neutral meaning within the community itself.

- Overall, **two different tasks**:
  - Task A - Textual Content: participants are provided only with the textual content of the tweet (Italian, Spanish, English).
  - Task B - Contextual Content: in addition to the textual content of the tweet, participants can use contextual information related to the author's profile, such as their biography (when available) (Italian, Spanish).

- **A single training set** (60% of the data) will be provided for both Tasks A and B.

- The **system to be developed** must be **run over the test data** (40% of the data).

Extra Project (3)

- Participants are invited to submit a **maximum of two runs to experiment with different models and architectures**, but discouraged from submitting slight variations of the same model.
- Systems will be evaluated using a **macro F1-score** computed over the Reappropriation binary label.
- Teams will have to **compile a report** describing the methodology used and the results obtained in detail.
  - Information about the template and format is available on the Evalita website: https://www.evalita.it/
- Further information is available at: https://multipride-evalita.github.io/pdfs/guidelines.pdf

Extra Project (4)

- The report will constitute a **genuine scientific article** that will then be submitted to EVALITA (which may be useful for those wishing to pursue a PhD).
  - The article will be reviewed and corrected together with me before being sent to the Workshop.

- **Important Dates**
  - September 29, 2025: release of training data
  - **November 21, 2025: Call for Interest deadline**
  - **November 27 – December 4, 2025: evaluation window**
  - December 15, 2025: results to participants
  - January 9, 2026: submission of participants' report
  - February 7, 2026: reviews to participants
  - February 16, 2026: camera-ready

CHOICE OF THE PROJECT/DATASET

Filling in the Google Sheet

- Groups are requested to fill in a **Google Sheet**, indicating:
  - Surnames and names of group members, separated by commas;
  - Project abstract → Short description of the project;
  - The dataset the group intends to use;
    - Please note that the same dataset can be used by a maximum of two groups.
  - For those interested in the extra project, please fill out the Google sheet and then contact me directly.

- **Link** to the Google Sheet:
  - https://docs.google.com/spreadsheets/d/1goXbWzJb2cad-Y3lNmZxue83dIF7sy2xaqBJG9tTqSc/edit?usp=sharing