NLP Project Description Spring 2025

Deadlines are listed below

1 Overview

This document contains the requirements of the NLP project and the final report for Spring 2025. Natural Language Processing is one of the most active research fields worldwide, new problems and challenges are introduced every day which motivates the need for creative ways to deal with them aiming to solve the problem and eliminate it or reduce its effect.

You can work in teams of max three while specifying the work done by each team member and their contribution, link for submitting the teams: Form The deadline for submitting the team is Saturday 1^{st} of March, 2025 at 11:59 pm. **Please refer to Section 3 for all the deadlines**

Note: You must also share your GitHub project link so we can track your progress as it is part of the evaluation. Kindly make sure your repo is visible by adding me, *mayaarosama* to the project or by making the repo public.

2 Project Requirements and Milestones

2.1 NLP Problems

The objective of this course is for you to take the lead with your choice of architecture and overall design. You are allowed to use different resources and tools to implement your project while keeping in mind the following:

1. You should have a solid understanding of your task, and implementation as you will be asked conceptual and design questions during the evaluation
2. You have to reference all the resources in your report
3. You can modify your code with access to documentation(s), without the help of conversational assistants e.g. ChaGPT, Copilot, etc..
4. You are aware of the content in your report and understand the output of your system and how to change it based on the task.

Your work and submission will be monitored through your GitHub repo commits. The deadlines are announced in advance to give you the flexibility to work at your own pace as long as you meet the given deadlines.

2.2 Milestones and Deliverables

This project is divided into three main milestones, for each you will submit a report and your code for the given task. The report should contain an explanation and reasoning for the choice of design, add your insights, analyze the output, and discuss the limitations of your framework. In milestone 1[1], the objective is to pre-process and analyze one of the following datasets. These datasets were collected as part of a research cluster over the summer:

- Spotify podcasts dataset, this dataset was collected by Gasser Ali, accessible through this link Spotify Podcasts

[1] this is your homework where you should research and argue the proper insights and analysis methods for the used dataset.

- Youtube manuscripts from famous channels, collected by Hamza Gehad, accessible through this linkYoutube Channels.

I would recommend you investigate the folders beforehand, you can also find a sample of the analysis for your reference. For the technical part in milestone 1, you will apply data analysis and provide a review of the dataset you are working with, preprocessing and cleaning the data to be usable for ML tasks; e.g. Topic modeling, Text Classification, Multi-label Tagging, etc..

In milestone 2, you will build a shallow neural network model, without using any pre-trained model, to solve the given problem. The task is to use one of the benchmark datasets for information retrieval and question-answering tasks;

- SQuAD (Stanford Question Answering Dataset)
- Natural Questions
- TriviaQA

Given a context, your model should be able to take a question and answer it from the given context. For computational purposes, you are only required to take a subset for the experiments no more than 20k rows, and no less than 5k rows. The evaluation will be based on the architecture of the network and how well you understand the task relative to your dataset, not on the performance of the model.

In milestone 3, you will build a chatbot using two pre-trained models and apply fine-tuning and/or transfer learning to improve the performance and evaluate the system on natural text unseen during the training phase. You would be evaluated on your choice of the models, the comparison between them, the choice of transfer learning or/and fine-tuning, the framework used, and your overall understanding of the system.

2.3 Weight Distribution

Milestone 1 is worth 10%, and Milestones 2 and 3 each is worth 7.5% of the course weight. Each Milestone would have a one-to-one evaluation with the team, the evaluation is conducted in a discussion manner. The exact time and date for the evaluations will be announced later during the semester based on your quizzes and other evaluations schedules.

In-class tasks and participation are worth 5%. There is a 5% bonus on the 30% for those who go beyond with their work.

3 Timeline

Kindly note that the submission should be done on your GitHub submitted repo, the last commit before the deadline is the one that your milestone will be evaluated on.

- Deadline for team submission is Saturday 1^{st} of March, 2025 at 11:59 pm.
- Teams announcement Monday 3^{rd} of March, 2025
- Reporting issues regarding team submission by Wednesday 5^{th} of March via GUC email: mayar.osama@guc.edu.eg with subject *NLP Project Team Issue*
- Milestone 1 deadline is 6^{th} of March, 2025 at 11:59pm
- Milestone 2 deadline is 17^{th} of April, 2025 at 11:59pm
- Milestone 3 deadline is 18^{th} of May, 2025 at 11:59pm

4 Useful Links

- Keras example notebooks
- LangChain Tutorials

- 10 Leading Language Models For NLP In 2022
- A Comprehensive Analysis of Various Tokenizers for Arabic Large Language Models
- A Survey on Evaluation of Large Language Models
- Understanding Deep Learning
- Huggingface
- ALUE: Arabic Language Understanding Evaluation
  PapersWithCode
- Arabic News Articles Dataset
- Arabic-named-entity-recognition
- https://huggingface.co/datasets/Fatima-Gh/GLARE
- https://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis
- https://arbml.github.io/masader/