NLP Project Description Spring 2023

Deadline is the 20^{th} of May, 2023 11:59pm

1 Overview

This document contains the requirements of the NLP project and the final presentation. 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.

In this project you are required to target one of the active tasks that appear in real applications, list of the task available in Section2 and feel free to propose your own by listing the complexity and importance of the task.

You can work in teams of two-three while specifying the work done by each team member and their contribution, link for submitting the teams: https://forms.gle/ekLvqE1Hy95FPG1P7

The deadline for submitting the team is Sunday 12^{th} of March 2023 at 11:59 pm.

Note: You are also required to share your GitHub project link so we can track your progress as it is part of the evaluation.

2 Project Requirements and Milestones

You are required to use a natural language dataset relevant to your project, and specify one main task to target in addition to two to three subtasks. Topics to discuss may include, but not limited to:

- Tweets Trending Topics Classification over Time
- Arabic Tweets Emotion Recognition
- Arabic Toxic Language Detection
- Arabic Speaking Personal Assistant
- Code Shifting Intent Detection
- Generating Subtitles for Arabic Movies
- Community Question Answering System
- Coreference resolution
- Image Generation from an Arabic text

The project is worth 15% and would be evaluated according to the size of the project, the amount of work done, and the overall understanding of the project and the main problem.

More than one team could work on the same project as long as they both have different angles, motivations, and challenges to target which we would discuss in the first presentation. The first milestone is worth 7% of the course weight and is divided into these three tasks:

- 40% collecting the data and preprocessing the natural language text
- 30% data analysis
- 30% system architecture

The second and final milestone is worth 8% of the course weight and is divided into these three tasks:

- 40% model training and fine-tuning
- 30% experiments and results
- 30% processing the output

3 Project Presentation and Report Requirements

In the first presentation, you will have 5-10 minutes to discuss an overview of your project the challenges, and the dataset that you would use, a corresponding report should also be submitted to explain the motivation and an initial plan of how to solve these challenges the report should be at most 5 pages. You may use the example project for latex to ensure the same format, a copy of the project is posted on the CMS.

In the final presentation, you would be done with the project so you need to discuss: the limitations of the dataset, the methodology and approaches that you used, the results, and a discussion on the other approaches that might be used for the same problem. The presentation should last 10-15 minutes and the report should be 8-12 pages.

The first presentation and the first report are worth 5%, and the second report and final presentation are worth 10%. You will be evaluated according to the following points:

- Delivery and clarity of the presentation
- Flow and organization of the manuscript
- Awareness of related work
- Technical quality of the paper

4 Timeline

- This document is to be posted on Monday 6^{th} of March 2023 and the deadline for the final project and presentation report is on the 20^{th} of May 2023 at 11:59 pm.
- Deadline for team submission is Sunday 12^{th} of March, 2023 at 11:59 pm.
- Reporting issues regarding team submission by Sunday 19^{th} of March via GUC email: mayar.osama@guc.edu.eg
- Final teams announcement Saturday 25^{th} of March, 2023
- First demo presentation and project proposal would be during Lab6 March 27^{th} (TBC)
- First milestone and report deadline 1^{st} of May at 11:59pm
- Second and Final Submission on the 20^{th} of May, 2023 at 11:59pm. kindly note that No edits on your GitHub project or presentation would be accepted after the deadline, any manipulation might result in a zero in the project and report grade.
- Final Presentation will be held on the 29^{th} of May, 2023. The teams' specific times and locations are to be announced.

5 Useful Links

- 10 Leading Language Models For NLP In 2022
  https://www.topbots.com/leading-nlp-language-models-2020/
- An Extensive Guide to collecting tweets from Twitter API v2 for academic research using Python3
  https://towardsdatascience.com/an-extensive-guide-to-collecting-tweets-from-twitter-api-v2-for-academic-research-using-python-3-518fcb71df2a
- How To Extract Data From The Twitter API Using Python
  https://towardsdatascience.com/how-to-extract-data-from-the-twitter-api-using-python-b6fbd7129a33
- Huggingface https://huggingface.co/models?sort=downloads
- Best 25 Datasets for NLP Projects
  https://www.kaggle.com/discussions/general/150720#845341
- ALUE: Arabic Language Understanding Evaluation https://aclanthology.org/2021.wanlp-1.18.pdf
- Reading list for Awesome Sentiment Analysis papers
  https://www.kaggle.com/getting-started/150145
- Papers with codes where you would find papers, state of the art, datasets, etc..
  https://paperswithcode.com/
- Arabic News Articles Dataset
  https://www.kaggle.com/datasets/haithemhermessi/sanad-dataset?select=Tech
- Twitter Data set for Arabic Sentiment Analysis Data Set
  https://archive.ics.uci.edu/ml/datasets/Twitter+Data+set+for+Arabic+Sentiment+Analysis
- Arabic-named-entity-recognition https://github.com/EmnamoR/Arabic-named-entity-recognition