Document Developed by: Islam Mahmoud
Email: imahmoud@conestogac.on.ca

CONESTOGA
Connect Life and Learning

PROG8245- Machine Learning Programming

Project: Applying NLP to Major Tasks

Rules-

- **The project is to be done in groups of 3 or 4 members based on the choice of the project, and all members should enroll in a group at EConestoga by the maximum deadline of Friday, March 29th, 2024.**
- **The project's submission deadline is Sunday, April 14th, 2024.**
- **The presentation of the project will take place during the last week in the scheduled session, Monday, April 15th,2024.**
- **All group members should be familiar with all parts of the code, as questions can be asked about any section during the presentation.**
- **All teams are required to be present in class during all presentations. Late arrivals will result in a 5% deduction from the project grade.**
- **Failure to present will result in a deduction of 30% from the project grade, with a maximum achievable grade of 70% if not presented.**

This project targets to make student experiment the interesting field of Natural Language Processing (NLP) by tackling one of the following topics: Language Translation, Email Filtering, Spam Detection, Simple Q/A Chatbots, and Sentiment Analysis. Through the project work, students are expected to gain hands-on experience in different stages of NLP, including data collection, preprocessing, analyzing textual data, all of which fundamental NLP techniques will be used.

Project Selection and Number of Students:

- Easy- **Sentiment Analysis:** You need to analyze sentiment of textual data and classify it as **at least 3 classes. This project can be run by a team of up to 3 students, if 4 students choose this project, it will result in a deduction of 20% of the project automatically.**
- Easy- **Spam Detection:** You need to build a system that is capable of identifying spam emails and messages, where you should be able to distinguish between spam and not spam emails. **This project can be run by a team of up to 3 students, if 4 students choose this project, it will result in a deduction of 20% of the project automatically.**
- Medium: **Email Filtering:** You need to build a system that is capable of filtering emails into three categories at least (Primary/important, Advertisement, Social Media). **This project can be run by a team of up to 4 students, no restrictions.**
- Medium: **Simple Q/A Chatbot:** You are required to build a basic question-answering chatbot which is capable of answering user queries based on closest match on trained dataset. **This project can be run by a team of up to 4 students, no restrictions.**
- Challenging: **Language Translation:** You are required to explore the challenges and techniques involved in translating text from one language to another one. Hence, students are expected to collect bilingual dataset or get one dataset and translate it using APIs to store the translation. **This project can be run by a team of up to 4 students, no restrictions.**

Note: The harder the level of the project, the more lenient the grading will be.

Document Developed by: Islam Mahmoud
Email: imahmoud@conestogac.on.ca

1. **Data Collection – (25):**
    - Gather **YOUR OWN DATASET** using webscraping or different APIs. (15) – Grade will be awarded when your code of collecting your own dataset is working fine.
        - Facebook provides an API but it is complicated. You can find it here.
        - You can make use of reddit API (PRAW) here.
        - Or you can make use of a Web Scraping techniques like Selenium, or BeautifulSoap libraries.
    - Annotate the dataset with labels based on the chosen project (sentiment analysis: positive, negative, neutral; spam detection: spam, not spam, …) based on collected data. (10)
        - You can use pretrained models from (https://huggingface.co/inference-api) to annotate your dataset.
    - **Using an available dataset with annotation will grant you only 8/25 marks.**
2. **Preprocessing (20):**
    - Perform necessary text preprocessing steps such as tokenization, stop-word removal, stemming/lemmatization, and lowercasing. (10)
    - Handle specific challenges of used text like hashtags, emojis, and slang. (10)
3. **Feature Extraction (20):**
    - Explore different feature representation methods such as bag-of-words, TF-IDF, word embeddings (e.g., Word2Vec or GloVe), or contextual embeddings (e.g., BERT or GPT). **Experiment with 3 different feature extraction techniques to capture meaningful representations of social media text where the 3 techniques should be of different word embedding categories.**
4. **Model Selection and Training (15 marks**, Part A can be skipped and a pretrained model can be used if you don't want to develop your own model **but you will lose 8 of the 10 marks of part a**.)
    a. Model Building: (10 marks)
    - Choose a suitable machine learning algorithm (e.g., Naive Bayes, SVM, or neural networks) or deep learning model for chosen task.
    - Split the dataset into training and testing sets.
    - Train the selected model using the training data, evaluate and record its performance on the training and testing data.

    b. Interpretation of results (5 marks)
    - Visualize your results explaining whether it is satisfactory or not.

Document Developed by: Islam Mahmoud
Email: imahmoud@conestogac.on.ca

5. **Deployment and Interface (10):**
    - Develop a simple user-friendly interface using any library (Tkinter is simple one for instance) that allows users to input prompts and obtain results in real-time.
    - Display a running real-time result on the interface and allow user to input prompts and classify it, **you must use your own data preprocessing and model from previous parts**.
6. **Documentation and Presentation (10):**
    - **Create a comprehensive report documenting the project's methodology, results, and findings.**
    - **Prepare a presentation to showcase the NLP project, discuss challenges faced, and highlight insights gained from the project. Including a live demonstration test cases that will be tested during the presentation which will be handled In Class.**