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 must be completed in groups of 2 or 3 members. All group members must enroll in a group on EConestoga by the deadline of Wednesday, November 27th, 2024.**
- **The submission deadline for the project is Monday, December 9th, 2024.**
- **Project presentations will take place during the last week in the scheduled session, Wednesday/Thursday, December 12th-13th, 2024 (based on your session date).**
- **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 the project presentations. Late arrivals will result in a 5% deduction from the project grade.**
- **Failure to present will result in a 30% deduction from the project grade, with a maximum achievable grade of 70% if the project is not presented.**

This project targets students to experiment with the interesting field of Natural Language Processing (NLP) by tackling the Sentiment Analysis topic (Please feel free to introduce your own topic in class if needed). 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 Description

- **Sentiment Analysis:** You need to analyze sentiment of textual data and classify it as **at least 3 classes. The project can be undertaken by a team of up to 3 students.**

1. **Data Collection – (20):**

    - Collect a dataset of product reviews. This can be obtained from APIs like Amazon Product API, Yelp API, or a public dataset such as the "**Sentiment140**" dataset or any other product review dataset available on platforms like Kaggle. (10)
        - 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 of positive, negative or neutral sentiment, based on collected data. (10)
        - You can use pretrained models from (https://huggingface.co/inference-api) to annotate your dataset.

- You can also use available libraries

2. Preprocessing (20):

- Perform necessary text preprocessing steps such as tokenization, stop-word removal, stemming/lemmatization, and lowercasing. (10)
- Remove any irrelevant columns, handle missing values, and clean text data by removing special characters, stopwords, and performing stemming/lemmatization.
- Handle specific challenges of used text like hashtags, emojis, and slang. (10)

3. Feature Extraction and Model Comparison (40):

1. 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. (10 marks)**
2. Model Building (20 marks)
    - Choose a suitable machine learning algorithm (e.g., Naive Bayes, SVM, or neural networks) or deep learning model
    - 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 (10 marks)

- Visualize your results
- Compare different feature representation
- Reach a conclusion on which is the best Embedding technique to use for your model

4. 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 <u>demonstration</u> test cases that will be tested during the presentation which will be handled In Class.**