SURV 622/SURVMETH 622 Fundamentals of Data Collection

Assignment #3: Machine Learning and Large Language Models (LLMs) with Reddit Posts

Due date: April 7, 2025, 3:30 PM.

Deliverable: Report that describes your work. Reports should not exceed 2,500 words of text (excluding any table or figures).

Introduction: In continuation of the work done in Assignment #2, your group is tasked with applying machine learning (ML) techniques and Large Language Models (LLMs) to analyze the stance of Reddit posts collected. This assignment will deepen your understanding of text analysis using R, Python, and LLMs.

Your task has two parts:

Part 1: ML coding of stance

1. **Feature Creation:** Create features that you can use in an ML model predicting stance. The first step is to carry out the basic steps of text pre-processing (e.g., removing punctuation and stop words). Then we recommend creating unigrams and removing rarely used words. The exact features are up to you but be sure to describe them in your report.
2. **Dataset Split:** Please split your corpus into 80% of the posts for the training and 20% for the test.
3. **Model Development:** Build an ML model to predict stance in the training set. Try at least two different models such as Support Vector Machine and Decision Trees. Compute performance metrics for each model (accuracy, precision, recall). Decide on the final model based on performance metrics.
4. **Model Evaluation:**
   - Using your preferred ML model, automatically code the stance in each post in the test set. Report the stance score.
   - Compute performance metrics for your model. Note: you are able to do this because the test set has manually-assigned labels.
   - Using the manually-assigned labels, compute the stance in the test set. How does it compare to the stance in the test set that you computed using the automatically-assigned labels?

Part 2: LLM coding of stance

1. **Model Selection:** Use Large Language Models (LLMs) to detect the stance of the Reddit posts, under Zero-Shot or Few-Shot setting. The LLM used in the example code is `google/gemma-3-12b-it`. You can use this model or any other LLM model. One reference for LLMs is the Hugging Face Leaderboard.
2. **Prompt Design:** Specify the prompt for the LLM using a Few-Shot or Zero-Shot scheme. The prompt should be designed to properly detect the stance of the Reddit posts. You can use the example code as a starting point.
3. **LLM Evaluation:** Run the LLM to detect the stance of the Reddit posts on the test set, and compare the performance metrics with the ML model.

You may find example R code by downloading “src/R/MLtexttutorial.R” from the “Assignment #3 materials” folder on the Canvas site. You will need to edit some code. Be sure to have the latest version of R installed.

You will also find examples of Python code by downloading “src/Python/LLMStanceTutorial.py” and “src/Python/LLMStanceTutorial.ipynb” from the “Assignment #3 materials” folder on the Canvas site. The Jupyter Notebook file will help you go over the LLM inference process step by step in a more interactive way. You can use the Jupyter Notebook file as a starting point for your work (i.e., testing the prompt and see the inference results). The Python script is pretty compact and can be used to run the LLM inference on the High-Performance Computing (HPC) cluster through the command line. The only thing you need to change in the Python script is the prompt your group came up with. The advantage of using the Python script is that it can be run on the Slurm job scheduler on the HPC cluster and automatically finish the inference process without any manual intervention (unless there are errors in your code).

Additional Notes:

- You will need to use GreatLakes HPC cluster to run the LLM. You can find the instructions on how to use the HPC cluster in the following website.
- Due to the HPC cluster being exclusively accessible to University of Michigan students, it's important to collaborate closely with the rest of your group to ensure that students from the University of Maryland also have the opportunity to gain experience with the HPC cluster.
- Feel free to reach out to Mao Li (maolee@umich.edu) with any questions. Also University of Michigan Advanced Research Computing offers consulting services for using the HPC cluster.