NLP 2023: Final Project Description

Rowan University

January 17, 2023

Instructions for submission: all code and results should be packaged and zipped. For large outputs, please save them as pickled files. Code should include enough information for me to reproduce your results. When questions ask for a specific answer or explanation, written answers should be included in a PDF formatted '{first_name}_{last_name}_Final.pdf'.

Please choose from one of the following options:

1. **Re-implementation** Propose a paper in the field of natural language processing (could be one we read in class or an alternate reading). In this exercise, you will re-create their experiments from scratch. This includes sourcing the correct datasets, pre-processing of the text data, implementation of a dataloader and model architecture and analysis of results. If there are extensive hyperparameters to train, choose those you feel are most impactful and run hyperparameter tuning, discussing the impact on the final results. Submit your code, the conda environment used to run that code and a written document detailing the process, including any challenges you faced and how you overcame them.

2. **Application** Propose a novel applied problem you feel could be solved using the techniques and models we learned in this class. Source a dataset from that problem domain, structure this dataset into train/dev/test splits and build features using a variety of models we learned over the semester. Compare and contrast these approaches on the problem you chose. Which solutions worked best, and why? Submit your code, the conda environment used to run that code and a written document detailing the process, including any challenges you faced and how you overcame them.

3. **Survey** Choose an application area outside those covered during the course (for instance, you may choose to study domain adaptation in NLP models). Build a survey of state of the art papers and methods used to solve this problem. Discuss the standard benchmarks used and the relative performance of each method over this data. What are the solved problems in this area, and what open problems still remain? Who are the relevant figures in this field of work, and how are their approaches similar/different? The deliverable should be a typeset survey (like the one here. The format should follow that of a standard conference paper and templates are available at this website.

In addition to the written paper (and code, if necessary), **you should prepare a 20 minute presentation to be given in the final in person meeting of the course**. This should include a re-statement of the problem you tried to solve, the data you used, the techniques and tools applied and an outlook on future directions.