Abbas Bakhshandeh

Riley Xin

Reetinav Das

Kevin Liu

CS 162, Winter 2024

Project Midterm Report

Team Name: Team25

Results after Milestone 2: 70% accuracy and 67% f1 score

Submission on Gradescope: Project Test Set Trial-5 (Week 7), by Abbas Bakhshandeh

Introduction:

The problem that our project is attempting to address is to use generative Natural Language Processing (NLP) techniques to both measure how well Large Language Models (LLMs) evaluate the fairness and factuality of language statements as well as to be able to improve their performance in this task by generating additional evidence. Specifically, we utilize Phi-2, a language model produced by Microsoft Research, in order to complete our examinations of this task as well as try to improve its performance as much as possible. The classification metrics that we use to measure our model's performance are accuracy and f1 score. We use the UniLC test set in order to quantify the performance of the model in distinguishing fairness and factuality of statements. Through the completion of Milestone 1 and 2, we have been able to achieve the baseline scores of 70% accuracy and 67% f1 score. In order to further improve the accuracy and f1 score of the statement predictions from the baseline, different optimization and model considerations can be taken which we plan on exploring throughout the remainder of the project.

Methods:

We were able to successfully complete Milestone 2 with Phi-2 model setup and zero shot evaluation. We implemented our prompting method based on the provided constant file as well as the training data. The current prompt setup could account for four types of prompt: phi_zero_shot_eval_prompt, phi_few_shot_eval_prompt, phi_zero_shot_evidence_prompt, and phi_zero_shot_evidence_eval_prompt. For milestone 2 particularly, the setup parses phi_zero_shot_eval_prompt type of input into *claim*, *task_type*, *language_generated*, and *domain*.

Results & Discussion:

The results on the test set was 70% accuracy and 67% f1 score, which achieved the baseline performance mentioned in the project description. The model and prompt combination we used was baseline phi-2 and phi_zero_shot_eval_prompt. The submission on Gradescope to achieve this score was made on “Project Test Set Trial-5 (Week 7)”, by Abbas Bakhshandeh, under the team name “Team25”. Regarding the future direction we plan to take, we will try to perform fine-tuning techniques to get the model to work even better.