COLUMBIA
UNIVERSITY

INTERNSHIP REPORT

Submitted as a part of COMS 6910 E Fieldwork

Submitted by: Ritvik Khandelwal

UNI: rk3213

Role: Data Science Intern

Company: Elsevier

Supervisor: Dr. Drahomira Herrmannova (NLP & Data Science Manager)

Supervisor's signature:

Due to the Non-Disclosure Agreement signed between me and Elsevier, I cannot disclose certain specifics of my work and the projects. Therefore, please reach out to me directly in case you need any additional information. I would be happy to provide the same.

CONTENTS

1. EXECUTIVE SUMMARY

Elsevier is a global information analytics business specializing in science and health. Elsevier helps researchers, educators and healthcare professionals' advance science and improve health outcomes for the benefit of society. They do so by facilitating insights and critical decision-making for customers across global research and health ecosystems. Elsevier operates all over the world with main hubs in Amsterdam, London, Philadelphia, Chennai, and Bengaluru.

I was hired as a Data Science intern to work in the health content operations team based in New York. My internship focuses on learning, working, and supporting the team with different tasks and high priority Data Science projects in the healthcare domain. In this report I go through 3 projects where I cover the description, approach, and results.

2. PROJECTS

During this internship I have had the opportunity to work on the following projects and tasks with a primary focus on supporting data development/generation for instruction tuning in LLMs guided by evaluations (Covered under point 2 listed below).

1) Sampling techniques for relation extraction
2) Prompt engineering and evaluation in LLM's for summarization and Named Entity Recognition
3) Developing and enforcing RAIDs (Responsible AI for Data Science) principles

2.1. Sampling techniques for relation extraction

Description

At the start of my internship. I was provided with a clinical dataset which had 2 concepts and their corresponding relation, and the task was to predict the relations between any 2 concepts. However, after doing EDA it found out that the data was highly imbalanced which hampered the ability to achieve great performance on this data.

Overview of my approach

To tackle the problem of this imbalanced data I used sampling techniques so to give equal representation to each of the classes. I did Oversampling, Undersampling, Synthetic minority oversampling (SMOTE) and Machine Translation. With oversampling I was able to increase the number of samples of the minority classes and with undersampling I reduced the number of samples of the majority classes. By applying SMOTE I was able to generate synthetic samples of the minority classes to make the number of samples of the minority classes equal to the majority classes. While these techniques helped, I decided to go a step further and try Machine translation to generate more samples of the minority classes. In this approach I converted a datapoint corresponding to the minority classes to French or any other language and then converted these back to English. This text was then appended as an additional datapoint to the dataset. Similarly, it was carried out for all datapoints corresponding to the minority class.

Results

Trained and evaluated several modes for relation extraction. Using the above sampling techniques significantly boosted the performance of all models. However, Among the above techniques SMOTE seemed to have limitations as it is not particularly helpful with textual data while the other 3 techniques gave impressive results.

2.2 Prompt Engineering and Evaluation in LLMs for summarization and Named Entity Recognition

Description

Instruction tuning is a crucial step in training Large Language Models effectively. In a company wide effort to develop Large Language models for over 10+ capabilities, my work involved developing data for instruction tuning for the Summarization and Named Entity Recognition capabilities followed by evaluations. To enunciate further, it was necessary to wrangle the data in the form of prompts for a language model to be trained on it.

Overview of my approach

To wrangle the data in the form of prompts, I created over 200+ prompt templates with each template being utilized to generate 10 datapoints each. This way I ensured to capture a large amount of variation in the type of prompts I created. While developing the prompts I broadly classified them into the following categories: 0 shot with indirect persona, 0 shot without persona (plain), 0 shot with direct persona, 0 shot with chain of thought, 2 shot without persona (no instruction), 2 shot without persona (with instruction), 2 shot with chain of thought, 2 shot with persona, 1 shot without persona (no instruction), 1 shot without persona (with instruction), 1 shot with train of thought, 1 shot with persona, 1 shot persona + train of thought, hallucination specific prompts. For keeping this report concise I will not be diving into what each of these categories look like and entail. However, please feel free to reach out to me for more information.

While I continued generating prompt templates under each of these categories, I developed an evaluation framework which would guide my prompt template generation. To elaborate I used an open source LLM to generate outputs to the prompts created by me. For summarization, I used BLEU, ROUGE and METEOR as the evaluation metrics and for Named entity recognition I used Precision, recall and F1 score. To give an example of how this was carried out, let's say I have 60 text files and their corresponding summaries (i.e 60 datapoints). Using one prompt template, these 60 datapoints were wrangled to generate 60 prompts and the LLM was prompted using these 60 prompts. The output was recorded, and this output was then compared with the ground truth using the evaluation metrics mentioned above.

Results

As a result of having an evaluation framework, I was able to determine how each, and every prompt performed and was able to generalize the performance of a particular category of prompts. Certain crucial patterns were observed in the evaluation. This helped make a more informed decision on the type of prompts that should be included. Perhaps, there should be more of the one's performing poorly so that the LLM built overcomes the shortcomings that these open source LLM's face. This framework developed by me will again be useful in post evaluations (i.e evaluating the final LLM built) and for every other project involving an LLM.

Another advantage of this systematic prompt development framework as opposed to randomized prompt development is that we kind of know beforehand to a certain extent about the type of prompts that could do well. This can help in guiding the end user on getting the best out of the LLM.

Since the evaluation is carried out on an open source LLM, a potential limitation could be that my prompt development is being biased by the results of just one LLM however this is an ongoing project and I plan to carry out more extensive evaluations on other LLMs as well.

2.3 Developing and enforcing RAIDs principles

Description

With the advent of LLM's it has become more important than ever to consider the real-world impact of our solutions on people, take actions to prevent the creation or reinforcements of unfair bias, be able to explain how our solutions work, create accountability through human oversight, respect privacy and champion robust data governance. Therefore, at each step of the project it is integral to look at things from the RAIDs perspective and see what things should be changed, removed, or mitigated. The goal was to do a RAIDs review on a lot of datasets and gauge whether they are suitable for use and their potential shortcomings.

Overview of my approach

Conducted initial round of reviews to in accordance with the responsible AI principles of the company followed by a deeper review which involved EDA. Through this I gauged the bias, toxicity, and representativeness of the data.

Results

- Identified datasets not suitable for commercial purposes.
- Identified shortcomings of each dataset and documented ways to mitigate them.
- Contributed towards developing and using the Algorithmic Impact Reflection tool (an internal tool developed for RAIDs).

3. Conclusion

Overall, this summer internship has been an enriching experience for me. Engaging in cutting-edge NLP projects and acquiring proficiency in new tools and technologies has been truly rewarding. Equally significant has been the privilege of collaborating with remarkably skilled and dedicated professionals. This internship stands as a pivotal stepping-stone, propelling my ambition to excel as a distinguished data professional within the healthcare sector. Above all, I extend my heartfelt gratitude to my supervisor for her unwavering support and guidance throughout this journey.