EE/CprE/SE 491 WEEKLY REPORT 1

11/12/2024 – 11/19/2024

Group number: sdmay25-proj031

Project title: Explainable AI for source code applications

Client &/Advisor: Arushi Sharma

Team Members/Role:

Manjul Balayar
Kellan Bouwman
Sam Frost
Akhilesh Nevatia
Ethan Rogers

- **<u>Weekly Summary</u>**

This week's main objective was to make first contact with our advisor, and formalize our expectations for the project. We met with Arushi for the first time on Thursday evening, and plan to do so weekly at 6:00 PM. We received more context on the project, as well as more technical details that we needed to be aware of. After the meeting, we were given a list of actions to complete to prepare for the next meeting.

- **<u>Past week accomplishments</u>**

Devin Alamsya:
- Read through the slides, repositories, and research papers that we currently have access to.

Manjul Balayar:
- Looked through all the resources trying to understand the process and outcome of the project.

Kellan Bouwman:
- Worked though notebooks and tutorials, read documents and other project related information

Sam Frost:
- Read information relating to project/skimmed repos, and getting group access to Pronto HPC

Akhilesh Nevatia:
- Skimmed through Arushi's Research Slides and attached papers, her draft paper related to our project on Overleaf, and a few of the Github Repositories presented to gather context.

Ethan Rogers:

- Prowled the shared GitHub repository and gave special interest to the provided Jupyter notebook. Also looked into various topics mentioned in our initial meeting such as tree-sitter, BIO labeling, and agglomerative and hyperbolic clustering.

Individual Time Contributions

Name	Hours This Week	Total Hours
Devin Alamsya	6	6
Manjul Balayar	6	6
Kellan Bouwman	4.5	9.0
Sam Frost	6	6
Akhilesh Nevatia	4	8
Ethan Rogers	6	6

.

○ **Plans for the upcoming week**

Manjul Balayar: Setting up Colab, getting familiar with NeuroX, pick project roles.
Kellan Bouwman:

- Task selection
- Git set up
- Gather context
- Start planning

Sam Frost:

- Continue researching project
- Communicate with necessary parties to get Pronto access for the group
- complete setup once Pronto access is granted
- begin working on portion of the project after task is confirmed

Akhilesh Nevatia:

- Gathering further context in the project, git and colab setup, starting work on a specific part of the project

Ethan Rogers:

○ **Summary of weekly advisor meeting**

Meeting Summary (9/12):

During this meeting, we focused on introductions, a project overview, and initial examples related to our work on "Explainable AI for source code applications." The project aims to develop and evaluate an auto-labeling pipeline for code datasets using Abstract Syntax Tree

(AST) tools, regular expressions, and large language model (LLM)-generated labels. We also touched upon the milestones and expectations for the first semester, including setting up the auto-labeling pipeline and preparing datasets for evaluation.

The advisor provided additional resources and documentation to help us familiarize ourselves with the tools and concepts we'll be using. Our next meeting on 9/19 will focus on discussing the materials we've reviewed and beginning the process of determining which specific sections each team member will work on.

Research related Slides

Raid Tool: Rapid creation of interpretability datasets

Recommended Research Papers to read / skim over:

Discovering Latent Concepts Learned in BERT (BERTConceptNet Dataset)

Analyzing encoded concepts in Transformer Language models

https://www.overleaf.com/read/stbrtdxfbfcd#24adf9 (WIP Draft)
https://docs.confident-ai.com/docs/metrics-llm-evals (DeepEval Metrics)
https://aclanthology.org/2024.eacl-long.48.pdf (Scaling up Discovery of Latent Concepts in Deep NLP Models)
https://neurox.qcri.org/projects/transformers-concept-net/ (Transformers Concept Net)
https://arxiv.org/html/2405.14535v1 (Exploring Alignment in Shared Cross-lingual Spaces)

Meeting Summary (9/19):

This meeting we went more in depth about specific roles and project goals. The project involves refactoring NeuroX to create NeuroX-Code, a new library focused on code analysis rather than NLP. Key tasks include making NeuroX Temp files accessible to CodeConceptNet, implementing CodeConceptNet scripts into NeuroXForCode, and following NeuroCode_text.ipnyb instructions for extracting and clustering activations. The RAID Tool will be used for lexical, syntactic, and semantic analysis, with a focus on evaluating clusters using various metrics. Priorities include Clustering Algorithm Evaluations, adding clustering evaluation metrics from sklearn, and implementing lexical metrics. The project also involves scaling by adding new models and datasets. Overall, the goal is to enhance NeuroX's capabilities for code analysis and evaluation.