ChatGPT - We See You

CS4350 Course Challenge 2023

To those of you interested, we propose the following timely challenge as an alternative to the final project. This challenge may however require more work than the normal proposal and we will also be more demanding in its evaluation since the task was given to you.

Challenge Description

ChatGPT is a large language model that can generate text from prompts among other things. This poses particular challenges to AI plagiarism checking. This challenge aims to detect ChatGPT written excerpts from human ones via graph machine learning techniques. In the following, we propose an approach for it.

You will use the word adjacency network (WAN) that transforms a text excerpt into a network [1], [(Sec.2.4); 2] and code repositories in [2,3] where you can also download the data. A WAN is an author-specific graph where each node is a function word (e.g., *of*, *the*, *and*) and an edge is established based on how the author uses pairs of these function words in the text. The node feature (or graph signal) is the frequency count of the words in an excerpt.
Data. The available data set contains text excerpts from 21 authors of the 18th century written in English. For each author are given some selected (varying between 212 to 1022 excerpts). These excerpts can serve as the human written text to you and each is affiliated with a particular poem/book of the author. *You need to generate the ChatGPT data yourself.* You can use your imagination about this. A couple of alternatives are:

- Prompt ChatGPT to generate a similar amount of excerpts for the same poems of the authors via the poem title. Divide them in AuthorGPT-specific to have author level comparisons.
- Prompt to create another similar excerpt form the given one but that is not in the poem.
- Ask ChatGPT to just rewrite completely a text of a given author.

You can also consider other text from human writing. E.g., master of science theses summaries of computer science students at TU Delft and prompt ChatGPT to generate similar summaries from the titles or differently. You can choose which human data to use for your dataset and the strategy to generate ChatGPT data. Note that this requires quite some work, so act in advance.

Remark *The WAN approach and the data generation are just some of the many strategies to approach this challenge. You do not need to strictly follow them if you have another idea (e.g., using graph of word approach as also used in [4] for constructing graph with word2vec or GLOVE based node features or other techniques). Notice however they may take substantial time, so keep it easy. For the course, it is important that the task is solved via graph machine learning techniques and that the following requirements are satisfied.*

Requirement

- Submit a project proposal detailing the way you will build the graph data and the methods. It is strongly recommended to have the data ready by the proposal.
- You need to have as baseline support vector machine see [2], another graph machine learning traditional method, and one graph neural network.
- You need to have all above methods implemented and running.
- You need to conduct an interpretability / explainability analysis
- You need to provide a creative idea (method, evaluation, experiment, analysis, challenge) of your own.
- You need to report the performance in at least three different metrics such as accuracy, recall, precision, F1, microF1 etc. You need to justify the metrics chosen.
- You need to create a Github repository that is readable and easily accessible containing all the data (both human and ChatGPT).
- You need to create a Jypiter Notebook as those in this course labs.

The team with the best work (including performance, creativity, accessability of Github and Jypiter notebook) get bonus points and their work will be used in the future editions of this course.

Reference

[1] M. Eisen and A. Ribeiro, *Authorship Attribution through Function Word Adjacency Networks*, IEEE Transactions on Signal Processing, 2015
[2] T. Sipko, *Identifying Author Fingerprints in Texts via Graph Neural Networks,* MSc Thesis TU Delft 2020 link Code repository: link
[3] Code repository from AlelabGNN Specific dataset and overall repository
[4] M. Rathee et al., *BAGEL: A Benchmark for Assessing Graph Neural Network Explanations, arXiv preprint arXiv:2206.13983*