Distinguishing human- from LLM-generated text: Stylometrics and deep learning

ALTA 2024 Shared Task

Lewis Mitchell

Project Outline

Recent advances in large language models (LLMs) have made it highly challenging to determine human-authored text from that generated by LLMs. This has implications in multiple contexts, from the education domain in which this project is situated(!) to global challenges like online social influence. This year's "Shared Task" competition for the Australasian Language Technology Association conference (ALTA 2024) is all about this challenge, and the aim of this project is to develop and submit a solution to the task. Participants are provided with a labelled training dataset containing texts where some sentences are written by humans and some by an LLM, and the goal is to develop a classification model which can predict which sentences are human-authored and which are LLM-authored on a held-out test set. There are many approaches that could be deployed here, and you will be free to experiment, however one potential “classic” approach is stylometry, which uses counts of linguistic features such as punctuation and “function words” to perform authorship attribution. This will make a good starting point for a prediction model. We can also use ensembling methods to make predictions based on multiple models.

Our aim will be to jointly make a submission to the ALTA 2024 Shared Task by the deadline of 6 October (Week 5!) and then iteratively improve each of our models over the remainder of the project.

Rough Timetable

- Week 1: Choose projects
- Weeks 2-3: Develop stylometric models for authorship attribution
- Weeks 4-5: Develop ensemble ML model, write-up and submit to Shared Task website
- Weeks 6-12: Further NLP analysis: sentiment analysis, entropy, ...
- Weeks 13-18: Further ML methods: theory, data analysis; further exploration of trends
- Weeks 19-24: Further analysis and report-writing

Further Information

The ALTA Shared Task website has a good collection of references to further “state-of-the-art” models.