Statistical Case Studies - Semester 2, Assignment 1

Background

Large Language Models (LLMs) have been increasingly prevalent since the initial release of ChatGPT in November 2022. Although LLMs can be beneficial for some tasks, there is increasing concern about the social risks they pose. For example, LLMs are thought to have increased plagiarism in schools and universities. There is also increased concern about the extent to which LLMs are used to generate content for newspapers, websites, social media, and other similar domains. As such, the task of detecting whether particular pieces of writing have been generated by LLMs is increasingly important.

In this assignment, you are asked to perform early prototyping for a company interested in creating "LLM detection" software. Your goal is to use stylometry to investigate whether text can be reliably classified as being LLM generated, versus human generated.

You are provided with a dataset which contains a variety of texts that are known to have been written by humans, and also some texts that are known to have been written by one of several LLMs (ChatGPT, Gemini, LLama). Each text has a single author i.e. it is either 100% human generated, or 100% generated by an LLM. No texts have mixed authorship.

The submission deadline is 4pm on the 20th February.

Task

Your goal is to train a supervised learning model which is capable of classifying texts as human-written vs LLM-written.

The human texts are written by a variety of different human authors, but you can treat them as all having the same author, i.e. all the human essays should be treated as being one single class. It is up to you to determine how you want to treat the LLM essays (should there be a different class for each LLM? Or should they all be grouped together into the same class?).

As a bonus question, how does your model perform if you use the "wrong" LLM to train the data? For example, suppose you train a binary classifier to classify "human-text" vs "ChatGPT-text". How does your model perform if the actual test set only contains "human-text" and "Gemini-text"?

Data

Download 'Assignment1.zip' from the course website. After unzipping it, there will be two folders: 'RawTexts' and 'FunctionWords'. The former contains the complete English language text of around 3,500 essays, grouped by author. It also contains a Python script which is used to count the function words. I am pro-

viding you with this simply to show you how function words can be extracted – you are not expected to use this folder as part of this assignment.

The other folder – 'FunctionWords' – contains the data which you should use. There are subfolders which contain the function word counts for each text (i.e. the extracted function words from the files in 'RawTexts'). Each file corresponds to one single text. 200 function words were used, so each vector has length 201. You can load this data into R using the loadCorpus function discussed in the lecture.

Your Report, and Marking Scheme

As in the assignments from last semester, you should submit a written report detailing your findings. This should be aimed at a non-technical audience, i.e. something which could be presented in a business context to intelligent people who do not have a mathematics background. You are free to structure your report however you like, but I would expect to see the following:

- A brief introduction containing an executive summary of your findings. Try to write this so that it would be understandable to someone who did not know the details of stylometry.
- A description of the data (you do not need to include more details than I have provided here) perhaps some visualisation and exploratory analysis.
- A short discussion of the methods which you will use for the analysis.
- Your results for how accurately LLM text can be identified (including the protocol you have used for evaluation – train/test set split, cross-validation, confusion matrices (or precision vs recall), etc) and a conclusion.

Your report should not exceed 5 pages.

0.1 Marking Scheme

- 80 – 100% A report that could be presented to the client or collaborator with little or no revision. Analysis is sound so that conclusions are well-supported statistically. Interpretation is reason- ably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors. The work is to a publishable standard.
- 70-79% A report that could be presented to the client or collaborator with little or no revision. Analysis is sound so that conclusions are well-supported statistically. Interpretation is reason- ably mature. The project should demonstrate a clear overview of the work, without getting lost in details, and be free of all but minor statistical errors.
- 60 – 69% A project that could be presented after a round of revision, but without having to re- do much of the actual analysis. Some flaws in the analysis or presentation (or minor flaws in both), but basically sound. A good grasp of the statistics and context, so that interpretation is reasonable.
- 50 - 59% Major re-working required before the project could be presented, but containing some sound statistics demonstrating understanding of statistical modelling and its application. Reasonable presentation and organisation.
- 40 – 49% Major flaws in analysis and presentation, but demonstrating some understanding of statistics, and a reasonable attempt to present the results.

- Fail (below 40%) Flawed analysis demonstrating little or no understanding of statistics, and/or incomprehensible or very badly organised presentation.