Statistical Case Studies - Assignment 2

Background

ChatGPT was released in November 2022 and was the first time that artificial intelligence language models had been made available to the general public. ChatGPT is an advanced AI language model created by OpenAI that uses machine learning techniques to understand and generate human-like text. It is designed to help with a wide range of tasks, from answering questions and explaining concepts, to assisting with brainstorming, and coding. Since ChatGPT was released, several other similar generative language models have also been publicly released, and are now widely used.

Although some have reported beneficial uses of ChatGPT, potential negative effects to society have also been identified. One of these is the risk that ChatGPT can be used for plagiarism in schools and universities, and also in the professional world where (e.g.) journalists may use ChatGPT to generate news articles and release these without attribution. This is possible because ChatGPT displays a remarkable ability to understand language and can generate very convincing (albeit superficial) essays and articles on essentially any subject. As such, it is important to try and produce methods for identifying whether a piece of writing has been written by ChatGPT.

In this assignment, your goal is to use the techniques from stylometry that you have learned in this course to investigate whether it is possible to detect whether a piece of writing was produced by ChatGPT. You can essentially consider this as an authorship attribution problem where one author is ChatGPT, while the second other is "human". For this purpose, you will need a large sample of texts written by both ChatGPT, and by humans.

Specifically, I will provide you with a data set which contains around 2000 human-written essays (each written by a different author), and 2000 ChatGPT-written essays. More information about how the data has been generated is provided later in this document. The essays are collected together based on their loose subject area, with there being approximately 100 different subject areas, for example "Architecture" and "Genetics"

Your goal is to investigate whether stylometry can be used to identify whether a particular piece of text (i.e. an essay) has been written by ChatGPT. In other words, given an essay for which the true author is unknown, can you identify whether the author is human? Most likely, this will involve building a suitable classifier, and then classifying each of the approximately 4000 essays in the dataset.

In particular, you should try to identify how the length of the essay affects the accuracy of your method. If an essay is 1000 words long then it should be easier to identify the author than if it were 50 words long. Your report should attempt to quantify how the accuracy of your approach varies based on the essay length.

There are also some related questions which you might want to consider:

- Does the writing style of ChatGPT depend on the topic which it is writing about? For example, if the goal is to determine the author of an essay in the "Architecture" category, should you only use the other "Architecture" essays to train your classifier? Or should all essays be used?
- Similarly, does the accuracy of your classifier suffer if you do not have access to human/ChatGPT

essays that are written on the same topic. For example, if the goal is to determine the author of an essay in the "Architecture" category, what would happen if there were no other Architecture essays available?

- Assuming you manage to obtain a high accuracy, how is your classifier actually achieving this? Are there a small number (e.g. one or two) of function words which are helping you identify ChatGPT text, or are your results being driven by small variations in all the function words? For this purpose, you might want to investigate how the accuracy changes when you only use a smaller number of function words (chosen in some appropriate manner).

Note that since each human essay has a different author, you will need to decide whether you want to concatenate all the human essays together into one large blob of 'human' writing (similar to how we previously combined all the texts written by a single author together), or whether you want to explicitly treat each essay as having a separate author. The former case would correspond to using $K = 1$ in KNN as before, whereas the latter case might correspond to a higher K (and similar for discriminant analysis, etc). A multi-dimensional scaling plot might help give insight about which approach makes more sense and, as usual, cross-validation performance should help determine which method is best.

All of the ChatGPT text was generated using the same version of ChatGPT (using the GPT4o-mini language model, for those who are curious).

Data

This section contains information about how the essay dataset was generated. The exact details will probably not be directly relevant for this assignment, but are included for completeness.

The 2000 human essays were collected from a site called Aeon.co. More specifically, it is this dataset: `https://www.kaggle.com/datasets/mannacharya/aeon-essays-dataset`. For each essay, the title and author of the essay is available, as well as the main essay text.

For each of the essays in the dataset, I have taken the title of the essay (which is usually around 10-20 words) and asked ChatGPT to write a 1000 word essay based on this title. Therefore, each of the ChatGPT essays corresponds to one particular human essay (i.e. the one with the title that was used to prompt ChatGPT). The ChatGPT essays are hence all around 1000 words long, while the length of the human essays varies. The specific prompt I used for ChatGPT was:

"Hi chatgpt I am going to give you a title for an essay, and I would like you to write a 1000 word essay on this subject. Please do not include anything in your response except for the essay. Also, your essay should not have section headings or a title. The title is:

The data is available on the course Learn page, in the "functionwords.zip" file. Once you unzip it, there will be 2 main folders, "titles" and "functionwords". The "functionwords" folder contains the function word counts for each of the human and GPTessays, while the "titles" folder contains their titles. I do not expect you to use the titles, they are simply included for completeness. It is the ''functionwords" folder which you will use.

If you are interested, you can also download the actual text of the essays, from which the function words were counted. These are available on Learn in the "rawtext.zip" file. All three folders ("functionwords", "titles" and "rawtext") contain a subfolder for each of the (approximately) 100 subject areas which the essays are grouped into. The filenames of the human essays correspond exactly to the filenames of the corresponding ChatGPT essays, For example, if you look in the 'Addiction' subfolder within the 'titles' folder, then you

will see that the file "Addiction - 1.txt" contains the following text: "The neuroscientific picture of addiction overlooks the psychological and social factors that make cravings so hard to resist". This is the title of the human essay called "Addiction - 1.txt" which is contained in the humanessays folder in rawtext.zip. The essay that ChatGPT produced when given this title is also called "Addiction - 1.txt" within the GPTessays folder. The respective function word counts for this essay are then given by the "Addiction - 1.txt" files in the "functionwords" folder. Example R code to load in this data is provided below.

Some notes:

- In general, each of the ChatGPT essays make sense, and stay true to the title that was used to generate it. However for a small number of essays this is not always the case, particularly when the titles are metaphorical and ChatGPT has misunderstood the essay content. **I do not expect you to worry about this problem**, just use the function word counts as given.
- The function word counts for each essay were counted by using a python script that I have included in the project zip file for reference. **I do not expect you to run this script**, it is provided for general interest. Again you can simply use the function word counts that I have provided.
- For the purpose of this assignment, you should assume that each essay has a different author. This is not strictly true, but it is a good enough approximation (there are actually around 1500 unique authors for the 2000 essays)

Some Sample R Code

The main goal of this assignment is to investigate how performance varies based on the length of the essay being studied. However most of the essays provided are over 1000 words long, so you will not be able to investigate shorter essays.

To fix this, I have included (on Learn) a function called reducewords() which can trim the essays down to a specified word count (e.g. 500 words). To keep things simple, this function just samples words randomly from the essay to make a new essay of the given size. You can hence use this function to explore how size impacts accuracy, for shorter essay lengths. Note that when you are investigating the effect that the number of words has on accuracy, it will most likely only be the size of the test set (i.e. the essays being classified) that needs to change, there is probably no need to reduce the size of your training set.

The following R code shows how to construct a matrix of function word counts for one particular topic (Architecture), you can use it as a starting point for your own analysis.

```
numwords <- 500 #number of words to trim the test set down into
topic <- 4 #Architecture

humanM <- loadCorpus("~/essays/functionwords/humanfunctionwords/","functionwords")
GPTM <- loadCorpus("~/essays/functionwords/GPTfunctionwords/","functionwords")

humanfeatures <- humanM$features[[topic]] #select the essays on this particular topic
GPTfeatures <- GPTM$features[[topic]]

features <- rbind(humanfeatures, GPTfeatures) #this is a matrix of both human and GPT essays

#here i use author=0 for human, and author=1 for ChatGPT
authornames <- c(rep(0,nrow(humanfeatures)), rep(1,nrow(GPTfeatures)))
```

```
#now reduce the essays down to numwords words
reducedhumanfeatures <- reducewords(humanfeatures,numwords)
reducedGPTfeatures <- reducewords(GPTfeatures,numwords)
reducedfeatures <- rbind(reducedhumanfeatures, reducedGPTfeatures)
```

You could then do leave-one-out cross validation on these essays by doing something like:

```
for (i in 1:nrow(features)) {
  train <- features[-i,]
  test <- reducedfeatures[i,,drop=FALSE] #note that only the test set size changes
  yourClassificationFunction(train,test)
}
```

Your Report, and Marking Scheme

As in the previous assignment, 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) and an exploratory analysis with appropriate visualisations that will justify your analysis.
- A short discussion of the methods which you will use for the analysis.
- Your results for how accurately ChatGPT text can be identified, and a conclusion

I am not going to be overly strict about word count, but I would expect something in the range of 10-15 pages. This is only a month long project which is quite open-ended so you will obviously not have time to analyse every single aspect of the data. It is more important to present a coherent and well-justified piece of analysis that answers a few simple questions well, than to submit a report that is full of dozens of ideas that are poorly executed.

There is no single correct analysis for this type of project, so you will not be marked on the basis of how close you get to some particular model answer. The marks are not subdivided, but will be allocated on a combination of statistical approach and justification, interpretation of results in context and presentation.

- 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.