GPT-2: Girl Detective

Analyzing AI-Generated *Nancy Drew* with Stylometry

Rebecca Lawson

IPHS 200 - Programming Humanity

Introduction

Writing styles are often viewed as unique to their writers–a compositional fingerprint of sorts. An analytical tool based upon this assumption is stylometry: the statistical analysis of the variations in the literary styles of works, often used to determine the most likely author of a particular work. Stylometric techniques abound in a multitude of fields, including history, literary studies, and even courts of law. Stylometry is often used as a form of evidence as to the identities of authors of written material pertaining to legal cases, a famous example being the conviction of the Unabomber based upon stylistic similarities between his earlier essays and his famous manuscript [1]. Thus, stylometric techniques are ascribed a lot of power. But, what if stylometry isn't as dependable as it is assumed to be? What if a writer's so-called "unique" style can be easily imitated to fool stylometric tools? In this project, we aim to analyze the ability of AI to generate text stylometrically consistent with the writer upon whom it was trained.

Methodology & Results

For this project, a GTP2 Natural Language Generator was trained on 18 of the classic *Nancy Drew* novels. The *Nancy Drew* series was famously ghostwritten, with author Carolyn Keene simply a pseudonym created by its producers, the Stratemeyer Syndicate, to create a sense of continuity within the series. Multiple writers wrote under Keene's name throughout the series' 55-year run, but the two writers with the most titles under their belt are Mildred Wirt Benson and Harriet Adams. Benson wrote 22 of the first 25 books and Adams wrote 26 of the following 28 books. Some work has been done analyzing the ghostwriting behind the series, but there is not much mystery there, as the Stratemeyer Syndicate kept records of the ghostwriters and the information can be easily found online [3].
The GPT2 is a transformer-based language model created by OpenAI with the goal of predicting the subsequent word based on the previous words in a given text. The model is fed an input of text and then generates synthetic samples of text in response to the input. It works by adapting to both the style and the content of the provided text. Here, we focus on its ability to mimic the style of the writer it was trained upon [6]. In this project, GPT2 was trained only on *Nancy Drew* novels ghostwritten by Benson. It was trained at differing periods of epochs–5,000 and 10,000–with a consistent temperature of 0.7, controlling the balance of randomness and conservativeness in generation. In the stylometric analysis, *Nancy Drew* novels ghostwritten by Adams are used as controls for style, as content is very similar, but style is presumably not. This presumption is somewhat confirmed with a stylometric analysis of works of the two ghostwriters, seen in the cluster analysis dendrogram below, which visually represents the statistical similarity of the given texts [5]:

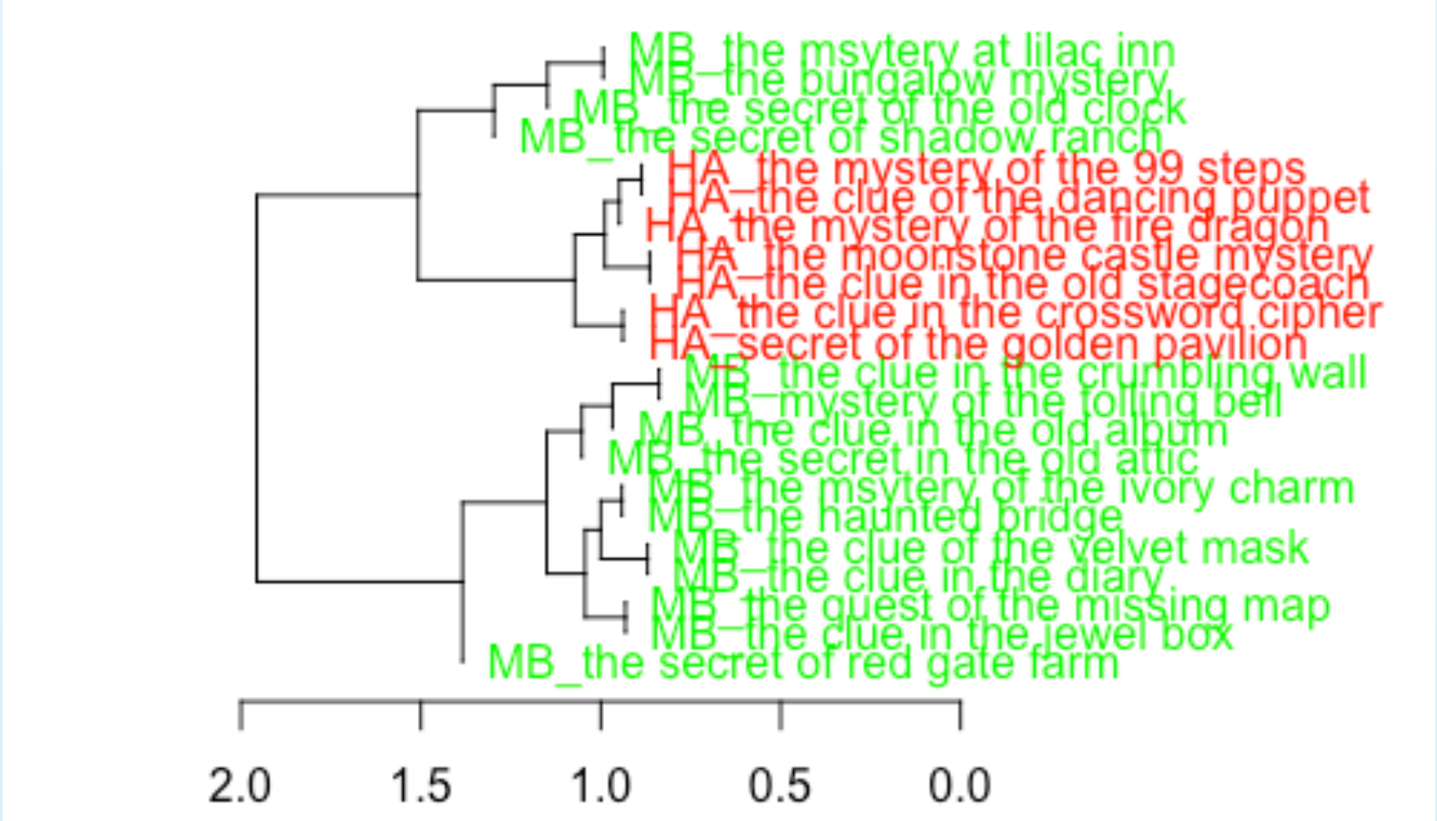

Fig. 1

In Fig.1, the green objects represent *Nancy Drew* novels written by Benson, and the red object represent *Nancy Drew* novels written by Adams. The analysis did differentiate between the 2 authors, clustering all of the Adams books together. It did seem to have some difficulty in correctly associating all Benson books together, as a clump appears to be initially clustered with the Adams books, but this may be caused by similar word frequencies, an important aspect of this cluster analysis, due to similar plot content. But, overall, when looking at the 4 main clusters, the analysis recognized the similarities between the Adams books' writing style and their variation in style from that of Benson. Benson did appear to have a little more overall variation of writing style in her books, as evidenced by their separation into 2 clumps, but this may be due to the fact that she wrote the books over a long period of time or the fact that some of the books were minorly revised by other authors and re-released in later years. When we generate a bootstrap consensus tree, the variation in style between the two writers is further confirmed.

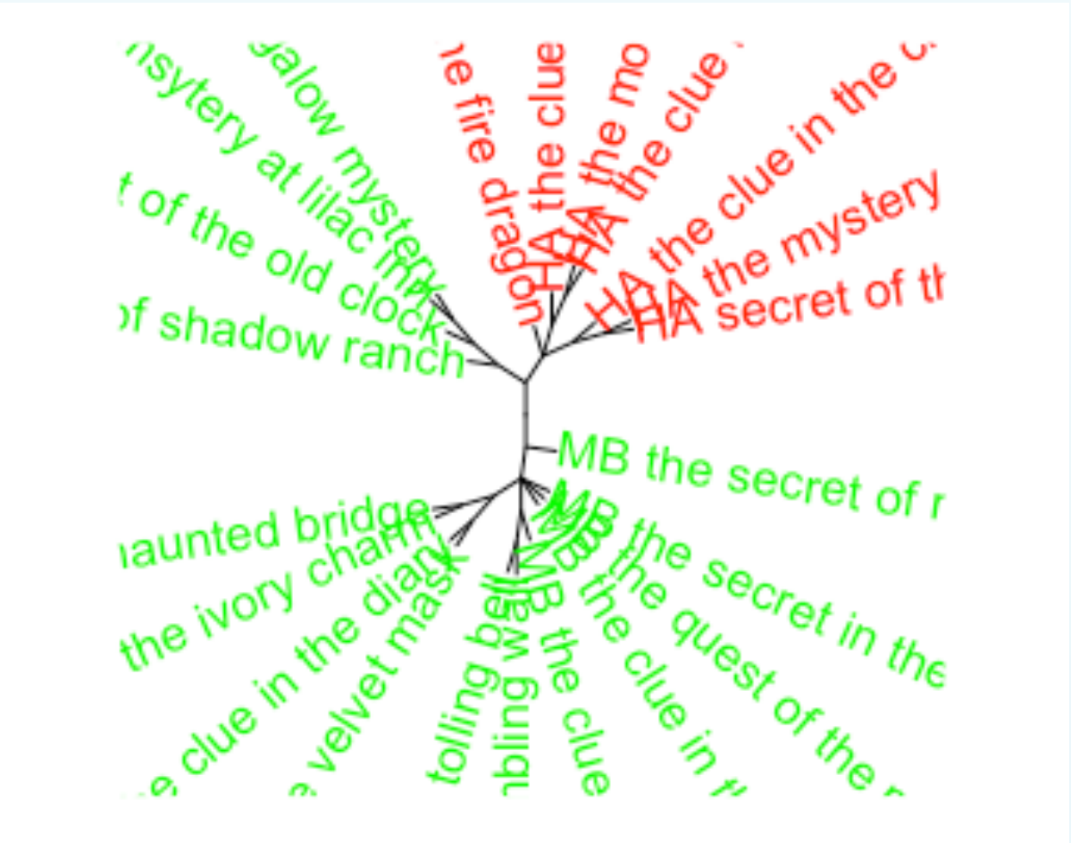

Fig. 2

This stylometric analysis uses a bootstrap sampling method to look at snapshots of each text and compare patterns across and throughout, assuming that with a large number of snapshots, true groupings will reappear [2]. And, in the case of Fig. 2, this approach effectively groups the Adams books together, separate from the Benson books. Again, as in our cluster dendrogram, we can see some variation in style within the Benson books, but ultimately they are classified as distinct from the Adams books.

Now, we look to analyze the GTP2-generated texts to determine how well GTP2 imitated Benson's writing style. As previously stated, 2 periods of epochs were used for training–5,000 and 10,000–resulting in the generation of two different texts, labelled 5K and 10K, respectively. Additionally, GPT2 was trained over 10,000 epochs and then instructed to generate its own text with no prompt or external influence. This text was labelled INF, for inference by the AI. The AI-generated texts consist of a compilation of the fragments GPT2 produced, as length of text does influence the classification methods of the stylometric analysis. The three compilations all approximately rivaled the average length of the *Nancy Drew* novels, mostly controlling for the variable of length. There was some concern about the fragmentary style of the AI-generated texts affecting the stylometric analysis, but this was not exactly the case, as observed in the results. Stylometric analysis requires a significant amount of text in order to properly recognize the complexities and idiosyncrasies of a particular writer's style, so it was determined that longer, fragmented texts would be of more analytical use than shorter, more compositionally cohesive texts. A cluster dendrogram of the Benson texts, the Adams texts, and the AI-generated texts reveals GPT2's ability to imitate Benson's style.

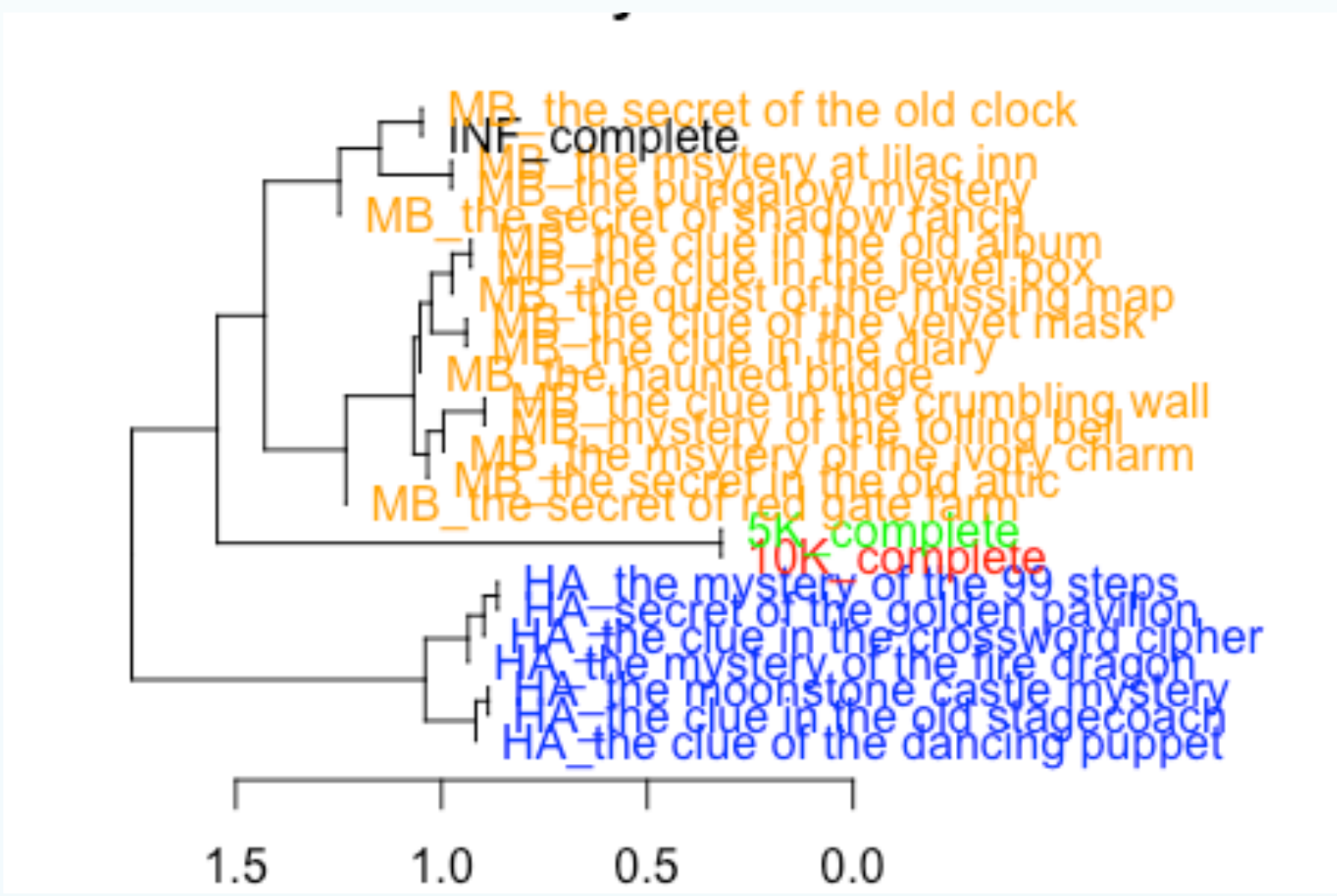

Fig. 3

Fig. 3 illustrates a lot about GTP2's strength's and weaknesses in imitating writing style. The green and red objects here represent the compilation texts generated by the AI trained at 5,000 and 10,000 epochs, respectively. The stylometric analysis clearly separates them from the Adams texts in the first cluster, grouping them instead with the Benson texts. But within this cluster of Benson books and AI texts, the analysis clearly finds a difference in style between the GTP2 and Benson, immediately separating the two AI-generated texts into their own cluster, and the Benson books into another. From this, we can conclude that while the trained GTP2 mimics Benson's style well enough to differentiate itself from another author of similar content, stylometric analysis is still able to differentiate between the human writer and the AI writer.
But, the most interesting result of this analysis concerns the compilation of text GTP2 was asked to generate on its own, free of prompt or control. This text, represented as the black object in Fig. 3, actually tricked the stylometric analysis into concluding it was most likely written by Benson. The analysis placed it in much closer proximity in style to four books actually written by Benson than to the other AI-generated texts. This result is further emphasized through the bootstrap consensus tree for this group of texts (Fig. 4).

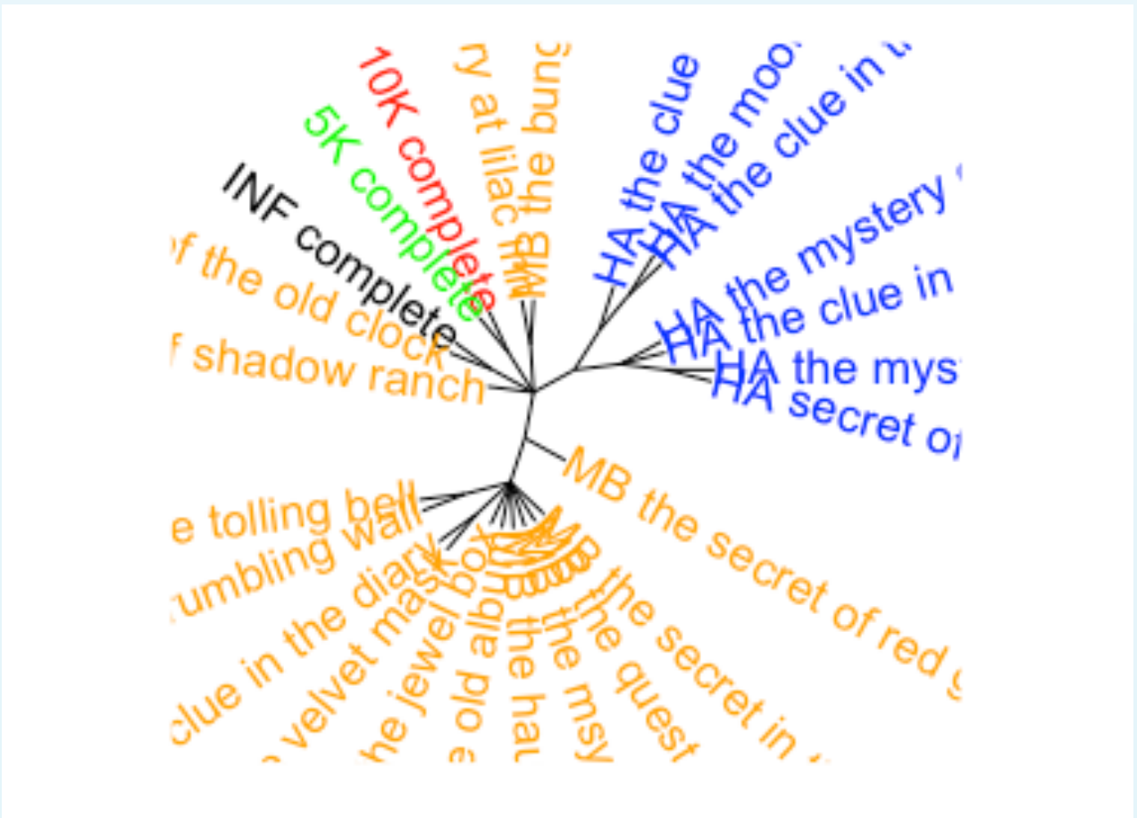

Fig. 4

This analysis does not associate the GTP2 inference text (black) with as many Benson books as the cluster analysis, but it still distinguishes between it and the two levels of training texts, placing them on separate branches. Additionally, it still classifies the inference text as more closely related in style to two Benson novels than to the other two AI-generated texts.
Even fragmented, the GTP2 inference text is statistically more similar in writing style to certain books written by Benson than even other books written by Benson are. Again, the apparent stylistic differences of different books written by Benson may be due to other factors, such as time of publication and later editing by other authors, but these analyses demonstrate GTP2's ability to match Benson's writing style closely enough to be considered more statistically consistent with books written by Benson than with the other AI-generated texts.

Conclusion

Thus, we conclude that when GTP2 is trained on one author's style over a large amount of epochs and allowed to generate text free of prompts or external control, it can come very close to effectively imitating the author's supposedly unique and identifiable writing style. It is extremely plausible that as GTP2 and other AI's evolve and develop finer learning abilities, they will be able to mimic personal writing styles so closely that stylometric analysis will not be able to distinguish between the two. This conclusion has possibly dangerous ramifications for our society. As stated earlier, stylometric analysis is based on the assumption that authors have unique complexities and idiosyncrasies in their writing style, and, so far, this has not been notably challenged. But, as the capabilities of and access to AI grow, this assumption may very well become unreasonable in the near future. The use of stylometric analysis as an evidential tool in courts of law may need to be reexamined, as writing styles become easier and easier to falsify with AI.

Works Cited

1. Barras, Colin. "Writing Style Fingerprint Tool Easily Fooled." New Scientist, 19 Aug. 2009, www.newscientist.com/article/dn17639-writing-style-fingerprint-tool-easily-fooled/.
2. "Bootstrap Consensus Networks." Computational Stylistics Group | Bootstrap Consensus Networks, computationalstylistics.github.io/projects/bootstrap-networks/.
3. "Carolyn Keene." Wikipedia, Wikimedia Foundation, 28 Sept. 2020, en.wikipedia.org/wiki/Carolyn_Keene.
4. Centre for Computing in the Humanities, King's College London. "Does Size Matter? Authorship Attribution, Small Samples, Big Problem." DH2010, 30 June 2010, dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-744.html.
5. Kirkwood, Jeffrey. "Using R for Stylometric Analysis with the Stylo Package." Vanderbilt University, Vanderbilt University, 4 Nov. 1970, www.vanderbilt.edu/digitalhumanities/using-r-for-stylometric-analysis-with-the-stylo-package/.
6. Radford, Alec. "Better Language Models and Their Implications." OpenAI, OpenAI, 3 Sept. 2020, openai.com/blog/better-language-models/.