Analyzing the Differences Between Sample Introductions

Anusha Bhat

October 13, 2023

1. Executive Summary

For this report, we have been tasked with identifying similarities and differences between student writing and writing generated by ChatGPT. We are also interested in identifying patterns employed by different categories of writers (students, academics, and ChatGPT). It is important to investigate this topic so professors can identify what lexical, grammatical, and stylistic aspects are unique to student writing to ensure that students are writing and submitting original works.

For our analysis, the Statistics and Data Science Department at Carnegie Mellon University (CMU) has provided us with a corpus of 300 introductions from sample statistic reports. In particular, 100 student intros were randomly sampled from a corpus of students papers collected from the introductory statistics course 36-200 "Reasoning with Data" prior to ChatGPT's launch. 100 ChatGPT introductions were generated using the prompts from the first data project from 36-200 as well. There are also 100 introductions from a corpus of open-access STEM journals.

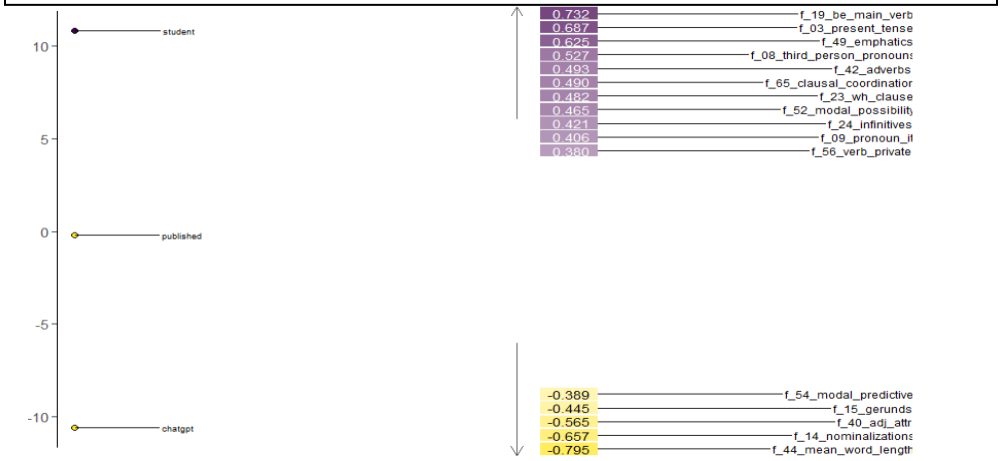
Dimensional analysis shows that journal writings are more involved and narrative, whereas ChatGPT writings are more informative, descriptive, and procedural, and student writings were a mix between these features. In addition, through principle component analysis (PCA), we find that modals, verbs, and determinants have more influence over the first and second principle component dimensions for student introductions than they do for journal and ChatGPT introductions. Professors may be preliminarily differentiating student writing from other writing by inspecting if there is a neutral usage of informative and narrative words and the frequency of modals, verbs, and determinants--proceeding with caution due to limitations in our analysis and conclusions.

We are limited in this analysis since we were not able to explore why these differences exist beyond identifying what the mere differences are. Future research can look into other sections beyond introductions to elucidate global patterns in writing between the three classes of writing.

2. Methods and Findings

2.1 Dimension Analysis

Figure 1: (+) Involved, Narrative vs. (-) Informative, Descriptive, Procedural



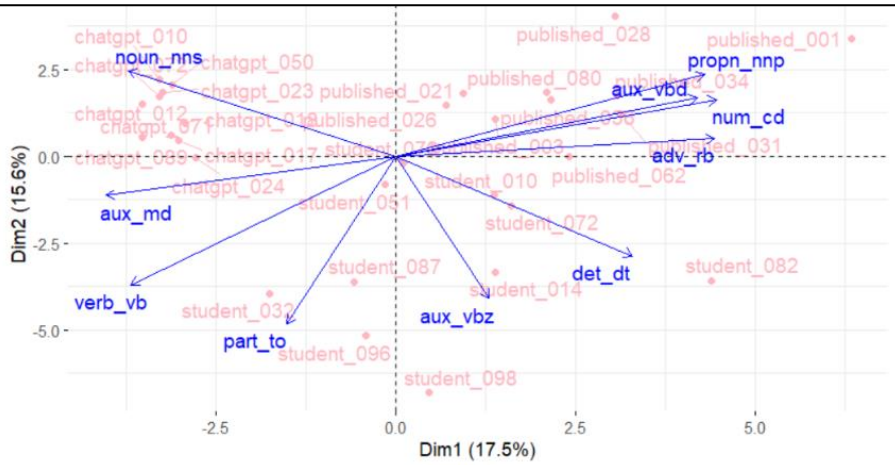
We will first look at dimension analysis and factor loadings for our corpus. We begin by annotating our corpus with parts-of-speech tagging and then perform additional tagging using the Biber loadings. For the purposes of this report, we retained the optimal amount of three factors (verified by a scree plot) and plotted a heat map of the corpus text classes along the first factor/dimension (Hardy, Romer, 2013).

From figure one, we can observe that ChatGPT is on the negative end of this dimension, indicating that these introductions were more informative and descriptive. Student papers were on the positive end of the dimension due to being more narrative (e.g., use of present tense, pronouns, and verbs), while journal introductions received a neutral score (Eberhardt, 2013).

Principal Component Analysis

Our next method will incorporate PCA. For plotting clarity, we randomly subsampled 10 student introductions, 10 ChatGPT introductions, and 10 journal introductions from the original corpus. Next, we annotated our data for parts-of-speech tags and plotted the variables with the 10 highest contributions to principal components one and two displayed on a loading plot.

Figure 2: PCA Biplot of Principle Components 1 and 2 for Sample Introductions



From figure two, we can observe that student writing is clustered towards the negative scores of dimension 2, ChatGPT writing is clustered towards the negative scores of dimension 1 and the positive scores of dimension 2, and journal writing is clustered towards the positive scores of dimension 1 and the positive scores of dimension 2. This may indicate that ChatGPT writings and journal writings have more differences between their principal components 1 and 2 than with student writing's components. Modal, verbs, and determinants have some of the greatest influence over student writing's principle components 1 and 2 (Redmond, Foucambert, 2023).

3. Future Outlooks

For further insight into the differences between student introductions and ChatGPT generated introductions, one can perform sentiment analysis, identify what parts of speech are the most commonly used in each genre, as well as examine length of different grammatical phrases. Comparing and contrasting these topics can help elucidate differences in stylistic choices between the two genres.

Finally, the researchers can include more student writing samples, ChatGPT generated samples, and STEM journal samples all corresponding to other parts of a statistical paper (e.g., methods, results, discussion) rather than only looking at introductory paragraphs/sections. This can help with comparing broader/overall trends in reports between the classes rather than looking at trends for a particular report section of interest.

References

[1] Eberhardt, S. (2013). "Identifying Multidimensional Patterns Across Register Variation". www.uni-Bamberg.de; University of Bamberg. https://www.uni-bamberg.de/fileadmin/eng-ling/fs/Chapter_21/Index.html?Multidimensionalanalysis.html

[2] Hardy, J. A., & Römer, U. (2013). Revealing disciplinary variation in student writing: a multi-dimensional analysis of the Michigan Corpus of Upper-level Student Papers (MICUSP). Corpora, 8(2), 183–207. https://doi.org/10.3366/cor.2013.0040

[3] Redmond, L., Foucambert, D., & Libersan, L. (2023). Language Corpora and Principal Components Analysis. Springer EBooks, 125, 117–132. https://doi.org/10.1007/978-3-031-29937-7_9