6-14-2024

Discerning Authorship: Distinguishing Student-Created Content from Generative AI Content Using Text Semantics

Roozmehr Safı
University of Missouri-Kansas City, safır@umkc.edu

" DISCERNING AUTHORSHIP: DISTINGUISHING STUDENT-CREATED CONTENT FROM GENERATIVE AI CONTENT USING TEXT SEMANTICS"

TREO Paper

Roozmehr Safi, University of Missouri – Kansas City, USA, safir@umkc.edu
Deepak Ayyasamy, University of Missouri – Kansas City, USA, danvy@umkc.edu

Abstract

Students increasingly rely on generative AI agents, such as ChatGPT, for their coursework. While these agents often significantly increase students' productivity, their use can suppress their creativity or lead to answers that are incorrect or biased. The ability to reliably discriminate between content written by students and that authored by AI is therefore critical. In this study, we use a typical college course assignment to investigate this topic. Our text classification method, based on a simple text mining approach, was able to perfectly discriminate between AI and student work. We then used elements from the domain of explainable AI (XAI) to identify semantic text features that contribute the most to correctly identifying text authored by AI. Our results offer a reliable method for identifying text generated by AI, along with insights into cues that can be used by human readers to identify such text.

Keywords: academic integrity, generative AI, text classification, SHAP analysis.

1 Introduction

Generative AI agents have arguably been among the most significant technological breakthroughs of recent times. By collecting and synthesizing information from a broad range of sources, these systems can answer questions from virtually all domains of human knowledge. Unsurprisingly, the use of these systems has exploded in various domains, including in education. Although there are many advantages to using generative AI in education, there are also risks, making it necessary to be able to detect content created by these agents (Westfall, 2023).

Detecting content generated by AI is important for several reasons. First, it allows readers to exercise the necessary caution when reading such content, and second, it helps to deter authors from receiving aid from generative AI when the use of these agents is not permitted (e.g., for class assignments). The goal of this study is to develop a system for that purpose. In the following sections, we discuss the details of an experiment we conducted to understand how texts created by AI can be detected using a transparent text classification technique and to identify semantic differences in content created by students versus AI that can help humans distinguish the two types of content.

2 The Experiment

We conducted an experiment involving responses to a typical college course assignment. Participants were 85 students in an introductory business analytics course at a medium-sized public university in the United States. They were instructed to study a two-page section of their textbook on the three major types of business analytics approaches and then asked to describe these approaches. Subsequently, we used ChatGPT to answer the same question, generating 85 responses to match the number of student responses. This process resulted in a total of 170 responses, evenly distributed across the two types of respondents.

3 Analysis and Results

To analyze our results, we used different elements of stylometry, the systematic study of personal styles in generating creative content (Neal et al., 2018). We first tokenized text by removing punctuation, numbers, and white spaces and converted all words to lowercase. We developed a custom list of stop words consisting of prepositions such as "in" and "for", articles such as "a" and "the", and other common words. We then removed these words from responses, as they generally contain no significant information. We retained some other commonly occurring terms such as personal pronouns, since these words can carry useful information in our context. We then represented the data in a Term Frequency (TF) matrix. In this tabular format, each row represents a response in the collection of responses (i.e., the corpus) and each column corresponds to a term present in the entire cleaned corpus.

By attaching a column containing author types to the TF matrix mentioned above, we can now use machine learning for authorship attribution. In this task, the author type serves as the binary outcome (student vs. AI) and the rest of the columns, each corresponding to a term in the corpus, serve as predictors. We then used logistic regression for classification. With 5-fold cross-validation, the accuracy achieved on our task was 100%. Accordingly, the model was capable of successfully assigning all cases to their respective classes.

To explain how the cases were classified, we used techniques from the field of explainable AI (XAI), specifically, the SHapley Additive exPlanations approach (a.k.a. SHAP Analysis; Lundberg and Lee, 2017). SHAP Analysis is a method for interpreting the role of predictors in a statistical learning model. In our study, we use it to identify the words that contribute the most to discriminating between response types. To generate SHAP values, we used an explainer object from the SHAP package in Python. The results are presented in Figure 1.

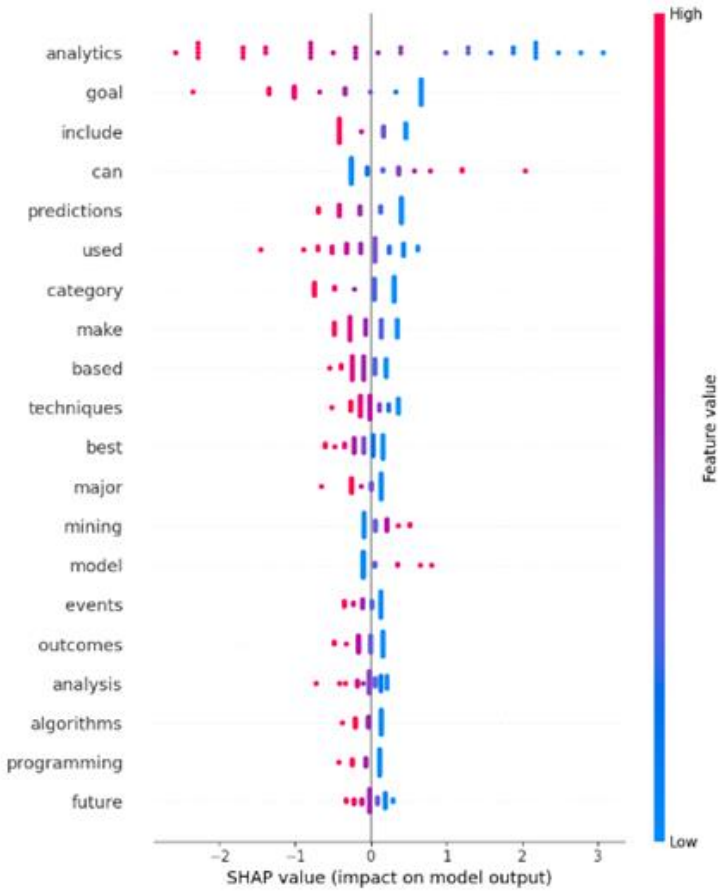

Figure 1. SHAP Summary plot for TF and TF-IDF vectorization.

The labels on the left-hand side of Figure 1 show the terms in order of their decreasing contribution to the classification task. According to the summary plots, the term “analytics”, followed by the term “goal” contribute the most to the classification task. Each dot in this visualization corresponds to a response in our dataset. The color of each dot indicates the value of the corresponding feature in the response: red indicates a high value, and blue indicates a low value. Small SHAP values positively contribute to classifying a case as written by AI whereas large SHAP values positively contribute to classifying a case as written by a student. Our model is therefore able to identify words whose usage rate can lead to predicting the author type: student or AI.

4 Discussion, Limitations, and Future Work

In this study, we devised a machine learning model for detecting text generated by generative AI. Despite its simplicity, the model was able to perfectly distinguish between student- and AI-generated text. We also used elements of explainable AI (XAI) to shed light on some semantic text features that can serve as cues for humans to detect AI-generated text.

This study has limitations that should be considered when interpreting the results. First, our dataset was limited to responses to a question from a specific domain. It would be interesting to see how the model we developed performs when it is applied to responses to a question from a different domain. Second, as discussed, different text features can be used for author attribution. To keep the scope of the study manageable, we focused our analysis on text semantic features. Future work could expand on these results by accounting for other text features, such as syntactic (e.g., the use of capitalization or punctuation), or structural features (e.g., paragraph average length or paragraph count).

References

Lundberg, S.M., Lee, S.-I. (2017). "A unified approach to interpreting model predictions," in: Guyon, I. et al. (eds.) *Advances in Neural Information Processing Systems* 30 (NIPS 2017).

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., Woodard, D. (2018). "Surveying Stylometry Techniques and Applications," *ACM Computing Surveys*, 50(6), 1–36. https://doi.org/10.1145/3132039.

Westfall, C. (2023). *Educators battle plagiarism as 89% of students admit to using OpenAI’s ChatGPT for homework.* URL: https://www.forbes.com/sites/chriswestfall/2023/01/28/educators-battle-plagiarism-as-89-of-students-admit-to-using-open-ais-chatgpt-for-homework/ (visited on August 23, 2023).