Exploration of Semantic Similarity for AI Detection

Tiffany Lee
Vanderbilt University
Tiffany.Lee@Vanderbilt.edu

ABSTRACT

With people using AI writing tools, there is a strong need to create an AI detection classifier to avoid plagiarism and continue to facilitate critical thinking skills through writing for the next generation. The purpose of this paper is to determine whether average semantic similarity between words can be used to detect whether an essay is AI-generated or human-written. The study will explore the average semantic similarity between words score in AI-generated and human-written texts to evaluate whether a particular group's text stays on-topic to a greater degree compared to other texts. We implemented a two-phase analytical approach to investigate the semantic similarities between words in AI-generated text compared to human-written essays. The initial phase involves examining the average semantic similarity between words distribution for both AI-generated and human-written essays. Following this, we conducted a post-hoc analysis to assess the influence of different writing prompts on average semantic similarity between words, determining whether specific prompts lead to greater topical adherence in texts. Our study then integrates both analytical findings through a multivariate regression analysis, which considers both the text generation method and the specific prompts used. The results from the study indicate that there is a linguistic characteristic difference between AI-generated and human-written texts in terms of whether a particular text stays on-topic to a greater degree compared to other texts. This implies that we can consider using average semantic similarity between words as a feature in a future AI detection classifier.

Keywords

NLP, AI detection classifier, semantic similarity, student essays

1. INTRODUCTION

With the rise and accessibility to OpenAI's ChatGPT and other large language models (LLMs), people can use generative AI for a variety of tasks, such as writing messages or essays. This helps save a user's time and creative mental effort in these tasks, but it has also led to the rise of potential plagiarism and become an obstacle in having inexperienced writers improve their writing skills due to their increased dependency on AI tools [1]. The rise in generative AI for writing poses a future challenge regarding teaching critical thinking skills to future generations as previous research suggests that writing increases a student's critical thinking performance for a variety of topics [2, 3]. Thus, there is a strong need to create an AI detection model to ensure that essays are truly sourced from the original writer and to continue instilling the critical writing skills needed in future generations.

Despite their sophistication, generative AI models still exhibit subtle differences in their use of language compared to humans as previous research has indicated that "writing style of the AI models exhibits linguistic characteristics that are different from those of the human-written essays" [4], such as sentence complexity and cohesion. For example, a joint study from Wuhan University and Worcester Polytechnic Institute used syntax, semantics, and pragmatic linguistic differences observed between AI-generated and human-written texts to create an AI detection model [5]. Subsequently, the linguistic characteristic differences between human-written versus generative AI texts are potential useful features when creating a classifier to conduct AI detection, which can combat potential plagiarism, help education systems continue to teach students how to improve their writing skills by reducing their dependency on AI tools, and check authenticity in future written works.

2. PURPOSE STATEMENT AND RESEARCH QUESTION

This analysis focuses on semantic similarity between words to indicate whether a particular text stays on-topic to a greater degree compared to other texts. Based on prior research suggesting that lexical diversity of humans is higher than that of ChatGPT-3 [4], which means that humans tend to use a wider range of vocabulary compared to ChatGPT-3, the study investigates if AI-generated texts stay on-topic to a greater degree than human-written essays due to humans using a wider range of vocabulary that results in lower average semantic similarity scores between words compared to generative AI.

The study used student-written and OpenAI's GPT generated essays, where each essay had a prompt about car-free cities or the electoral college. We assessed each essay's topicality using spaCy to calculate the average semantic similarity between words to determine how effectively an essay stayed on topic throughout the text. We compared the average semantic similarity scores between words distribution for both AI and student-written essays to identify if there are any distribution differences between the two groups. In addition, a post-hoc analysis was produced where we examined the average semantic similarity between words distribution for both AI and student-written essays when accounting for specific prompts as different sub-groups. To analyze statistical significance, we used both t-test and a regression analysis to assess whether the linguistic characteristic could be useful in an AI detection classifier. The research question that guides this study is the following:

- Can average semantic similarity between words be used to detect whether an essay is AI or not?

3. LITERATURE REVIEW

Recently, there has been a growing amount of research examining the impact of AI writing tools on human writing capabilities. This interest stems from concerns about people's overreliance on AI tools that impacts their writing skills and the increasing amount of plagiarism due to using generative AI. Many studies focus on identifying the linguistic characteristic differences between human-written and generative AI writing to either have a better understanding on how generative AI is affecting people's cognitive skills or to use as features in an AI detection classifier to alleviate the growing concerns about writing integrity and authenticity. Specifically, previous research has investigated how AI influences various aspects

of writing, including content organization, syntax, semantics, and pragmatics [1, 5, 6, 7].

Nevertheless, the differences from generative AI influencing a person's writing style have also led to beneficial outcomes within the realm of writing. For example, tools like ChatGPT have helped people "to brainstorm writing ideas, get individual grammar & phrasing feedback, and write more polished articles" [8]. The increased accessibility to AI writing tools has overall provided individuals more opportunities to get personalized and immediate writing support, which further cultivates confidence in a person's writing skills.

Unfortunately, generative AI usage has also led to several drawbacks where the increased AI dependency has caused people to "lose the capacity or willingness to think by themselves" [8], resulting in diminished creativity and reduced critical thinking skills. This becomes an overarching problem as generative AI, which is trained on existing human-generated content, inherently recycles past ideas. Consequently, this reliance perpetuates a cycle of lack of originality, limiting the potential for innovative human thought from writing [9]. Another drawback is the inherent biases present in the generative AI's initial training algorithms. These biases lead to AI tools excelling in certain linguistic characteristics while underperforming in others. As a result, individuals may adapt their writing styles to align with the AI's strengths, further narrowing an individual's capability to express diversely in their own unique manner. The adaptation not only limits personal writing development, but it reinforces the homogeneity in writing encouraged by generative AI if left unchecked for people learning to write [6, 8].

The use of generative AI in writing has also extended to the realm of research, presenting unique challenges. The research community is facing increasing pressure to address situations where generative AI is producing high-quality fraudulent papers that not only cite incorrect reference materials but also mimic the style and format of credible research. This situation makes it difficult for reviewers and readers to identify falsified data or fabricated studies, which risks the integrity of scientific publication. Additionally, the ease with which generative AI can generate plausible research findings spreads misinformation within academic fields. This creates a significant problem as it becomes easier for individuals to create and disseminate fraudulent scientific content that can easily go undetected [9]. To combat these issues, journals are recommending alternative methods and stricter publication protocols to further enhance fraudulent research identification [10].

Although alternative methods such as stricter rubrics and review processes and increased ethical penalties can help discourage overreliance on AI writing tools, the significant disadvantages associated with the use of generative AI in writing highlight the need for robust AI detection classifiers. An AI detection classifier is needed to differentiate between human-written and AI-generated content to ensure there is regulation involved in contexts where authenticity, critical thinking, and originality are crucial. Numerous research studies have explored various natural language processing techniques when developing an AI detection classifier by identifying linguistic characteristic differences in AI-generated text. The linguistic characteristic differences explored in other research studies include parts of speech (POS), sentence length, stop word ratio, semantics, coherence, and pragmatics [5].

3.1 Expanding the Scope via Current Study

In this study, we will also examine whether there is a linguistic characteristic difference regarding average semantic similarity between words for human-written and AI-generated text, which can lead to using average semantic similarity between words as a feature for an AI detection classifier. The difference that sets this study apart from other AI detection classifier studies is that previous literature focused on linguistic characteristic differences in the realm of scientific papers, such as biology, computer science, and medicine [5, 7]. This results in creating an AI detection classifier that works in a limited context, but it is unknown whether the identified linguistic characteristics can work effectively in other contexts.

Given the accessibility of generative AI to individuals across all age groups, not just those from scientific or research-focused backgrounds, it is essential to extend these analyses to additional domains, such as student essays. This is particularly important given the established drawbacks of using generative AI in writing, which can impact critical thinking and creativity. By examining whether the difference in average semantic similarity between words is consistent across various types of writing, this study aims to refine AI detection classifiers for broader applications. This research differentiates itself by specifically focusing on middle school and high school student essays to determine if average semantic similarity between words can serve as a crucial feature in an AI detection classifier. This focus aims not only to address issues of over-reliance on AI writing tools and plagiarism, concerns well-documented by previous research, but also to extend these concepts to safeguard the development of critical thinking skills among younger students who are learning how to write.

4. METHODS

4.1 Data

Two datasets were used in this study. The first dataset is derived from the LLM - Detect AI Generated Text Kaggle Competition, which includes 1375 human-written essays by middle and high school students. These essays were part of a controlled setup where students were instructed to respond to given essay prompts after reading various source texts. The competition and the dataset are the product of collaborative efforts between Vanderbilt University's Peabody College and the Learning Agency Lab, with funding support from the Bill & Melinda Gates Foundation, Schmidt Futures, and Chan Zuckerberg Initiative. Access to the Kaggle dataset used in this study can be found at: https://www.kaggle.com/competitions/llm-detect-ai-generated-text/overview

The second dataset consists of 700 AI-generated essays, created by Radek Osmulski for the same LLM - Detect AI Generated Text Kaggle competition. This dataset mimics the same student-written response conditions where generative AI is given the same source texts and prompts to ensure comparability between human-written and AI-generated essays. The generative AI essays come from two versions of OpenAI's GPT models: 500 essays from GPT 3.5 Turbo and 200 essays from GPT 4. The essays were generated from using OpenAI's API to give the generative models the respective source texts and prompts. Access to the Kaggle dataset used in this study can be found at: https://www.kaggle.com/datasets/radek1/llm-generated-essays/data

This analysis particularly focused on whether average semantic similarity between words can be utilized to discern AI-generated essays from human-written ones. By concentrating on semantic similarity between words, the study aimed to determine the degree to which a text maintains topical relevance compared to other essays. The analysis considered essays responding to prompts on car-free cities or the electoral college, ensuring a consistent basis for comparison between human and AI-generated texts. The

methodology enables a controlled comparison, which is essential for isolating the effects of authorship on the linguistic characteristic differences that may exist between different essay prompts, where the study can assess differences in semantic similarity resulting from differences in authorship. Associated code and additional details about the data used in this study can be found at: https://github.com/Math1019/nlp_ai_detection

For this study, both datasets have been merged into a single comprehensive dataset categorized by various features. These features include id for essay identification, prompt id indicating whether the essay addresses the car-free cities or the electoral college prompt, essay text, and a categorical variable specifying whether the essay was human-written or AI-generated, and further details on the specific generative AI model used, if applicable. As a result, the study's dataset has prompts generally balanced based on each generated source where no single source, human or AI, dominates the essay count for each prompt category as seen in figure 1.

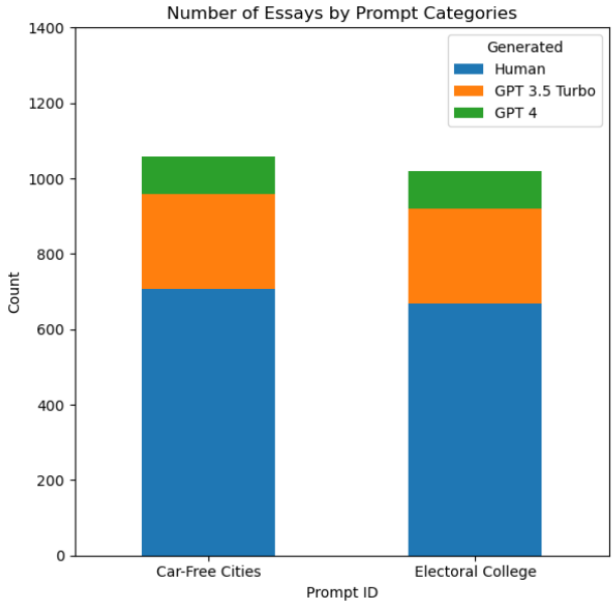

Figure 1. Bar chart for essay distribution by prompt in dataset

4.2 Semantic Similarity Analysis

In this study, the analysis of semantic similarity between words was designed to evaluate each essay's topicality by calculating the similarity between words within each document. This approach involved two primary steps: preprocessing the text to focus on content words and calculating average semantic similarity between all the content words in the document. The text preprocessing was carried out using the spaCy natural language processing library, specifically leveraging the large English model. This large English model was chosen for its suitability involving semantic understanding due to its comprehensive vocabulary and sophisticated word embeddings [12].

The initial step involved using spaCy to tokenize the text, followed by the removal of stop words, which includes words like "a," "the," and "is", and punctuation. Stop words were filtered out because they carry minimal semantic load due to their high frequency in any general text, which would skew our analysis focusing on topicality with content words. Once we isolated the content words, we utilized them to calculate the semantic similarity between words. The spaCy model employs Word2Vec embeddings, which represent words as dense vectors within a continuous vector space, to capture the semantic relationships based on contextual usage across the document when calculating the semantic similarity.

Semantic similarity between words was quantified by iterating through pairs of content words in a document. This computation utilizes spaCy's capability to measure similarity, which is based on the cosine distance between word vectors. The resulting similarity scores from each pair of content words, where the scores have floating point numbers ranging from 0 to 1 with 1 indicating perfect semantic alignment, were then averaged to produce an overall average semantic similarity score for each document.

4.3 Statistical Analysis

Initial evaluations involved assessing the distributions of average semantic similarity between words for both human-written and AI-generated essay groups in the main analysis, which focused on the general comparison between human-written and AI-generated texts. Additionally, these evaluations were applied in the post-hoc analysis, where the data was further segmented by prompts to explore subgroup differences. To ensure comparability and address potential scale disparities, skewness and kurtosis tests were applied to assess the symmetry and tail distribution of the data. Bartlett's test was then conducted to evaluate variance differences between the groups, leading to the selection of the appropriate statistical test, either the two-sample t-test or Welch's t-test, based on the variances observed. These analyses were further complemented by calculating Cohen's D to measure the effect size.

However, to mitigate the risk of Type I error from multiple t-tests and to manage the family-wise error rate, a comprehensive multiple linear regression analysis was performed as a final step in the statistical analysis. This analysis was facilitated by creating dummy variables for categorical predictors, such as prompt and whether an essay was human-written or AI-generated. Ordinary Least Squares (OLS) regression was used to get a nuanced understanding of the impact of AI generation and prompt type on semantic similarity between words across the corpus. The multiple linear regression analysis further enhanced the robustness of the findings found through the t-tests by controlling for multiple comparisons in a single model framework.

5. RESULTS

5.1 Independent T-Tests

We initially conducted t-tests on the general group, focusing solely on the comparison between human-written and AI-generated essays without considering specific prompts as sub-groups. Subsequently, in our post-hoc analysis, we extended this approach to include different prompts, allowing for a more detailed examination of subgroup variations. We first observed that the average semantic similarity scores between words scores for both the general and prompt-specific sub-groups are normally distributed. This was confirmed by histogram visualizations, with an example of the general distribution presented in Figure 2. Additional visualizations are available on the study's GitHub page. The normally distributed findings were further supported by the skewness and kurtosis values listed in Table 1.

Table 1. Skewness and kurtosis for all groups

Group	Skewness	Kurtosis
General - Human	-0.209	3.547
General - AI	-0.331	3.236
Car-Free - Human	-0.225	3.854

Group	Skewness	Kurtosis
Car-Free - AI	-0.150	2.666
Electoral - Human	-0.829	4.817
Electoral - AI	-0.661	3.455

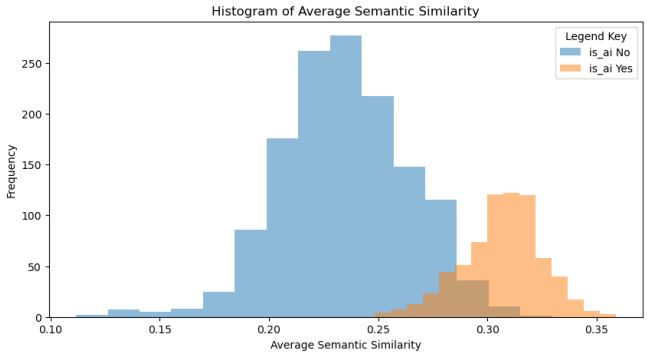

Figure 2. Histogram Distribution by General Authorship

In our analysis of the general group, which consists solely of human-written and AI-generated essays without accounting for specific prompts, we established our null hypothesis to be that there is no significant difference in the average semantic similarity scores between words in AI-generated texts and human-written essays, which would indicate that authorship does not impact how well a text stays on-topic. Based on significant variances identified by Bartlett's test, Welch's t-test was deemed appropriate and subsequently indicated that there is a statistically significant difference with the average semantic similarity between words scores for the two groups, so we rejected our null hypothesis. Furthermore, the large positive effect size indicated by Cohen's D suggests that AI-generated essays maintain a higher degree of topical relevance compared to human-written essays, as evidenced by their higher average semantic similarity scores between words.

Further exploration through post-hoc analysis involved examining the average semantic similarity between words scores within specific prompts to identify distribution differences between AI-generated and human-written texts. Consistent with our general analysis, where we are using the same null hypothesis from the general analysis, but we are now applying it to specific prompt contexts, we observed a recurring pattern of rejecting the null hypothesis across the two prompt categories, each supported by large effect sizes. This indicates a robust trend where AI-generated essays generally exhibit greater topic consistency, which is the average semantic similarity between words score, than their human-written counterparts regardless of the two prompts. The results from the independent t-tests when determining whether the results are statistically significant are in Table 2 and 3.

Table 2. Bartlett test results for different analysis groups

Group	Bartlett's Test Statistic	Bartlett's p-value	T-Test Type
General	191.66	$p < 0.001$	Welch's
Post-hoc: Car-Free	2.46	0.117	Independent Two-Sample
Post-hoc: Electoral	214.96	$p < 0.001$	Welch's

Table 3. T-test and Cohen's D results by analysis group

Group	T-Statistic	Test p-value	Cohen's D	Effect Size Type
General	-68.71	$p < 0.001$	2.751	Large
Post-hoc: Car-Free	-56.50	$p < 0.001$	3.689	Large
Post-hoc: Electoral	-43.48	$p < 0.001$	2.340	Large

5.2 Multiple Linear Regression

To mitigate the potential type I error due to family-wise error rate, we also performed a multiple linear regression analysis in the post-hoc analysis where we used the essay's authorship, which is whether an essay was human-written or AI-generated, and the specific prompt, consisting of 'Car-Free Cities' and 'Electoral College' as categorical predictors to predict the average semantic similarity between words score.

The multiple linear regression analysis revealed similar results that we saw in when conducting using t-tests in our post-hoc analysis on prompt sub-groups. We saw that the model accounts for 66.9% of the variability in average semantic similarity between words scores. This indicates that these two factors are significant contributors to the differences in semantic similarity, as evidenced by the substantial Adjusted R-squared value of 66.8%. This value suggests that the model's explanatory power is robust, unaffected by the number of predictors. The model's overall significance is further indicated by an F-statistic of 2093.0, with a p-value approaching zero, providing strong evidence to reject the null hypothesis in favor of the alternative hypothesis that there is a significant difference in the average semantic similarity scores between words in AI-generated and human-written essays for a given specific prompt.

We see that the authorship coefficient is 0.0716, implying that AI-generated texts exhibit a higher average semantic similarity than human-written texts, holding the prompt constant. This finding, supported by a p-value of close to 0, indicates a statistically significant effect of AI authorship on increasing average semantic similarity between words scores. Similarly, the prompt is 0.0170 suggests a slight increase in average semantic similarity for essays responding to the electoral college prompt compared to those on car-free cities, when controlling for authorship. The p-value of close to 0 also confirms that the specific prompt also has a significant, albeit small, effect on average semantic similarity between words scores. Thus, the multiple linear regression analysis indicates that both AI authorship and prompt type play a significant role in determining a text's topicality where we see that AI authorship plays a much higher effect compared to the prompt type. The results from the multiple linear regression model are seen in Figure 3.

OLS Regression Results

Dep. Variable:	avg_semantic_similarity	R-squared:	0.669
Model:	OLS	Adj. R-squared:	0.668
Method:	Least Squares	F-statistic:	2093.
Date:	Tue, 23 Apr 2024	Prob (F-statistic):	0.00
Time:	11:32:38	Log-Likelihood:	4743.3
No. Observations:	2078	AIC:	-9481.
Df Residuals:	2075	BIC:	-9464.
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	0.2264	0.001	266.593	0.000	0.225	0.228
is_ai	0.0716	0.001	62.525	0.000	0.069	0.074
prompt_id	0.0170	0.001	15.723	0.000	0.015	0.019

Figure 3. Multiple linear regression model results

The model's scatter plot found in Figure 4 visualizes the relationship between the observed and model-predicted average semantic

similarity between words scores, providing insight into the model's predictive accuracy. The red linear line illustrates perfect prediction, indicating points where the model's predictions exactly match the observed values. A significant number of points are closely clustered around this line, affirming the model's effectiveness in forecasting average semantic similarity based on the authorship and prompt predictor variables. The figure also shows four distinct clusters where each cluster corresponds to an authorship and prompt combination. We see that combinations where an essay is AI-generated will have a higher average semantic similarity between words scores compared to its human-written counterpart for any given prompt. Also, we see the relative influence of the prompt on average semantic similarity between words. Essays pertaining to the 'Electoral College' are positioned higher than essays from the 'Car-Free Cities' prompt suggesting that the 'Electoral College' prompt will have a slightly higher average semantic similarity between words when the authorship factor is held constant. Thus, the findings from the scatter plot visually reinforce the multiple linear regression model results.

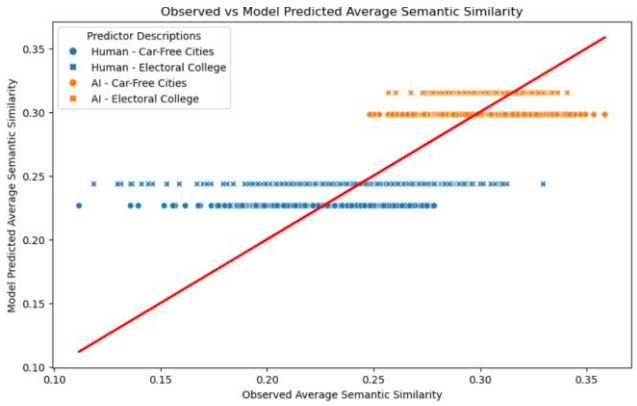

Figure 4. Multiple linear regression scatterplot by predictor

We also created an interaction plot, which is Figure 5, to examine the potential interaction effect between authorship and prompt on the average semantic similarity between words score as this will help us determine if generative AI's impact on average semantic similarity between words is consistent across both prompts.

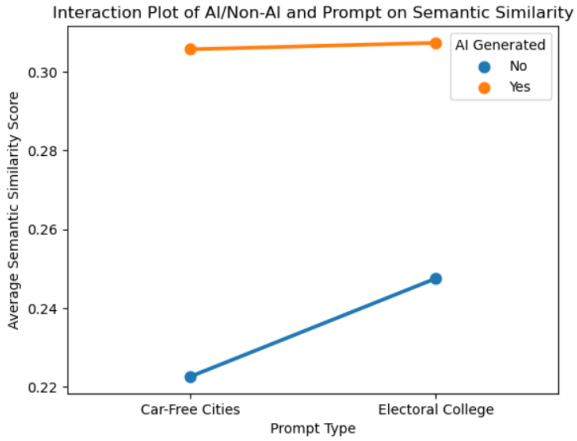

Figure 5. Interaction plot by authorship and prompt

The parallel lines across the 'Car-Free Cities' and 'Electoral College' prompts suggest a uniform influence of AI on average semantic similarity between words scores. The lack of intersection between these lines indicates that the increased average semantic similarity between words scores associated with AI-generated texts are steady across both prompts. This implies that AI's ability to maintain higher topical adherence is a stable characteristic that is not influenced by the prompt subject matter. Since we see that the contribution of AI authorship to average semantic similarity between words is not contingent on the essay prompt, this implies that the effects of AI authorship and prompt type are essentially additive as each predictor factor independently contributes to the overall average semantic similarity between words score without interacting significantly with the other. Thus, the figure further underscores our understanding that AI-generated essays consistently yield a higher average semantic similarity between words by maintaining topic relevance across the two prompts.

6. CONCLUSION AND FUTURE WORK

In this study, we explored average semantic similarity between words in AI-generated and human-written texts in a general analysis and in a post-hoc analysis to evaluate whether a particular group's text stays on-topic to a greater degree compared to other texts. The study was conducted through t-tests and a multiple linear regression model. Our findings across all the tests suggest that AI-generated essays consistently exhibited higher average semantic similarity between words scores compared to human-written essays, which indicates that AI-generated essays have greater topicality relevance than those written by humans across our study's two prompts.

The general comparison between AI-generated and human-written essays showed that there was a statistically significant difference in the average semantic similarity between words, underscoring AI's capacity to produce more topically related essays. This pattern persisted across the prompt-specific examinations within the post-hoc analysis for both t-tests and multiple linear regression model, implying that the observed discrepancy is not prompt-dependent but rather an intrinsic characteristic of AI-generated texts.

The results showing that there is a linguistic characteristic difference between AI-generated texts and human-written essays regarding average semantic similarity between words is an aspect that other AI detection studies have previously noted when creating an AI detection classifier [5]. The difference with this study compared to other studies is that we specifically explored whether the average semantic similarity between words linguistic characteristic difference holds when the prompts are specifically focused on material aimed at middle school and high school students, whereas previous research studies focus on highly educated scientific backgrounds [5, 7]. We deemed it crucial to explore this aspect due to the growing need for an AI detection classifier capable of regulating students' access to generative AI. Such a tool is essential to curb plagiarism and to prevent an overreliance on generative AI that could potentially diminish students' critical thinking abilities [8, 9]. Thus, the results from this study indicate that we can consider using average semantic similarity between words score as a feature for AI detection to hopefully counter plagiarism and continue supporting written text from human thought.

Through the findings in this study, we recognize there are a number of limitations and additional potential to explore in identifying whether average semantic similarity between words is a linguistic characteristic difference that can be used in an AI detection classifier. The first recommendation is to conduct a cross-analysis using different groupings, such as by generative AI model. Future research could benefit from disaggregating the data by generative AI model to examine the nuances of AI-authored text compared to human-written essays. Such analysis could reveal additional linguistic

distinctions associated with different AI models, which remained unexplored in the current study. Another recommendation is to assess the semantic similarity at the sentence level rather than between individual words. According to other research studies, there is an interest in calculating a text's average cohesion score [5, 6] which is the semantic similarity score between sentences. Expanding on the methodologies outlined in this paper to calculate the average cohesion metric, we can evaluate whether it is also a linguistic characteristic difference that can be used as a feature in a future AI detection classifier. The last recommendation is to consider using different prompts in a future study. The dataset used in our study, though comprehensive, did not include all the different possible prompts provided to the middle and high school students. A future consideration is to acquire or use additional student prompts to see if there are any result differences to the post-hoc analysis conducted in this paper to get a better understanding on whether the linguistic characteristic difference with average semantic similarity between words score continues to hold between AI-generated and human-written essays.

7. REFERENCES

[1] Bašić, Ž., Banovac, A., Kružić, I., & Jerković, I. (2023). CHATGPT-3.5 as writing assistance in students' essays. *Humanities and Social Sciences Communications, 10*(1). https://doi.org/10.1057/s41599-023-02269-7

[2] Bouanani, N. (2015). Enhancing Critical Thinking Skills through Reflective Writing Intervention among Business College Students. *IOSR Journal of Research & Method in Education (IOSR-JRME),* 5(1). https://www.iosrjournals.org/iosr-jrme/papers/Vol-5%20Issue-1/Version-3/I05135055.pdf

[3] Quitadamo, I. J., & Kurtz, M. J. (2007). Learning to improve: Using writing to increase critical thinking performance in General Education Biology. *CBE—Life Sciences Education, 6*(2), 140–154. https://doi.org/10.1187/cbe.06-11-0203

[4] Herbold, S., Hautli-Janisz, A., Heuer, U., et al. (2023). A large-scale comparison of human-written versus ChatGPT-generated essays. *Scientific Reports, 13,* 18617. https://doi.org/10.1038/s41598-023-45644-9

[5] Ma, Y., Liu, J., Yi, F., Cheng, Q., Huang, Y., Lu, W., & Liu, X. (2023, February 12). AI vs. human -- differentiation analysis of Scientific Content Generation. *arXiv.org*. https://arxiv.org/abs/2301.10416

[6] Marzuki, Widiati, U., Rusdin, D., Darwin, & Indrawati, I. (2023). The impact of AI writing tools on the content and organization of students' writing: EFL teachers' perspective. *Cogent Education*, *10*(2). https://doi.org/10.1080/2331186x.2023.2236469

[7] Cooperman, S. R., & Brandao, R. A. (2024, February 15). *Ai tools vs AI text: Detecting ai-generated writing in foot and ankle surgery*. AI tools vs AI text: Detecting AI-generated writing in foot and ankle surgery. https://www.sciencedirect.com/science/article/pii/S2667396724000077?ref=pdf_download&fr=RR-2&rr=8792ccc2aa18ad6b

[8] Chan, C. K., & Hu, W. (2023). Students' voices on Generative AI: Perceptions, benefits, and challenges in Higher Education. *International Journal of Educational Technology in Higher Education*, *20*(1). https://doi.org/10.1186/s41239-023-00411-8

[9] Iskender, A. (2023). Holy or Unholy? Interview with Open AI's ChatGPT. *European Journal of Tourism Research, 34*, 3414. https://doi.org/10.54055/ejtr.v34i.3169

[10] Májovský1, M., Černý, M., Kasal, M., Komarc, M., Netuka1, D., Neurooncology, Department of Neurosurgery and, & Májovský, C. A. (2023, May 31). *Artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened*. Journal of Medical Internet Research. https://www.jmir.org/2023/1/e46924

[11] Májovský, M., Mikolov, T., Netuka, D., Neurooncology, Department of Neurosurgery and, & Májovský, C. A. (2023, August 31). Ai is changing the landscape of academic writing: What can be done? authors' reply to: Ai increases the pressure to overhaul the scientific peer review process. comment on "artificial intelligence can generate fraudulent but authentic-looking scientific medical articles: Pandora's box has been opened." Journal of Medical Internet Research. https://www.jmir.org/2023/1/e50844

[12] Slimani, T. (2013, October 30). *Description and evaluation of semantic similarity measures approaches*. arXiv.org. https://arxiv.org/abs/1310.8059