Detecting AI Authorship: Analyzing Descriptive Features for AI Detection

Christopher M. J. André[1,*,†], Helene F. L. Eriksen[1,†], Emil J. Jakobsen[1,†], Luca C. B. Mingolla[1,†] and Nicolai B. Thomsen[1]

[1]*Copenhagen Business School, Frederiksberg, Denmark*

Abstract

Motivated by the growing role of AI in text generation and the potential misuse of generative tools, this study investigates key features that differentiate AI-generated text from human-authored content. We produce a corpus of AI-generated counterparts to 2.100 research paper abstracts, in order to compare formal linguistic and stylometric characteristics such as perplexity, grammar, n-gram distributions and function word frequencies between human- and AI-generated texts. Key findings indicate that human-written abstracts tend to exhibit higher perplexity, greater grammatical error, and more diverse n-gram distributions. To distinguish between the two types of texts we employ various machine learning algorithms, with our Random Forest implementation achieving a precision of 0.986 on unseen data. Notably, feature importance analysis reveals that perplexity, grammar, and n-gram distributions are highly influential in AI-detection classification. Our research contributes a nuanced study of discriminating characteristics of AI-generated text to the increasingly important field of AI authorship attribution.

Keywords

Authorship attribution, Generative AI, Descriptive AI, AI authorship, GPT-3.5-turbo, Machine-learning.

1. Introduction

In the ever-evolving landscape of artificial intelligence (AI), the advent of Large Language Models (LLMs), championed by instruction-based ChatGPT, has attracted significant attention since late 2022. The remarkable capabilities of these pre-trained models enable them to generate coherent and arguably insightful paragraphs on virtually any conceivable topic. Notably, the recent successful completion of both the medical exam and uniform bar re-exam highlights the remarkable performance at a high academic level [2]. Consequently, this extraordinary performance raises important concerns regarding authorship attribution. In the education system, in particular, as students and researchers may be tempted to not only seek inspiration from such generative tools but to also claim its generated work as their own.

The motivation for this research came from the growing role of AI in text generation and its potential misuse. This paper extends current research on what defines AI-generated text and what strategies could be useful in tackling the related challenge of authorship attribution. In the following, we examine the differences between human-written and AI-generated text,

NL4AI 2023: Seventh Workshop on Natural Language for Artificial Intelligence, Nov 6-7th, 2023, Rome, Italy [1]

*Corresponding author.

†These authors contributed equally.

christopher@andre.bz (C. M. J. André)

and conduct a comparative study of text- and feature-based machine learning approaches for detecting AI-generated text.

Specifically, we analyze a corpus composed of 2.100 human-written abstracts and their AI-generated counterparts, created using the titles of the original abstracts in the prompts [1]. To evaluate the performance of the proposed modelling approaches, we favor precision – the proportion of correctly predicted AI-generated abstracts from all the predicted positive cases [3] - as, intuitively, the cost of a false positive is significantly higher than that of a false negative. The choice to solely focus on abstracts of research papers allows for a focused scope, and a specialized context for the model to consider and for us to understand.

2. Related Works

In the works of Gao et al.[4] they compared abstracts generated by ChatGPT to human-written abstracts, using both blinded human reviews and a GPT-2 Output Detector from OpenAI[2]. The output detector underscores the prospects of AI, correctly classifying 38% more AI-generated abstracts than humans. Due to the human involvement, the study was limited to 50 abstracts of each type, a restriction avoided in our case. Levin et al.[5] also used ChatGPT to generate 50 abstracts, analyzing the differences through Grammarly, a online typing assistant service. They found that ChatGPT had fewer grammatical errors and that it tends to use more unique tokens.

Various researchers have also generated larger corpora, in order to build more reliable AI authorship attribution models. Guo et al.[6] discovered that ChatGPT-written text tends to be more formal than humans, who write in a more colloquial language (using i.e. humor, slang, metaphors, and antiphrasis), while Mitrovic et al.[7] discuss the benefits of using an ML model over a perplexity score-based classification approach.

Uchendu et al.[8] explore different approaches to determining authorship attribution, on a corpus parted into eight sections with one human-written and seven different generative AIs. Using a version of the transformer model RoBERTa[9], fine-tuned with 20% of their data, they were able to achieve immaculate scores on GPT-2. Chen et al.[10] also utilize a RoBERTa model. The research showcases a strong classification model, achieving a high accuracy on their test dataset consisting of human-written and rephrased content generated by ChatGPT. While the benefits of multi-headed attention in NLP tasks are evident, the black box challenge of transformer models requires interpretability studies to extract and differentiate key features.

Gehrmann et al.[11] developed GLTR, a visualization tool for the identification of anomalies in texts. The tool analyzes what GPT-2 would have predicted at each token position. Trained on GPT-2, GLTR reports on any subject in its library, whereas our scope is narrowed to abstract structures.

[1]Full list of prompts is available at Github, URL: https://github.com/ChrisMJAndre/Detecting-AI-Authorship-Analyzing-Descriptive-Features-for-AI-Detection/blob/main/DatasetCreation.py

[2]HuggingFace: https://openai-openai-detector–w994j.hf.space/

3. Dataset

All analyses and experiments described in the paper were conducted in English. The linguistic features, including perplexity, grammar, n-gram distribution, type-token ratio, average token length, and frequency of function words, were examined within the context of the English language. The dataset[3] crafted for this research is composed of two parts: one part of human-written abstracts, and the other of AI-generated abstracts using GPT-3.5-turbo. The key challenge lies in finding any distinguishing features between the two sets. We chose to use abstracts from academic research papers, for both the human and AI-generated datasets. Abstracts generally conform to a more standardized structure and length, and the comparable datasets provide us with the ability to better understand distinguishing features. Based on Kirchner et al.[12] we set a minimum character limit of 500 for the abstracts, trying to balance the challenges of short texts with the computational load of processing larger documents.

The corpus of academic research papers was obtained via Kaggle from the arXiv Database an open-access archive managed by Cornell University[13]. For the human-written dataset, we sampled 10,000 papers, excluding cases where the abstracts contained mathematical equations, were under 500 characters, or where the paper was labeled as "no abstract", "withdrawn", or "et al." due to GPT-3.5-Turbo's limitations with writing references. We also ensured only papers updated before 2021-01-01 were included to eliminate potential GPT-3.5-Turbo-generated content. This left us with 2,100 human-written abstracts. Subsequently, when creating the AI-generated abstracts, we employed OpenAI's ChatCompletion GPT-3.5-turbo v. 0.27.6 model[4]. Model temperature was set to 0.7 to strike a balance between deterministic behavior and the creativity of generated content. For creative tasks, the most common temperatures lie within the 0.7 and 0.9 range[14], and in the context of academic abstracts, we wanted the AI to be coherent and logical, but also introduce some variability, which is why we chose the lower range of the commonly chosen temperatures. 'Top_p' was set to 1, the recommended setting for adjusted temperature setting. To simulate real-world queries and introduce further variation, we sample each query from a set of 60 slightly varying prompts. Ultimately, we had a dataset, with 2.100 human-written abstracts and 2.100 AI-generated abstracts. However, due to GPT-3.5-turbo occasional omission of abstracts, we excluded 147 AI abstracts, leaving us with a final dataset of 2,100 human-written and 1,953 AI-generated abstracts.

4. Feature Understanding

The challenge in detecting AI-generated text lies in its resemblance to human writing. Hence, we aimed to contribute to the current understanding of how specific linguistic characteristics differentiate both AI and human-written text. Specifically, we conduct a comparative study of how perplexity, certain grammatical structures, n-gram distributions, type-token ratios, as well as the stylometric features average token length and frequency of function words, vary between human- and AI-generated text.

[3]The dataset is available at Kaggle: https://www.kaggle.com/datasets/heleneeriksen/gpt-vs-human-a-corpus-of-research-abstracts

[4]The code is available at GitHub: https://github.com/ChrisMJAndre/Detecting-AI-Authorship-Analyzing-Descriptive-Features-for-AI-Detection

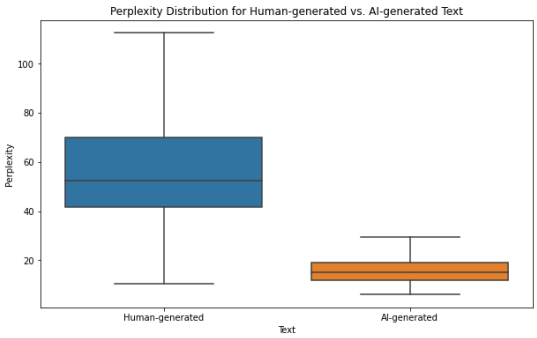


Figure 1: Perplexity distribution for human vs. AI-generated text

4.1. Perplexity

Perplexity quantifies the ability of a given probabilistic model to predict a given sample. Effectively, it encapsulates the uncertainty of a model's prediction: The lower the perplexity score, the higher the model's confidence in its outputs[15].

In comparing the AI-generated and human-written abstracts, we hypothesized that a perplexity score might capture subtle differences in token-level predictability, and thereby randomness of tokens used in both types of abstracts. Given that state-of-the-art generative models like ChatGPT are fine-tuned to minimize perplexity[16], we would expect this metric to be comparatively lower for AI-generated abstracts.

To examine our hypothesis, we used the GPT-2 model, an autoregressive language model from OpenAI through the HuggingFace API [5]. This pre-trained model made it possible to compute perplexity scores for each text sample within our corpus, offering a consistent and reliable metric to understand the token-over-token predictability of each abstract.

Our analysis revealed that human-written abstracts consistently show a higher perplexity than their AI-generated counterparts, as shown in Figure 1.

This implies that, based on the vocabulary of GPT-2, human writing tends to be significantly more diverse at a token level. The AI-generated abstracts consistently exhibit higher conformity, leading to comparatively higher levels of token-over-token predictability. This finding shows that human-written abstracts consistently register higher perplexity scores than their AI-generated counterparts, which, in turn, suggests perplexity as a valuable discriminator in AI authorship attribution.

While the vast vocabulary of generative models like GPT-3.5-turbo allow them to produce coherent and contextually relevant text, they inherently conform to patterns of high statistical likelihood, leading them to exhibit token-level regularity and predictability. In contrast, human writing introduce novelty. This results in comparatively lower perplexity scores for AI-generated text.

[5]https://huggingface.co/docs/transformers/perplexity

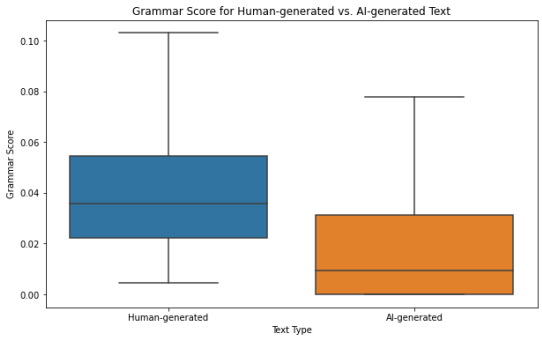


Figure 2: Grammar score

4.2. Grammar

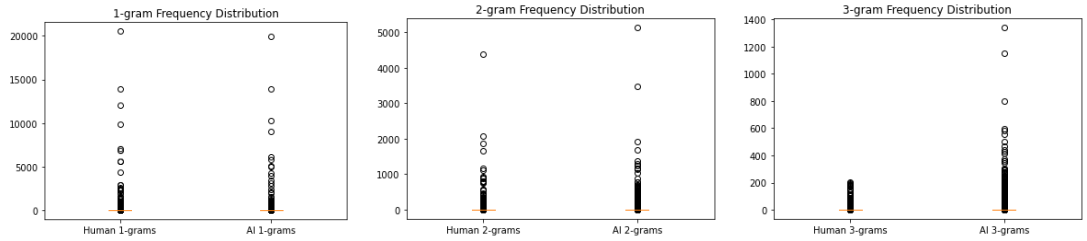
Grammar refers to the set of structural rules that dictate the composition of sentences, phrases, and words in any language. Since LLMs like GPT-3.5-turbo are trained on vast corpora of text , they typically exhibit high grammatical accuracy [17]. Our hypothesis is that by incorporating grammatical analysis as a feature, we might capture discrepancies between human and AI-generated abstracts.

Using language_tool_python, an open-source library for detection of grammatical and spelling errors, we identified the number of errors in each abstract in our corpus. This enabled us to compute a grammar score. The grammar score is calculated by taking the number of errors detected by the language_tool_python and dividing it by the number of tokens in each abstract. Our comparison of grammatical correctness in human- and AI-generated texts reveals that AI-generated abstracts consistently contain fewer grammatical errors, as shown in Figure 2.

The grammatical aptitude of GPT-3.5-turbo may be attributed to the sheer volume of data on which it is trained. Exposure to large amounts of textual data enable LLMs to understand, or convincingly mimic, the rules and structures of languages.

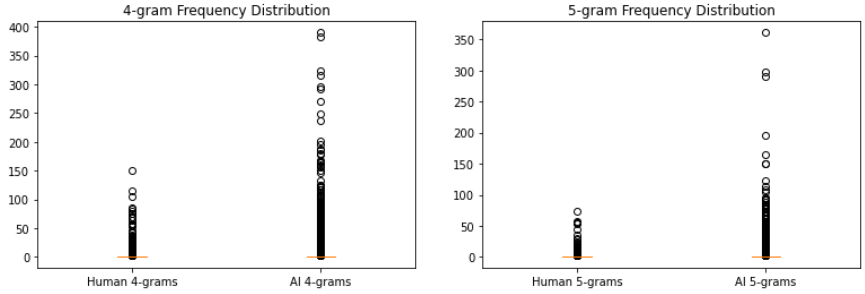
4.3. N-Gram Distributions

Although eclipsed by emergent LLMs, n-grams[18] remain a reliable technique for the analysis of linguistic patterns. We hypothesize that LLMs may favor specific token sequences due to their training data, which will lead to a higher number of n-grams compared to in human-written abstracts. To analyze the distribution of n-grams in our corpus we tokenize the texts using NLTK, and log the prevalence of n-grams for each value of n in the range [1,7] for each individual abstract. These scores are then aggregated across all texts to understand the overall distribution. The intent behind n-gram distribution is to discern token sequences that might be comparatively more prevalent in either human-written texts or AI-generated texts. Our findings, as visualized in figures 3-5, show that AI-generated abstracts exhibited a higher frequency of the same n-grams, especially in higher n-gram ranges compared to human-written abstracts. The increasing frequency of identical n-grams in AI-generated abstracts, particularly in the higher n-gram ranges, offers interesting insights into the behavior of generative models. Notably, the



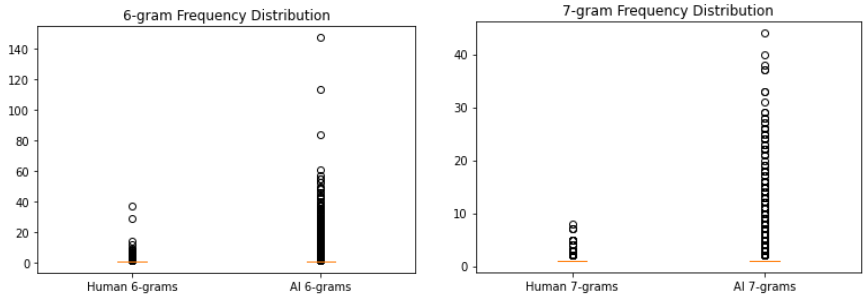
(a) 1-grams (b) 2-grams (c) 3-grams

Figure 3: (a) Distribution of 1-grams. (b) Distribution of 2-grams. (c) Distribution of 3-grams.



(a) 4-grams (b) 5-grams

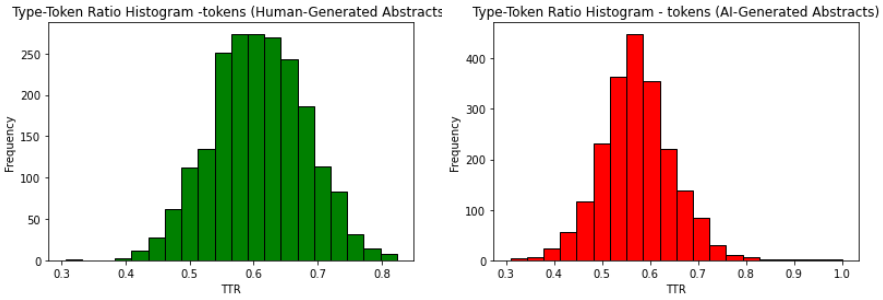
Figure 4: (a) Distribution of 4-grams. (b) Distribution of 5-grams.



(a) 6-grams (b) 7-grams

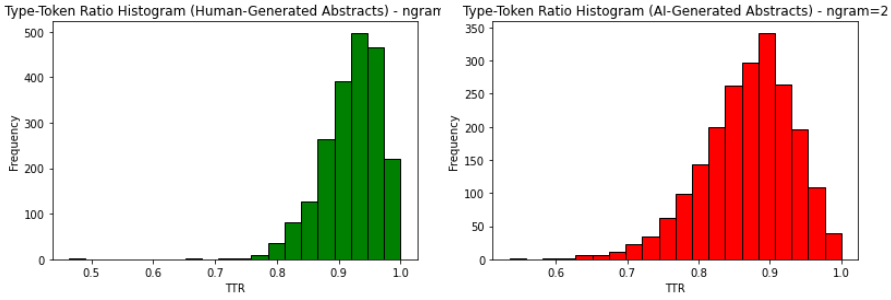
Figure 5: (a) Distribution of 6-grams. (b) Distribution of 7-grams.

disparity in n-gram distributions, most visible in the pronounced difference from the 3-gram range onward, underscores the conformity of current AI-generated text. LLMs are trained and evaluated on their ability to identify statistically significant patterns, which becomes particularly evident in the dataset by the repetition of certain 5-grams (>100 occurrences) and 7-grams (>20 occurrences). Such frequent repetitions indicate that the model has identified these sequences as ”safe bets” for language generation, resulting in their high use. Conversely, the more varied distribution of n-grams in human-written abstracts reflect the relatively more nuanced ways



(a) Human 1-gram (b) AI 1-gram

Figure 6: (a) Type-token ratio for human text (b) Type-token ratio for AI text



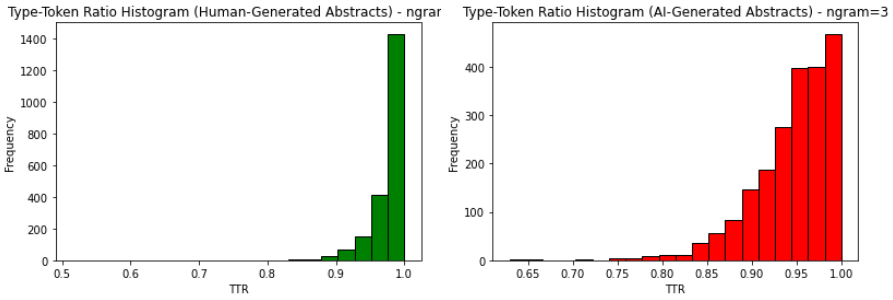
(a) Human 2-gram (b) AI 2-gram

Figure 7: (a) Type-token ratio for human text (b) Type-token ratio for AI text

humans tend to express themselves, especially for higher values of n. Humans draw from unique perspectives, which leads to a richer diversity in phrasing and structuring as evident in our findings. The relative scarcity of repetitive n-grams in human abstracts suggests a broader array of stylistic choices. Within the field of AI authorship attribution, our findings suggest that n-gram distributions can serve as a discriminant measure for detection, as high recurrence of specific higher-order n-grams might be indicative of AI authorship.

4.4. Type-Token Ratio

Type-Token Ratio (TTR) is used to measure the lexical diversity within a text[19]. In this research, we focus on TTR with different n-gram sizes. This design choice allows us to analyze lexical diversity at different scales, looking into choices of individual tokens as well as the complexity of a three-token sentence. In our research, we chose to use type-token ratio to further understand the n-gram distribution in our feature-based model, as TTR produces a descriptive feature at an individual abstract level. Like with n-gram distributions, our hypothesis was that LLMs would exhibit lower TTR compared to human-written texts. Our analysis reveals a pronounced difference in the TTR between AI-generated and human-written abstracts, particularly for higher range n-grams, as shown in Figure 6, Figure 7, and Figure 8. Specifically,



(a) Human 3-gram (b) AI 3-gram

Figure 8: (a) Type-token ratio for human text (b) Type-token ratio for AI text

AI-generated abstracts exhibited a lower TTR score, indicating less variety in both the tokens and structures used. This concurs with the conclusions of the n-gram distributions.

4.5. Average Token Length and Frequency of Function Words

The average token length offers insights into tendencies in writing styles. Upon analysis, AI-generated abstracts display a consistent bias towards longer tokens relative to those used by humans[6], possibly a result of the LLMs comprehensive vocabulary, or its formal word choice in general. Function words, primarily comprising of prepositions, pronouns, and conjunctions, serve as the backbone for grammatical relationships within sentences, but bear minimal lexical meaning. Our hypothesis aligned with the findings of Boukhaled & Ganascia [20], which illustrate that frequency could be used in discerning stylistic nuances for authorship attribution. Our results show a broader distribution of function words in human-written abstracts compared to their AI-generated counterparts, hinting at differences in their diversity of writing styles[7].

5. Modeling

Based on our analyses, we construct two variants of machine learning approaches for AI authorship attribution: A text-based approach and a feature-based approach. The former relies solely on the text of the abstracts, while the latter uses the precomputed features described in our analyses. The text-based approach is employed as a benchmark to assess the efficacy of the feature-based approach, given the complexities of deriving a comprehensive view solely from a textual analysis of abstracts. We evaluate the two approaches using three machine learning algorithms for classification: Logistic Regression, Random Forest, and Multinomial Naïve Bayes. A most-frequent label classifier baseline, i.e., a zero-rule baseline strategy of predicting the majority class, is included for comparison

[6]graphs available on Github: Visualizations/Average Token Lenght
[7]graphs available on Github: Visualizations/ Distribution of function words

Table 1
Model Performance Metrics

Model	Set	Accuracy	Precision	Recall	F1-score
Random Forest	Train	0.993303	0.992811	0.993525	0.993168
Random Forest	Validation	0.970395	0.963636	0.970696	0.967153
Random Forest	**Test**	0.983553	**0.986111**	0.979310	0.982699
Logistic Regression	Test	0.981908	0.982699	0.97911	0.981002
Most-frequent label classifier baseline	Test	0.501645	0.478827	0.506897	0.492462
MultinomialNB	Test	0.978618	0.975945	0.978310	0.977625

5.1. Feature-based Approach

The feature-based approach utilizes the precomputed features to predict authorship attribution, namely: perplexity, grammar, type-token ratio for 1-, 2-, and 3-grams, frequency of function words, and average token length. Per our earlier analyses, these features are shown to hold significant discriminatory value. Hyperparameter tuning was performed using grid search with 5-fold cross-validation, optimizing for precision scores. [8] Post-training, we extracted the most influential features from the top-performing model.

5.1.1. Evaluation

Our models report consistently high results with Random Forest achieving a precision score of 0.986 on the test data[9], as shown in Table 1. While other classifiers like Logistic Regression and MultinomialNB were competent, they were unable to outperform Random Forest.

5.1.2. Feature Importance

Upon examination of the feature importance calculated based on the top-performing Random Forest model, as shown in Table 2, we observe that perplexity dominates the decision function with a weighted score of 0.71. Next, the grammar score, and TTR_3ngram had significant influence on the predictions. The grammar scores' significant importance suggest that the grammatical structure of the abstract is a distinguishing factor. It's worth noting that abstracts might be comparatively less prone to grammatical errors, as a result of the process of peer-review before publication. Arguably, the importance of grammar score may be even higher in the classification of texts from other domains. Our results indicate that both perplexity, grammatical structures and type-token ratios could be key features for distinguishing between human and AI-generated abstracts.

[8]Hyperparameters for the best performing model: max_depth: 50, min_samples_split: 10, n_estimators: 20.
[9]Distribution of dataset: Train (70%), Test (15%), Validation(15%)

Table 2
Feature Importance from Random Forest

Feature Importance	
Perplexity	0.709898
Grammar Score	0.100167
TTR_3ngram	0.095967
TTR_1ngram	0.029586
Average token length	0.028081
TTR_2ngram	0.026615
frequency_of_function_words	0.009686

5.2. Text-based Approach

For comparison, we also implemented a text-based approach. During this process the abstract text underwent normalization, tokenization, stop word removal[10], and punctuation elimination. We employed the TF-IDF vectorizer to numerically transform the text, emphasizing n-grams to capture token sequence contexts. Like our primary approach, we optimized hyperparameters through grid search with 5-fold cross-validation, and based model selection on precision scores. [11] The Logistic Regression model proved superior, exhibiting a test precision score of 0.988.

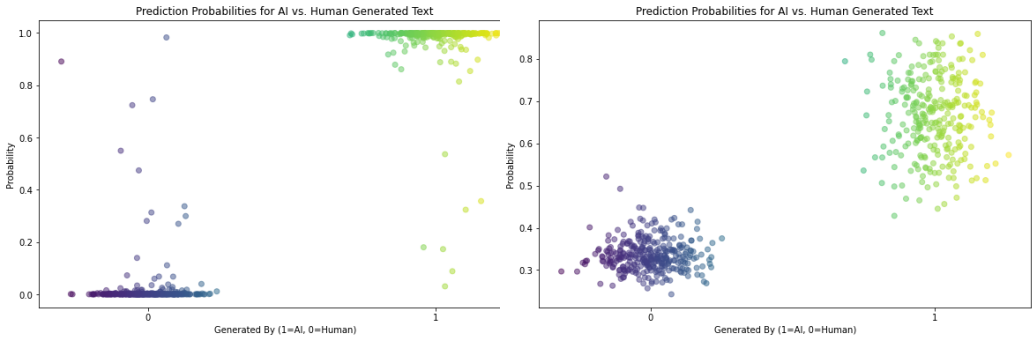
6. Overall Model Evaluation

Overall, the evaluation of the text-based and feature-based approaches provide insights into discriminative differences between human and AI-generated abstracts. Both approaches led to highly impressive results when evaluated on precision. However, when observing the log probabilities of the predictions, the feature-based approach show a substantially more robust discrimination between the two classes, as seen in Figure 9. This is demonstrated by the predicted probabilities which are primarily clustered near the extremes, either 0 or 1, suggesting a strong confidence in the classification. Conversely, the text-based approach's predicted log probabilities exhibited a more evenly spread distribution, mainly residing around 0.3 and 0.7, indicating a lesser degree of certainty in its predictions.

The feature-based approach, especially with the incorporation of the perplexity feature, offers a robust and reliable indication of whether an abstract is AI-generated or not. In contrast, the text-based model has multiple predictions with probabilities around 0.5, meaning the classification is largely an arbitrary decision between the two classes. Samples within 0.4 to 0.6 could be considered likely "flukes", where the model is closer to guessing that predicting.

The high performance of the feature-based approach across training, test, and validation data speaks to the performance and generalizability of our approach. Overfitting, a common pitfall in machine learning, typically manifests as high performance on the training data, but a significant dip in performance on unseen data. However, our model's ability to maintain consistent high

[10]Using a custom list, Github File: DataUnderstanding & Modeling.py
[11]Hyperparameters for best performing model: C: 0.1, penalty: 12, solver: liblinear, tfiaf_max_df': 0.9, tfidf_ngram_range: (5,6).



(a) Feature-based approach (b) Text-based approach

Figure 9: Difference in Prediction probabilities for AI (1) vs human-generated text(0) from Text-based model and Feature-based model

precision across all datasets counters this issue, indicating that it is not merely memorizing the training data, but learning underlying patterns that apply to new unseen data. Conversely, the attainment of exceptionally high performance across training, test, and validation data, such as 95%+ precision, raises some considerations that need further consideration. The simplicity of our task, or the informative nature of the chosen features, could explain the exceptionally high performance.

7. Discussion

While perplexity is the most deciding factor for the feature-based approach, as shown in section 4.1 and 5.1.2, the low perplexity scores for the AI-generated abstracts may be due to the narrow scope of the research. This suggests the need for a multi-domain evaluation. Uchendu et al. [8] highlights a similar challenge, caused by using a single topic corpus and argues that basic topical analysis cannot differentiate between human- and AI-generated text across domains. Yet, as our domain, research paper abstracts, span multiple topical domains, the domain influence may be reduced. Similarly, the establishment of a reliable character minimum for AI authorship attribution is still an open research question. Kirchner et al. [12] recommend a 1,000-character threshold, which conflicts with the structures of most abstracts, due to their concise and formulaic format.

7.1. Limitations

One notable limitation of our study is the use of perplexity to differentiate human-authored abstracts from those generated by GPT-3.5-turbo. The potential limitation arises due to the shared lineage between GPT-2, used to calculate perplexity, and GPT-3.5-turbo. The similarities in training data might introduce biases in our results, and different alternatives to calculate perplexity could have yielded varying perplexity scores due to differences in training nuances. A further limitation in our data generation and modeling process is the absence of a specified random seed, which affects the replicability of our findings. The lack of a fixed seed for random

number generation introduces variability between runs, potentially influencing the consistency of our results.

8. Future Research

The selected temperature setting in LLMs holds considerable influence over both the variability and coherence of the generated text. In this study, we chose a temperature value of 0.7 to strike a balance between coherence and textual variation. It is important to note, however, that this choice—falling in the lower spectrum of the 0.7 to 0.9 range—may limit diversity and perhaps induce slight repetition, as corroborated by the frequent n-gram repetitions discussed in section 4.3. For future research, we recommend examining the effects of varying temperature settings on text generation, particularly focusing on aspects like perplexity and repetition. A more nuanced understanding of temperature's influence on text generation could offer valuable insights into the patterns of AI-generated text and improve methods for authorship attribution. To facilitate a potential in-depth exploration of our feature-based approach's performance, we have highlighted the top 3 most challenging classifications for both AI and human abstracts available at GitHub[12]. This data would allow future research to examine what feature combinations that present challenges for detecting AI authorship. Furthermore, it is important to acknowledge that the text generated by LLMs can be altered to avoid detection. Techniques such as swapping out key phrases, inserting unexpected vocabulary, or embedding grammatical errors into the abstracts could all impact the probability of the abstracts being misidentified as human writing. Considering the societal impact, it remains imperative that the field of AI authorship attribution develops in parallel with the evolution of LLMs.

9. Conclusion

In the rapidly progressing field of artificial intelligence, the emergence of sophisticated generative models poses both opportunities and challenges. This research sheds light on potential key differences in human-written and AI-generated academic abstracts. Through a comprehensive analysis of various textual features like perplexity, grammar, n-gram distributions, and type-token ratios, the study reveals clear discriminatory patterns. The AI-generated texts are shown to exhibit higher token-level predictability and grammatical accuracy when compared to their human-written counterparts. Our approaches, especially the feature-based one, demonstrate remarkable precision and certainty in distinguishing between human and AI-generated content. Such precision emphasizes the significance of the chosen features, with perplexity standing out as a particularly influential metric in tackling the challenge of AI authorship attribution.

Acknowledgments

We would like to express our gratitude to professor Daniel Hardt for his invaluable guidance, support, and mentorship throughout the duration of the research project.

[12]Github File: Worst classified AI and Human Abstracts.xlsx

References

[1] E. Bassignana, D. Brunato, M. Polignano, A. Ramponi, Preface to the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI), in: Proceedings of the Seventh Workshop on Natural Language for Artificial Intelligence (NL4AI 2023) co-located with 22th International Conference of the Italian Association for Artificial Intelligence (AI* IA 2023), 2023.

[2] OpenAI, Gpt-4 technical report, 2023. arXiv:2303.08774.

[3] A. Bamania, When should you use accuracy, precision, recall & f-1 score?, https://levelup.gitconnected.com/4-important-metrics-for-classification-machine-learning-models-when-how-to-use-them-6aa7c85d7665, 2022. Last accessed: September 10, 2023.

[4] C. A. Gao, F. M. Howard, N. S. Markov, E. C. Dyer, S. Ramesh, Y. Luo, A. T. Pearson, Comparing scientific abstracts generated by chatgpt to real abstracts with detectors and blinded human reviewers, npj Digital Medicine 6 (2023) 1–5. URL: https://doi.org/10.1038/s41746-023-00819-6. doi:10.1038/s41746-023-00819-6.

[5] L. G, M. R, K. E, B. Y., Identifying chatgpt-written obgyn abstracts using a simple tool., Am J Obstet Gynecol MFM 5 (2023). URL: https://pubmed.ncbi.nlm.nih.gov/36931435/. doi:10.1016/j.ajogmf.2023.100936.

[6] B. Guo, X. Zhang, Z. Wang, M. Jiang, J. Nie, Y. Ding, J. Yue, Y. Wu, How close is chatgpt to human experts? comparison corpus, evaluation, and detection, 2023. arXiv:2301.07597.

[7] S. Mitrović, D. Andreoletti, O. Ayoub, Chatgpt or human? detect and explain. explaining decisions of machine learning model for detecting short chatgpt-generated text, 2023. arXiv:2301.13852.

[8] A. Uchendu, T. Le, K. Shu, D. Lee, Authorship attribution for neural text generation, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020, pp. 8384–8395. URL: https://aclanthology.org/2020.emnlp-main.673. doi:10.18653/v1/2020.emnlp-main.673.

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, 2019. arXiv:1907.11692.

[10] Y. Chen, H. Kang, V. Zhai, L. Li, R. Singh, B. Raj, Gpt-sentinel: Distinguishing human and chatgpt generated content, 2023. arXiv:2305.07969.

[11] S. Gehrmann, H. Strobelt, A. M. Rush, Gltr: Statistical detection and visualization of generated text, CoRR abs/1906.04043 (2019) 1–6. URL: http://arxiv.org/abs/1906.04043. doi:arXiv:1906.04043.

[12] J. Kirchner, L. Ahmad, S. Aaronson, J. Leike, New ai classifier for indicating ai-written text, 2023. URL: https://openai.com/blog/new-ai-classifier-for-indicating-ai-written-text, last accessed: September 10, 2023.

[13] C. University, J. Tricot, devrishi, B. Maltzan, S. Brinn, arxiv dataset, 2023. URL: https://www.kaggle.com/datasets/Cornell-University/arxiv, last accessed: September 10, 2023.

[14] J. A. Kolar, A simple guide to setting the gpt-3 temperature, 2020.

[15] M. Bernstein, Perplexity: a more intuitive measure of uncertainty than entropy., 2021. URL: https://mbernste.github.io/posts/perplexity/, last accessed: September 10, 2023.

[16] R. Kadiyala, Medium.com: Gptzero vs chatgpt — a gray story., 2023. URL: https://

raj-k-kadiyala.medium.com/gptzero-vs-chatgpt-a-gray-story-901b825e0666, last accessed: September 10, 2023.

[17] H. Wu, W. Wang, Y. Wan, W. Jiao, M. Lyu, Chatgpt or grammarly? evaluating chatgpt on grammatical error correction benchmark., 2023. Last accessed: September 10, 2023.

[18] S. Srinidhi, Medium: Understanding word n-grams and n-gram probability in natural language processing., https://towardsdatascience.com/understanding-word-n-grams-and-n-gram-probability-in-natural-language-processing-9d9eef0fa058, 2019. Last accessed: September 10, 2023.

[19] D. Thomas, Type-token ratios in one teacher's classroom talk: An investigation of lexical complexity, 2005. Last accessed: September 10, 2023.

[20] M. Boukhaled, J.-G. Ganascia, Authorship attribution, 2017. URL: https://www.sciencedirect.com/topics/computer-science/authorship-attribution, last accessed: September 10, 2023.