Text Authorship Attribution: Stylometric Insights into Human and LLM-Generated Text

Shifali Agrahari
a.shifali@iitg.ac.in
Indian Institute of Technology
Guwahati, Assam, India

Samridhi Bisht
samridhibishtt@gmail.com
Vivekananda Institute of Professional
Studies, Delhi, India

Sanasam Ranbir Singh
ranbir@iitg.ac.in
Indian Institute of Technology
Guwahati, Assam, India

Abstract

The widespread use of large language models (LLMs) has raised concerns about the potential misuse of AI-generated text for deceptive purposes, such as disinformation and spam. While there is research on detecting human and AI-generated text, specific investigations into the differences between texts generated by individual LLMs remain limited. This study aims to explore the distinctions between human-written text and text produced by various LLMs, including Gemini, GPT-Neo, Falcon, LLaMA, and Bloom. Through a comprehensive analysis of lexical and stylistic features, we have found that LLM-generated text tends to be longer, more structured, and less lexically diverse than human-written content. Cross-model classification experiments revealed that although models like Gemini and ChatGPT closely align with human text, detecting AI-generated content across different models remains a challenge. Our findings underscore the need for better detection techniques to effectively distinguish between human and AI-generated text.

Keywords

Text Authorship, LLMs, AI Text, Stylometric Features

ACM Reference Format:
Shifali Agrahari, Samridhi Bisht, and Sanasam Ranbir Singh. 2024. Text Authorship Attribution: Stylometric Insights into Human and LLM-Generated Text. In *8th International Confernce on Data Science and Management of Data (12th ACM IKDD CODS and 30th COMAD) (CODS-COMAD Dec '24), December 18–21, 2024, Jodhpur, India.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3703323.3703712

1 Introduction

Recent advancements in large language models (LLMs) have led to their widespread use across diverse applications, resulting in the development of a variety of LLMs. While many applications of LLMs are benign, growing concerns surround their misuse for deceptive purposes such as phishing attacks, disinformation, spam, and plagiarism. As such, distinguishing LLM-generated text from human-written text is increasingly important, particularly given the proliferation of different LLMs. Despite the growing body of research focused on differentiating AI- and human-generated text, there has been limited exploration of how text produced by various LLMs compares to one another and how their features differ from human-written content. Existing studies primarily focus on statistical and stylistic features, such as word count, vocabulary richness, and readability [9, 3, 8, 7, 6, 4], or on zero-shot detection methods using metrics like top-k, top-p, and entropy [1]. Additionally, black-box methods, leveraging machine learning and deep learning models, have been employed to detect AI- and human-generated text [5]. However, a few crucial questions remain: *Can detectors trained on text from one LLM effectively identify text generated by other LLMs? How stylometically similar or dissimilar are the texts generated by different LLMs?* To address the aforementioned questions, this paper focuses on exploring and analyzing text generated by five LLMs: Gemini, GPT-Neo-2.7B, Falcon-7B-Instruct, LLaMA3-70B-8192, and Bloom-1B7.

CODS-COMAD Dec '24, December 18–21, 2024, Jodhpur, India

ACM ISBN 979-8-4007-1124-4/24/12
https://doi.org/10.1145/3703323.3703712

2 Dataset

For our task, we have utilized the HC3 dataset [2], which contains data from five different domains. Each domain dataset consists of three columns: a question, a human-generated answer, and a Chatgpt-generated answer. For this task, we have only considered Open Question Domain. To extend this dataset, we have generated additional answers using five different LLMs: Gemini, GPT-Neo-2.7B, Falcon-7B-Instruct, LLaMA3-70B-8192, and Bloom-1B7. For each question in the dataset, we have provided the same prompt, "Answer the following question: <Question>," to these models to generate their respective responses.

Table 1: Lexical analysis results based on the features outlined in the paper [6]

Feature	Human	Chatgpt	Gemini	GPT-neo	Falcon	Llama	Bloom
# Character	247.63	2246.18	1245.47	1788.25	2101.91	402.72	3687.01
# Words	40.35	365.34	183.44	315.19	340.20	63.44	661.44
# Sentence	1.09	13.47	11.12	23.22	19.20	4.55	45.04
# Quotation	0.59	3.91	1.74	8.29	2.61	0.85	11.46
# Unique Words	28.92	122.05	98.06	93.91	113.14	44.59	143.82
Unique Words Rel.	0.75	0.35	0.60	0.33	0.37	0.72	0.23
# Special Char	7.01	36.76	58.65	28.64	34.14	8.66	50.14
Personal Pronoun Rel.	0.01	0.02	0.02	0.04	0.03	0.02	0.06
# Personal Pronoun	0.45	10.00	6.33	17.23	11.49	1.88	46.43
# Grammar Error	3.22	7.80	3.39	11.88	10.10	0.81	22.91

3 Feature Analysis

We have focused on two types of analysis: lexical and stylometric. The goal of this analysis is to examine the differences between human-written text and text generated by various LLMs, as well as to understand how the features of text generated by two different LLMs compare.

Stylometric features are used to identify different stylistic patterns in a text, particularly for detecting shifts in writing style within literary works. In text analysis, two main types of features are typically considered: statistical and stylistic. Statistical features include lexical aspects such as average word length, function word count, punctuation, and sentence length. Stylistic features focus on readability, vocabulary diversity, and richness.

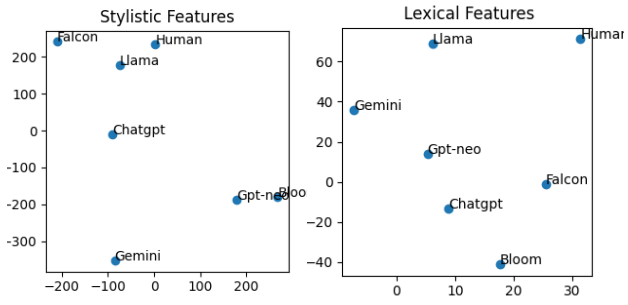

Figure 1: t-SNE visualization of feature across models

Table 2: Stylistic feature analysis results of the dataset based on the features outlined in the paper [6, 8]

Feature	Human	Chatgpt	Gemini	Gpt-neo	Falcon	Llama	Bloom
Flesch reading ease	49.55	50.06	45.86	82.14	62.27	53.31	80.72
Flesch kincaid grade	11.7	11.5	11.1	5.4	8.9	10.3	6
Gunning Fog	12.34	9.47	8.76	6.44	7.82	9.63	6.78
Dale Chall	9.14	5.86	6.55	5.18	5.49	7.5	1.27
Smog Index	13.9	13.3	13.3	9.2	12	12.5	9.5
Automated Readability	14.8	13.4	14.3	6.7	11.5	11.8	7.1
Linsear Write Formula	18.5	14.8	13.5	5.44	5.77	13.4	5.11
Coleman Liau Index	12.3	11.31	13.45	6.72	10.84	10.73	6.67
TTR	0.28	0.08	0.11	0.05	0.06	0.19	0.03
Log TTR	0.86	0.78	0.8	0.75	0.76	0.83	0.72
Root TTR	34.48	29.64	30.82	19.83	23.28	29.92	16.84
Mass TTR	24.38	20.96	21.79	14.02	16.46	21.15	11.91
Herdan C	0.86	0.78	0.8	0.75	0.76	0.83	0.72
MAAS	0.0135	0.0179	0.0173	0.0209	0.0197	0.0161	0.0219
Yule K	118.51	140.63	345.58	133.62	91.81	124.4	133.92

3.1 Lexical Features

One way to distinguish AI-generated text from human-written text is through its lexical structure as mentioned in the study [6]. Table 1 provides a comparative lexical analysis between text generated by large language models (LLMs) and human-written text. Key metrics include character, word, sentence, and quotation counts, as well as unique word counts and ratios, which assess lexical diversity. Additional indicators such as special character counts, personal pronoun usage, and ratios help identify conversational tones. The number of difficult words and grammar errors further highlight complexity and fluency differences between AI and human text.

3.2 Stylistic Features

Readability scores and measures of vocabulary diversity assess the reading level of a text, often based on the education level required for easy comprehension. Human and AI-generated texts differ in readability feature. as mentioned in the study[6, 8] due to varying approaches to language. Human texts consider factors like cultural relevance, context, and style, while AI models focus on textbook definitions, leading to differences even when conveying the same meaning. Table 2 readability formulas like Flesch Reading Ease, Flesch-Kincaid Grade, Gunning Fog, and Dale-Chall evaluate sentence length and word difficulty. Other indices, including the Smog Index, Automated Readability Index, and Linsear Write Formula, focus on text complexity. Lexical diversity measures like Type-Token Ratio (TTR) and Yule's K provide further insight into vocabulary richness, with higher values indicating more complex, varied language use.

Table 3: Accuracy (Acc) and F1 Scores of RoBERTa-based Model for classification of human vs each LLMs Models

Open	GPT-neo		ChatGPT		Falcon		Llama3		Gemini	
Question	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
GPT-neo	**99.3**	0.95	99.51	0.98	56.14	0.22	92.33	0.92	80.35	0.76
ChatGPT	67.08	0.51	99.98	0.99	53.99	0.15	50.06	0.00	52.64	0.10
Falcon	72.69	0.63	99.44	0.99	99.40	0.89	51.07	0.05	63.56	0.63
Llama3	94.29	0.94	98.12	0.98	72.34	0.62	97.27	0.89	69.62	0.57
Gemini	89.08	0.88	99.66	1.00	73.06	0.63	51.03	0.05	99.00	0.94

4 Detectors

To classify human-written text versus text generated by LLMs, we split each dataset into a 1:4 ratio for training and testing. We used the RoBERTa model to train classifiers to distinguish between human text and text generated by each LLM. After training, we tested our detector on the remaining LLM-generated text to evaluate its generalization across different models.

5 Results Analysis

In response to above question, the analysis in Table 1 and Fig. 1 shows that AI text, especially from GPT-Neo and Bloom, is longer and more structured compared to human-written text, with higher character, word, and sentence counts. These models also use more quotations, special characters, and personal pronouns, contributing to a conversational tone. Despite generating more unique words, AI text has less lexical diversity than human text and exhibits more grammatical errors, particularly in GPT-Neo and Bloom.

In terms of stylistic features Table 2, human text and ChatGPT have similar readability scores around 50, whereas GPT-Neo and Bloom produce text with readability scores above 80, indicating simpler language. Human text requires a higher Flesch-Kincaid grade level for comprehension and shows greater lexical richness, with higher TTR and Herdan C values. Gemini and Bloom generate more repetitive text, reflected in their lower TTR. Overall, AI models are easier to read but lack lexical diversity of human writing.

Table 3 shows that models perform well in detecting their own outputs, with GPT-Neo, ChatGPT, and Gemini achieving high accuracy and F1 scores. However, cross-model detection is challenging, with significant performance drops when classifying outputs from other models, particularly Falcon and LLaMA3. Gemini and GPT-Neo show better generalization across models, while ChatGPT struggles the most in cross-model scenarios.

6 Conclusion & Future work

This study examines differences between human-written text and text generated LLMs.The findings indicate that LLM-generated content is generally longer, more structured, and less diverse than human text. Our future work aims to enhance text detection techniques by incorporating advanced stylistic and contextual features. We plan to expand our dataset to include a broader range of LLMs and human-authored content across various domains, which will improve the generalizability of detection methods. Additionally, we will explore the integration of real-time detection systems to better address the risks associated with the misuse of AI-generated text. This approach will help in developing more effective tools to differentiate between human and AI-generated content in practical applications.

References

[1] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. 2019. Gltr: statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*.

[2] Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How close is chatgpt to human experts? comparison corpus, evaluation, and detection. *arXiv preprint arXiv:2301.07597*.

[3] Tharindu Kumarage, Joshua Garland, Amrita Bhattacharjee, Kirill Trapeznikov, Scott Ruston, and Huan Liu. 2023. Stylometric detection of ai-generated text in twitter timelines. *arXiv preprint arXiv:2303.03697*.

[4] Tharindu Kumarage and Huan Liu. 2023. Neural authorship attribution: stylometric analysis on large language models. In *2023 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC)*. IEEE Computer Society, 51–54.

[5] George K Mikros, Athanasios Koursaris, Dimitrios Bilianos, and George Markopoulos. 2023. Ai-writing detection using an ensemble of transformers and stylometric features. In *IberLEF@ SEPLN*.

[6] Lorenz Mindner, Tim Schlippe, and Kristina Schaaff. 2023. Classification of human-and ai-generated texts: investigating features for chatgpt. In *International Conference on Artificial Intelligence in Education Technology*. Springer, 152–170.

[7] Hoang-Quoc Nguyen-Son, Ngoc-Dung T Tieu, Huy H Nguyen, Junichi Yamagishi, and Isao Echi Zen. 2017. Identifying computer-generated text using statistical analysis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 1504–1511.

[8] Aditya Shah, Prateek Ranka, Urmi Dedhia, Shruti Prasad, Siddhi Muni, and Kiran Bhowmick. 2023. Detecting and unmasking ai-generated texts through explainable artificial intelligence using stylistic features. *International Journal of Advanced Computer Science and Applications*, 14, 10.

[9] Irene Solaiman et al. 2019. Release strategies and the social impacts of language models. *arXiv preprint arXiv:1908.09203*.

www.Grammarly.com/edu

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009