HOMOGENEITY OF TOKEN PROBABILITY DISTRIBUTIONS IN CHATGPT AND HUMAN TEXTS

Dragica Ljubisavljević*[1], Marko Koprivica*[1], Aleksandar Kostić[2] and Vladan Devedžić[1]
[1]*Faculty of Organizational Sciences, Department of Software Engineering, Belgrade University*
Jove Ilica 154, Belgrade, Serbia
[2]*Faculty of Philosophy, Department of Psychology, Belgrade University*
Čika-Ljubina 18-20, Belgrade, Serbia

ABSTRACT

This paper delves into statistical disparities between human-written and ChatGPT-generated texts, utilizing an analysis of Shannon's equitability values, and token frequency. Our findings indicate that Shannon's equitability can potentially be a differentiating factor between texts produced by humans and those generated by ChatGPT. Additionally, we uncover substantial distinctions when studying the most frequent tokens.

KEYWORDS

Shannon Equitability, NLP, ChatGPT, Education

1. INTRODUCTION

Large language models (LLM) have created disturbance in a lot of fields, by making it possible to get knowledge as a response to a user's prompt. ChatGPT, a dialogue system (chatbot) developed on top of GPT-3 and GPT-4 LLMs (Liu et al., 2023) is the most widely known example, with successful usage in mathematics (Frieder et al., 2023), software industry (Surameery & Shakor, 2023), medicine (Biswas, 2023), meteorology (Zhu et al., 2023), education (Bishop, 2023), and many other fields.

This technology can be abused in education, as students can cheat by using ChatGPT to write essays, do assignments, and solve problems. In response to a user's question, ChatGPT can create a text output of up to 4096 characters. But, with just a little hack, we have managed to generate texts with more than 4,000 words. This is more than enough for basic student reports, as Benton's research (Benton, 2017) finds the average of 700 words to be sufficient for the top grade.

In their study, Kostić and Vitić (2021) examined the differences in probability distribution of inflected cases of nouns in the contemporary and medieval Serbian language. The obtained results indicate a conspicuous difference in probability distribution between the two samples. To determine whether there is a parameter that is conserved over time, they calculated the ratio of the obtained and the maximum entropy for the two samples. This metric specifies the distance from the maximum entropy of a system and is known as *Shannon's equitability*. Applying Shannon's equitability metric, a minimal (statistically nonsignificant) difference between the two samples was obtained. This finding suggests that despite large changes in the probabilities of noun cases over time, the distance from the maximum entropy remained conserved. It's worth noting that Levshina (Levshina, 2019) expands the purview by utilizing Shannon's equitability calculation to compare the entropies of entire texts, rather than focusing solely on individual cases.

The primary aim of this paper is to employ Shannon's equitability as a statistical measure to discern between texts generated by ChatGPT and texts composed by humans. Our study sets forth a null hypothesis (H_0) that assumes the population means of Shannon's equitability in both groups are equal. In contrast, the alternative hypothesis (H_1) suggests that the population means of Shannon's equitability differ between the two groups. It

*These authors contributed equally to this work.

is important to emphasize that all texts examined in this study are written in the English language, ensuring consistency in linguistic context.

The rest of the paper is organized as follows. Section 2 presents some related work. Section 3 explains how data is collected, and Section 4 describes the methodology used to calculate Shannon's equitability. In Section 5 we present our findings and draw conclusions in Section 6.

2. RELATED WORK

Considering the broad range of question domains to which ChatGPT responds with seemingly human-like answers, justified concerns arise regarding potential misuse. Numerous research studies have been conducted using various methodologies aiming to determine the authorship of a text - ChatGPT or human? Khalil and Er in their research (Khalil & Er, 2023) state that this powerful and easily accessible technology has led to concerns about plagiarism and cheating in educational institutions. In their study, they examined short essays generated by ChatGPT using two popular plagiarism-detection tools, iThenticate and Turnitin. The results of this study showed that out of 50 essays generated by ChatGPT, 40 exhibited a high level of originality, emphasizing the importance of this topic. Mitrović, Andreoletti and Ayoub have also investigated the detection of text generated by ChatGPT (Mitrović et al., 2023). In their research, they developed a Transformer-based machine learning (ML) approach that classifies text as either generated by ChatGPT or by a human. The results revealed that their ML model can achieve satisfactory performance with an accuracy of approximately 79%. Justin Diamond offered a different perspective on this topic in his research (Diamond, 2023), posing the question: "Do languages generated by ChatGPT statistically look human?". In his research, Diamond utilized Zipf's law, which is commonly used in the field of statistical linguistics. According to Zipf's law, the word frequencies in a text corpus are inversely proportional to their rank in the frequency table (Diamond, 2023). In the context of his experiment, the procedure unfolded as follows: once all texts had been generated, collected, and/or compiled, their Zipfian distributions and correlations were computed and visually represented to facilitate comparison. His findings indicate that both the text generated by ChatGPT and the texts written by humans closely adhere to Zipf's law, and that the text generated by ChatGPT exhibits statistical properties similar to those written by humans. The ultimate purpose of his research is to enhance our understanding of technology and linguistics.

Our research revolves around a fundamental question: Do the word frequencies in ChatGPT and human writings share a common origin? To investigate this, we employ Shannon's equitability as a tool to uncover insights and seek an answer to this question.

3. DATASETS

The corpus utilized for our experiment comprises essays generated by AI as well as essays written by humans. The AI-generated text was acquired from a freely accessible version of ChatGPT, which was made available to all students. The specific task assigned to the AI was to compose a random essay. As the experiment progressed and the results began to circulate, we introduced additional keywords to the essay request, including "Europe," "Africa," "America," "Asia," "love," "war," "betrayal," "famine," and "happy ending." These keywords were incorporated to prompt diverse and varied responses from the AI-generated essays. To accommodate the size limitation of 4096 characters, the AI-generated responses were structured as a series of chapters, with each chapter containing a few sentences that provided an overview of the plot. In order to expand upon these initial responses, we employed an iterative approach by using ChatGPT's own generated responses as input, feeding them back into the system for each individual chapter. Through this process, we were able to generate 12 extended texts, allowing for a more comprehensive exploration of the essay topics and enhancing the overall depth of the AI-generated essays.

We selected twelve random essays written by final-year undergraduate students from the MICUSP corpus (Römer & Wulff, 2008) for our analysis of human texts. These essays were specifically chosen based on certain criteria: they belonged to the English discipline and were argumentative essays with a minimum length of 1900 words.

4. METHODOLOGY

Our approach is grounded in information theory, a field that may appear disconnected from linguistic studies but finds practical applications in understanding language phenomena. We started with the input text and conducted the *tokenization process* (Loper & Bird, 2002) to split the text into individual words. The resulting sequence of individual words formed the foundation for subsequent analysis. In addition, we incorporated punctuation marks, recognizing their importance in human writing as aids for sentence comprehension and clarity (Hill & Murray, 2000). Then, we calculated the occurrence of each token in the defined corpus. By analysing the frequencies, we identified the dominant tokens.

Then we directed our attention to one of the fundamental concepts in information theory - *entropy*. In information theory, entropy (or Shannon's entropy) can be understood as a measure of the orderliness of a system or as the average amount of information emitted by a system.

A way to compare the entropy of two systems, regardless of the number of their elements, is by applying a measure known as *Shannon's equitability*, expressed as the ratio of the obtained entropy (H) and the maximum entropy (H_{max}).

We have developed a straightforward Python function to compute Shannon's equitability, which serves as the ultimate metric. Utilizing this function, we have evaluated Shannon's equitability for various systems and conducted thorough comparisons.

Finally, we have computed the T-test using the collected results of Shannon's equitability.

5. FINDINGS

Table 1 presents the processed data derived from a set of twelve human-written texts. The data includes Shannon's equitability, the token with the highest frequency, and the word with the highest frequency within these texts. One notable commonality among these texts is the consistent appearance of the article "the," which is the most frequently used word in the entire corpus. Furthermore, we observe the significance of commas, which emerge as the most frequent token in almost half of the texts. Overall, the Shannon's equitability values range between 0.8 and 0.84.

Table 1. Data gathered after processing human texts: Shannon's equitability, most frequent token (top token) and most frequent word (top word). Values are rounded to 4 digits

Human	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 10	Text 11	Text 12
Shannon's equitability	0.8331	0.8271	0.8309	0.8261	0.8047	0.8033	0.8276	0.8196	0.8292	0.805	0.8074	0.8362
Top token	the	the	the	,	the	,	,	,	,	the	the	the
Top word	the	the	the	the	the	the	the	the	of	the	the	the

In contrast, the word "and" emerges as the most frequently occurring word in ChatGPT texts. Notably, commas surpass even that frequency and dominate as the most common token.

When considering Shannon's equitability values, it becomes apparent that ChatGPT texts demonstrate a broader range of dispersion compared to human texts. Specifically, the Shannon's equitability values for ChatGPT texts span from 0.78 to 0.83.

ISBN: 978-989-8704-52-8 © 2023

Table 2. Data gathered after processing ChatGPT texts: Shannon's equitability, most frequent token (top token) and most frequent word (top word). Values are rounded to 4 digits

ChatGPT	Text 1	Text 2	Text 3	Text 4	Text 5	Text 6	Text 7	Text 8	Text 9	Text 10	Text 11	Text 12
Shannon's equitability	0.7804	0.781	0.8159	0.8292	0.8157	0.8203	0.8031	0.8162	0.8192	0.7999	0.8014	0.8052
Top token	the	,	,	,	,	and	,	,	,	,	,	,
Top word	the	the	and	and	and	and	and	and	and	and	the	and

In our endeavour to discern disparities between texts composed by humans and those generated by ChatGPT, we identified the top 6 most frequent tokens within both datasets. Subsequently, we calculated the standard deviation of each token's frequency. To illustrate these findings, we employed a Star chart in Figure 1. Each spoke on the chart corresponds to one of the selected tokens, with its length representing the token's standard deviation of occurrence. The standard deviation of token frequency in ChatGPT is generally higher than in human text, which indicates a greater dispersion of tokens within the texts produced by ChatGPT compared to those generated by humans.

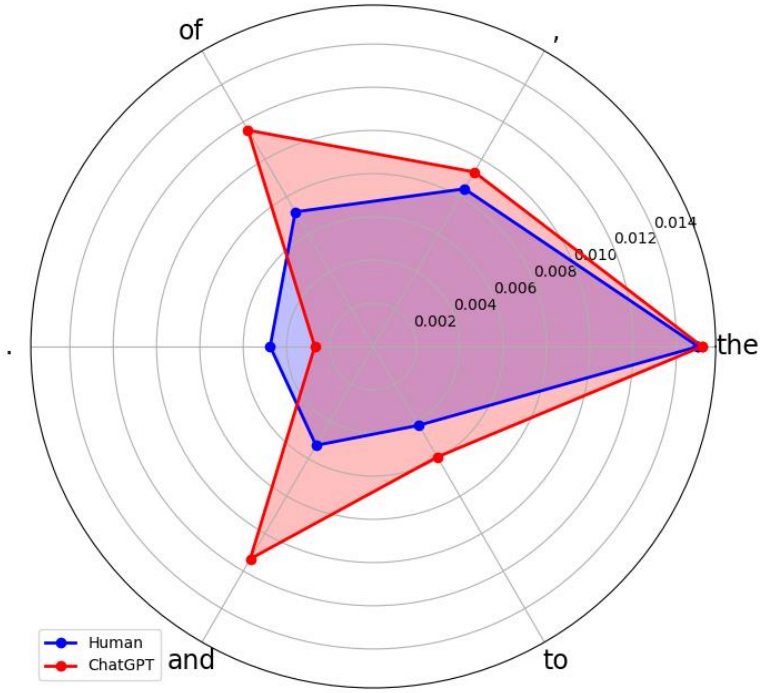

Figure 1. Comparison of standard deviation of frequency for six most frequent tokens in Human and ChatGPT texts

Moreover, a contrast between the mean Shannon's equitability values for each of the two samples (i.e. humans vs Chat GPT) indicated statistically significant difference ($t(22) = 2.40$, i.e. $p = 0.026$). Notably, this p-value falls below the commonly accepted threshold of 0.05 for statistical significance. As a result, we are able to reject the null hypothesis (H_0) that posits no difference in means between texts authored by humans and those generated by ChatGPT. This statistical evidence further supports the notion that there exists a significant distinction in terms of Shannon's equitability between human-generated texts and ChatGPT-generated texts.

Table 3. Mean and variance of Shannon's equitability values for humans and ChatGPT, along with the difference between these numbers. The values are rounded to 4 digits

Shannon equitability	Mean	Variance
Human	0.8208	0.000139
ChatGPT	0.8073	0.000212
Difference	0.0136	

6. FUTURE WORK

In our next research phase, we aim to develop a versatile web application focused on quantitative text analysis. This application will empower users to manipulate and analyse text in various ways, offering both file upload and text generation capabilities. Figure 2 displays the upcoming application's conceptual design.

Key features of the application include:

1. *Upload Text*: Users can upload files or request text generation by ChatGPT.
2. *Tokenization*: Users will have access to standard text tokenization and advanced tokenization, which allows custom rules and word exclusion.
3. *Token Frequency Analysis*: The application also calculates token frequencies. They can analyse token frequencies to identify the most common ones, helping them gain insights into their text.
4. *Statistical Text Analysis*: The application provides tools to calculate text entropy and Shannon's equality, aiding users in gauging the complexity of their text.
5. *Data Visualization*: Users can visualize data distribution within the text through Zipf's distribution graphs and other relevant graphical representations. Various relevant graphical representations will facilitate the understanding of complex language patterns for all users.
6. *Data Export*: Based on the analysis, users can conveniently download the generated data.

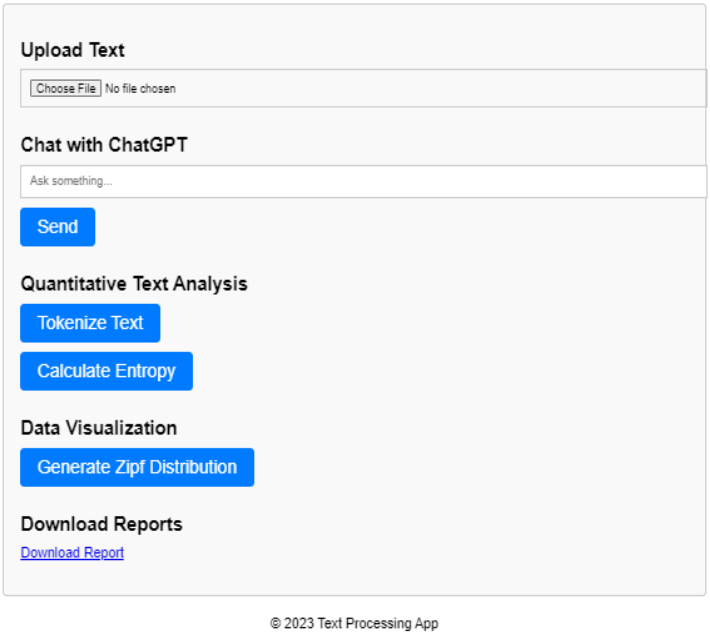

Figure 2. Conceptual design overview of upcoming web application focused on quantitative text analysis

The idea is for this web application to serve as a foundation for future innovations in text analysis. The ultimate goal is for it to become *a comprehensive tool for quantitative text analysis*, initially for Serbian and English languages.

7. CONCLUSIONS

This paper focuses on conducting a statistical analysis to discern the disparities between texts written by humans and those generated by ChatGPT. Our study examines various metrics, including Shannon's equitability values, and the occurrence of frequent tokens and words, extracted from essays authored by students as well as those generated by ChatGPT.

Through our analysis, we have reached a significant conclusion. The examination of Shannon's equitability values allows us to reject the null hypothesis (H_0), which posits that the means of Shannon's equitability are equal between human-written and ChatGPT-generated texts. This finding suggests that Shannon's equitability can effectively serve as a distinguishing factor between texts produced by humans and those generated by ChatGPT.

Furthermore, our investigation of the most frequent tokens reveals notable distinctions. In ChatGPT-generated texts, the word "and" predominates as the most frequent word, while commas emerge as the most frequently occurring token. Conversely, in human-written texts, "the" emerges as the most common word, with a comparable proportion of commas and "the" as the most frequent tokens.

Overall, our analysis sheds light on the statistical differences between human-written and ChatGPT-generated texts, demonstrating the efficacy of Shannon's equitability as a differentiating factor. Additionally, our examination of frequent tokens provides insights into the distinct usage patterns observed in both text categories.

Regarding the limitations of our research, we encountered a small dataset comprising merely 24 texts. These texts were authored by both humans and generated by ChatGPT. Furthermore, the diversity of text topics necessitates that we prompt ChatGPT to generate text aligned with the topics of the human-authored texts. Lastly, a comprehensive analysis of the Zipf's distribution within the two text groups is imperative.

In our future endeavours, alongside rectifying the limitations inherent in this research we plan to carry out additional experiments that involve comparing texts from larger corpora, written in diverse styles including journalism, everyday speech, slang, and tailored for various disciplines such as engineering, biology, philosophy, physics, and more. These experiments will provide us with valuable insights and comparisons across different contexts.

PROGRAM AVAILABILITY

The source code of the program written in Python for this experiment and data results is available from: https://github.com/koprivica/chatGPTvsHumanEntropy

REFERENCES

Benton, T. (2017). How much do I need to write to get top marks? *Research Matters: A Cambridge Assessment publication*, 24, pp. 37–40. Available at: https://www.cambridgeassessment.org.uk/our-research/data-bytes/how-much-do-i-need-to-write-to-get-top-marks/index.aspx (Accessed: 23 June 2023).

Bishop (formerly Lea Shaver), L. (2023). A Computer Wrote this Paper: What ChatGPT Means for Education, Research, and Writing. *Research, and Writing*. Available at: https://doi.org/10.2139/ssrn.4338981.

Biswas, S. (2023). ChatGPT and the Future of Medical Writing, *Radiology*, 307(2), p. e223312. Available at: https://doi.org/10.1148/radiol.223312.

Diamond, J. (2023). ”Genlangs” and Zipf’s Law: Do languages generated by ChatGPT statistically look human? arXiv. Available at: https://doi.org/10.48550/arXiv.2304.12191.

Frieder, S. et al. (2023). Mathematical Capabilities of ChatGPT. arXiv. Available at: https://doi.org/10.48550/arXiv.2301.13867.

Hill, R.L. & Murray, W.S. (2000). Chapter 22 - Commas and Spaces: Effects of Punctuation on Eye Movements and Sentence Parsing, in A. Kennedy et al. (eds) *Reading as a Perceptual Process*. Oxford: North-Holland, pp. 565–589. Available at: https://doi.org/10.1016/B978-008043642-5/50027-9.

Khalil, M. & Er, E. (2023). Will ChatGPT get you caught? Rethinking of Plagiarism Detection. arXiv. Available at: https://doi.org/10.48550/arXiv.2302.04335.

Kostić, A., & Vitić, Z. (2021). Probability distribution of noun cases in writings of St. Sava and the contemporary Serbian language, *Digital humanities and Slavic cultural heritage: international scientific conference, Vol. 1*, Belgrade, 6-7 May 2019, Belgrade: Savez slavističkih društava Srbije, pp. 19-36. Available at: https://doi.org/10.18485/mks_dh_skn.2021.1.ch2.

Levshina, N. (2019). Token-based typology and word order entropy: A study based on Universal Dependencies, *Linguistic Typology*, 23(3), pp. 533–572. Available at: https://doi.org/10.1515/lingty-2019-0025.

Liu, Y. et al. (2023). Summary of ChatGPT/GPT-4 Research and Perspective Towards the Future of Large Language Models. arXiv. Available at: https://doi.org/10.48550/arXiv.2304.01852.

Loper, E. & Bird, S. (2002). Nltk: the natural language toolkit. arXiv. Available at: https://doi.org/10.48550/arXiv.cs/0205028.

Mitrović, S., Andreoletti, D. & Ayoub, O. (2023). ChatGPT or Human? Detect and Explain. Explaining Decisions of Machine Learning Model for Detecting Short ChatGPT-generated Text. arXiv. Available at: https://doi.org/10.48550/arXiv.2301.13852.

Römer, U. & S. Wulff. (2008). Applying corpus methods to written academic texts: explorations of MICUSP, *Journal of Writing Research*, 2 (2), pp. 99–127. Available at: https://doi.org/10.17239/jowr-2010.02.02.2.

Surameery, N.M.S. & Shakor, M.Y. (2023). Use Chat GPT to Solve Programming Bugs, *International Journal of Information Technology & Computer Engineering (IJITC) ISSN : 2455-5290*, 3(01), pp. 17–22. Available at: https://doi.org/10.55529/ijitc.31.17.22.

Zhu, J.-J. et al. (2023). ChatGPT and Environmental Research, Environmental Science & Technology [Preprint]. Available at: https://doi.org/10.1021/acs.est.3c01818.