UDC 004.738.5
DOI: 10.20535/SRIT.2308-8893.2025.1.10

ASSESSING THE IMPACT OF AI-GENERATED PRODUCT NAMES ON E-COMMERCE PERFORMANCE

O. BRATUS

Abstract. This paper studies the impact of Large Language Model (LLM) technology on the e-commerce industry. This work conducts a detailed review of the current implementation level of LLM technologies in the e-commerce industry. Next, it analyzes the approaches to detecting AI-generated text and determines the limitations of their application. The proposed methodology defines the impact of LLM models on the e-commerce industry based on a comparative analysis between indicators of machine-generated texts and e-commerce product metrics. Applying this methodology to real data, one of the most relevant data collected after the release of ChatGPT, the results of statistical analyses show a positive correlation between the studied indicators. It is proved that this dependence is dynamic and changes over time. The obtained implicit indicators measure the influence of LLM technologies on the e-commerce domain. This influence is expected to grow, requiring further research.

Keywords: large language models, AI-detection, e-commerce, product performance.

INTRODUCTION

Since the release of the first version of ChatGPT on November 30, 2022, LLMs have become integral across numerous aspects of human activity. The capabilities of these models to search for information, serve as assistants, and analyze data have made them widely applicable in various sectors, including business and industry [1]. Particularly in e-commerce — a field where Natural Language Processing (NLP) techniques were already well-integrated before the advent of LLMs — these models have found applications at every stage of interaction among customers, sellers, and products. The introduction of LLMs has inevitably transformed e-commerce practices, significantly changing the industry. Given that the presence of LLMs in a business isn't always immediately apparent, the challenge of assessing their impact on e-commerce closely ties into the ability to discern whether textual data was generated by an LLM or not.

Perplexity per token is a key metric for assessing the predictive power of language models, including prominent transformer models like BERT and GPT-4, among other LLMs. This metric has been crucial for comparing different language models on the same dataset and fine-tuning hyperparameters, though it is sensitive to linguistic characteristics and sentence length [2]. Despite its central role in developing language models, perplexity has limitations. Notably, it does not reliably characterize speech recognition performance and may not effectively indicate overfitting and generalization capabilities [3; 4]. This has led to questioning the merit of solely focusing on perplexity optimization.

Furthermore, while perplexity is a common baseline for differentiating between machine-generated and human-generated text, it is often inadequate when used alone, leading to a shift away from methods solely reliant on statistical signatures. Instead of relying solely on raw perplexity scores, a more nuanced approach involves comparing the perplexity measurement with cross-perplexity [5]. This method assesses how unexpected one model's next token predictions are to another, providing a more distinct separation between machine and human text than perplexity alone.

Thus, to investigate the impact of LLM technology on e-commerce, the following research questions are formulated:

RQ1: Do text perplexity-based statistical indicators and e-commerce product metrics correlate?

RQ2: Does the relationship between text perplexity-based statistical indicators and e-commerce product metrics evolves over time?

This research contributes to the understanding of LLMs' influence on e-commerce. The key contributions are as follows:

1. To the best of the author's knowledge, this study is among the first to assess the impact of LLM models on e-commerce, with the introduction of a unique approach and then using it on real-world data.

2. This paper explored the relationship between text perplexity-based statistical indicators and product metrics and found a positive correlation that, as verified by statistical techniques, appears to change over time.

The structure of this paper is organized as follows: Section 2 reviews related work, Section 3 describes the methodology, Section 4 details the experiments and results, and Section 5 concludes the paper and proposes directions for future research.

RELATED WORK

LLM in NLP. Recent advancements in NLP have been significantly shaped by (LLMs like GPT-2, GPT-3, and BERT, which have established new benchmarks in various NLP tasks due to their ability to produce coherent and human-like text [6; 7]. These models have demonstrated their effectiveness beyond benchmarks and have been successfully utilized in real-world applications such as automated customer support, conversational systems, and text summarization [8; 9].

More recently, advanced LLMs, including GPT-4 [10], Gemini [11], and Llama 2 [12], have shown remarkable proficiency in natural language processing tasks [1], information retrieval [13], and various other domains [14; 15].

NLP in e-commerce. NLP techniques have been extensively utilized in e-commerce for various tasks, including sentiment analysis, recommendation systems, and search engine optimization [16; 17]. Previous research has investigated using NLP to extract product attributes, create stylistic variations of product descriptions, and generate multilingual descriptions [18; 19]. Although these methods show promise, they have yet to achieve the scalability needed to produce high-quality, human-like results. While NLP applications in business settings are not a novel concept, there has been limited exploration into their tangible effects on revenue and customer engagement.

LLM in e-commerce. The integration of LLM technology into e-commerce has not only surpassed existing NLP solutions but has also been instrumental in addressing a broader range of challenges. Key applications of LLMs in this domain (Fig. 1) include advanced customer support, content generation (such as product descriptions, blog posts, comments, and reviews), content evaluation (including ratings and sensitivity analysis of user feedback), recommendation systems, and search engines [20].

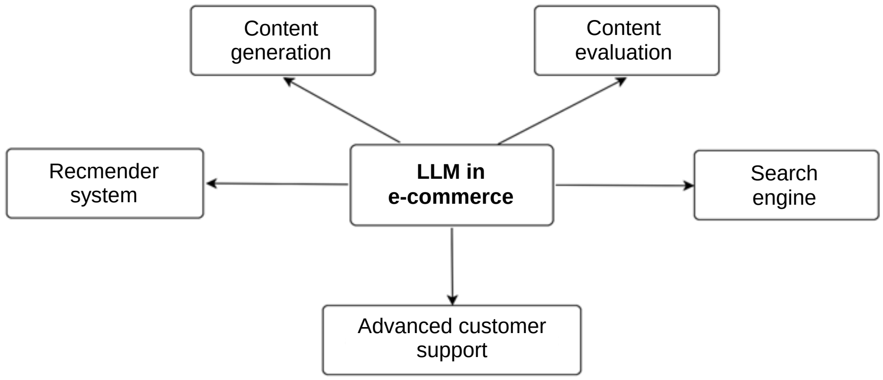

Fig. 1. Applications of LLMs in e-commerce

One notable trend is the fine-tuning of state-of-the-art LLMs for specific e-commerce tasks. For instance, LLMs created for automating product description generation enhance click-through rates and significantly reduce the manual effort required in content creation [21]. Similarly, employing LLMs for analyzing product reviews offers substantial benefits to e-commerce stakeholders — such as owners, managers, marketers, and data analysts — by providing quicker responses to customer feedback, thereby improving the overall effectiveness of e-commerce strategies [22]. In search engine optimization, LLMs are utilized for keyword selection and content enhancement [23].

Additionally, there is a growing trend towards developing families of LLM models tailored specifically for e-commerce applications. These models are not designed to be generalists across multiple domains but are specialized and optimized for e-commerce tasks, training exclusively on relevant data and targeting e-commerce metrics [24; 25]. Given the widespread adoption of LLMs in the e-commerce sector, exploring how this technology impacts the industry is crucial.

AI-generated text detection. Early efforts to detect machine-generated text have shown potential, particularly with models whose outputs are not convincingly human-like. However, the advent of transformer models for language generation [6; 7; 12; 26] has rendered many of these basic detection mechanisms ineffective. One strategy is to record [27] or watermark all generated text [28], but such preemptive measures require complete control over the generative models.

In response to the growing prevalence of machine-generated text, primarily through platforms like ChatGPT, a wave of research has focused on post-hoc detection methods. These approaches do not rely on cooperation from model developers. Detection methods can be broadly categorized into two types. The first involves training detection models, where a pre-trained language model is fine-tuned for the binary classification task of detecting machine-generated text [29–31]. Techniques such as adversarial training [32] and abstention [33] are also employed. Alternatively, instead of fine-tuning the entire model, a linear classifier can be applied to fixed learned features, allowing for the integration of commercial API outputs [34].

The second category includes methods based on statistical signatures characteristic of machine-generated text. These approaches typically require little or no training data and can be easily adapted to new model families [35]. Examples include detectors based on perplexity [33; 36; 37], perplexity curvature [38], log-rank [39], intrinsic dimensionality of generated text [40], and n-gram analysis [41]. While this overview is not exhaustive, recent surveys can reveal further details [42–45].

From a theoretical standpoint, the main limitation of detection is that fully general-purpose language models, by definition, would be impossible to detect [46–48]. However, even models approaching this ideal may still be detectable with a sufficient number of samples [49]. In practice, the relative success of detection methods, including those proposed and analyzed in this work, provides evidence that current language models are still imperfect representations of human writing and, thus, detectable.

RESEARCH METHODOLOGY

The proposed methodology employs a specialized approach that examines the statistical properties of texts, particularly those that indicate the extent to which a text is machine-generated, and compares these with product metrics. The goal is to identify potential relationships between the two characteristics. This methodology is structured into three distinct stages (Fig. 2): 1) calculating the machine-generated characteristics of text features; 2) assessing the e-commerce product metrics; 3) conducting a statistical analysis to determine any significant correlations.

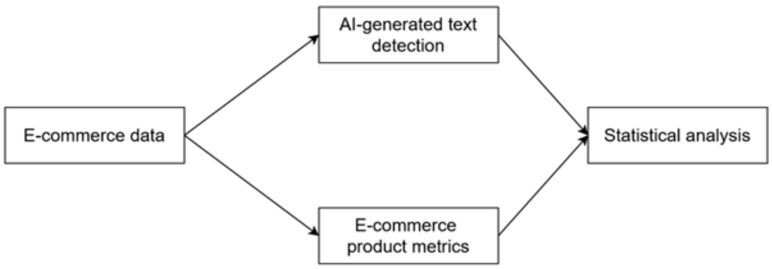

Fig. 2. Proposed methodology

AI-generated text detection. As described in related works, one of the approaches to detecting machine-generated text involves calculating specific statistical indicators of the texts and comparing them to predefined threshold values. This paper follows two critical conditions to choose a model for detecting machine-generated text. First, there is the absence of a training dataset to fine-tune classifiers for machine-generated text recognition. Second, there is no information on whether LLM models were used in generating the texts and, if so, which specific models. Therefore, a detection model that does not rely on training (zero-shot model) and is agnostic to any LLM model is required. The method, called Binoculars, meets these criteria and utilizes the binoculars score, which calculates the ratio of perplexity to cross-perplexity [5]:

$$B_{M_1,M_2}(s) = \frac{\log PPL_{M_1}(s)}{\log X - PPL_{M_1,M_2}(s)},$$

where perplexity, $\log PPL_{M_1}(s)$ is defined as the average negative log-likelihood of all tokens in the given sequence s cross-perplexity, $\log X - PPL_{M_1,M_2}(s)$, is defined as the average per-token cross-entropy between the outputs of two models, M_1 and M_2 when operating on the tokenization of the sequence s.

In other words, the numerator in this method is the perplexity, which quantifies how unexpected a string is to model M_1. Conversely, the denominator assesses how unexpected the token predictions of model M_2 are when evaluated by M_1. Intuitively, this means that a human is expected to diverge from M_1 more than M_2 could, assuming that the LLMs M_1 and M_2 re more similar to each other than they are to a human. This approach achieves state-of-the-art accuracy without requiring any training data. It can detect machine-generated text from various modern LLMs without needing model-specific adjustments. Therefore, this work utilizes the Binoculars score as a statistical signature for identifying machine-generated text.

E-commerce product metrics. Using an e-commerce dataset that captures interactions between customers and products, various metrics can be calculated to provide valuable insights into product performance and customer behaviour. Metrics related to sales and revenue include sales volume, revenue, conversion rate, and profit margin. Another category of metrics focuses on user experience, encompassing indicators such as product return rate, customer reviews, and ratings. The scope of product metrics is not confined to these examples; it is instead determined by the availability of specific features in the dataset that enable the calculation of particular metrics.

Statistical analysis. The third and final stage of the proposed methodology is a statistical comparison of machine-generated text characteristics and product metrics. Spearman's rank correlation coefficient is used to determine any relationship. It is a nonparametric measure of rank correlation that assesses how well the relationship between two features can be described using a monotonic function.

A bootstrap method is used to answer this research's second question and determine whether the relationship between the studied metrics has changed. It estimates the confidence intervals and significance of the difference between two Spearman coefficients. Bootstrapping can provide a flexible and robust way to handle non-parametric statistics without relying on normality assumptions. Bootstrap involves repeatedly resampling the data with replacement. For each bootstrap sample, the Spearman correlations for each of the two datasets are calculated, and the difference between them is computed. Then, the differences from all bootstrap samples (1000 samples in this work) are collected to form a distribution of differences and determine the confidence interval. A 95% confidence interval is used, which means the 2.5th percentile and the 97.5th percentile of the bootstrap differences are found. Suppose the 95% confidence interval does not include zero. In that case, it indicates a statistically significant difference between the two correlation coefficients and, therefore, a statistically significant change in the relationship between machine-generated text characteristics and product metrics. Otherwise, if the interval includes zero, no significant difference exists between the correlations at the chosen confidence level.

EXPERIMENTS

Dataset and Preprocessing. One of the challenges in researching the effects of LLM technology on e-commerce is the scarcity of accessible, complete, and up-to-date datasets. Given that ChatGPT was only released in November 2022, and considering the gradual integration of LLMs within the e-commerce sector, it will take some time to accumulate and publish comprehensive datasets.

The MerRec [50], introduced in March 2024, is one of the first datasets that meets these requirements. It encapsulates detailed records of user interactions on the Mercari e-commerce platform, tracking millions of users and products over six months, from May to October 2023. MerRec not only captures basic features such as user attributes (user_id, sequence_id, session_id) and product attributes (item_id, product_id) but also includes specialized data like timestamped action types, product taxonomy, and textual product descriptions, making it a rich resource for analysis.

This analysis focused on products listed during the dataset's initial (May) and final (October) months. Given limited computing resources and to minimize data skew from outliers or abnormal product behaviour, the data is preprocessed with specific criteria: only those products are selected whose names contain at least five words and are purchased at least once.

Generalized word shift graphs were utilized to enhance the clarity of product names analyzed in this research. Such visualizations provide a meaningful and interpretable summary of how individual words contribute to variations observed between two distinct text corpora [51].

For instance, the product names in the Women category for October 2023 were analyzed, featuring low ("AI-generated") and high ("human-generated") binoculars scores. Examples of top 20 product names with the lowest and the highest binoculars scores are presented in Table. Names scoring low on the binoculars scale exhibited higher standardization, including consistent word order, capitalization of each word, and numerical size descriptors. In contrast, names with high binoculars scores (likely human-generated) displayed a less structured word order, lacked punctuation and used words to describe sizes (e.g., small).

Top 20 product names in the Women category for October 2023 with the lowest and the highest binoculars scores

Top 20 product names with the lowest binoculars score	**Top 20 product names with the highest binoculars score**
Converse size 7.5 women's shoes womens ugg boots size 9	Sebek Zigvolt Acrylic Stud Earrings
Keychain Wallet, Wristlet, Bangle, Bracelet, ID Card Holder, Purse, Key Chain, G	FIGS rose joggers size Small Petite
Christian Louboutin Women Black Heels Shoes Size 8.5 (38.5)	J crew midi floral sun Dress
Vtg Sterling Silver 925 Hinged Bangle Bracelet	Motel Olivia faux leather biker jacket white
Polo Ralph Lauren Women's V-Neck T-Shirt - Size Medium - Navy Blue	She Darc sweatshirt! Size small
UGG Brookfield Brown Sheepskin Leather Boots Size 8.5	Grae Cove linen short sleeve waist tie pockets mini dress blush women's XL
Avatar: The Last Airbender Aang & Katara Mini Backpack	Hot topic rob zombie hoodie XS
Womens Old Navy Fleece Jacket Size Small	Famous magic land couples OS leggings
Nike air max 270 women size 7	August Silk womens colorful funky patterned Shorts
Old Navy Active Fleece Jacket	Kate spade Pitch Purrfect Piano Crossbody KC729 NWT
Lululemon Long Sleeve - Size 10	Beautiful Disaster Tribe Jacket Size L
Purple Hooded Long Sleeve Sweater	Express Low rise columnist pants
Tory Burch Black Leather Boots Size 10.5	New sweatshirt hoodie Jeffrey Star
Victoria's Secret PINK Bling Leggings	Hades Disneyland Spirit Jersey Small
The Nightmare Before Christmas Jack Skellington	Save for Rosemary Special love lot
Nike Air Max 2X (Women)	Hot Topic Mushroom Collar dress
Super cute and comfy pajamas	Coach Wyn Logo Plaque Small Wallet
Tommy Hilfiger Women's Medium Red and White Striped Dress	NEW bundle Victoria Secret underwear
Costume Jewelry Lot - 25 pieces - Necklaces, Bracelets, Earrings	Cat In a pumpkin earrings
	Waffle Debut Retro Sneaker leopard

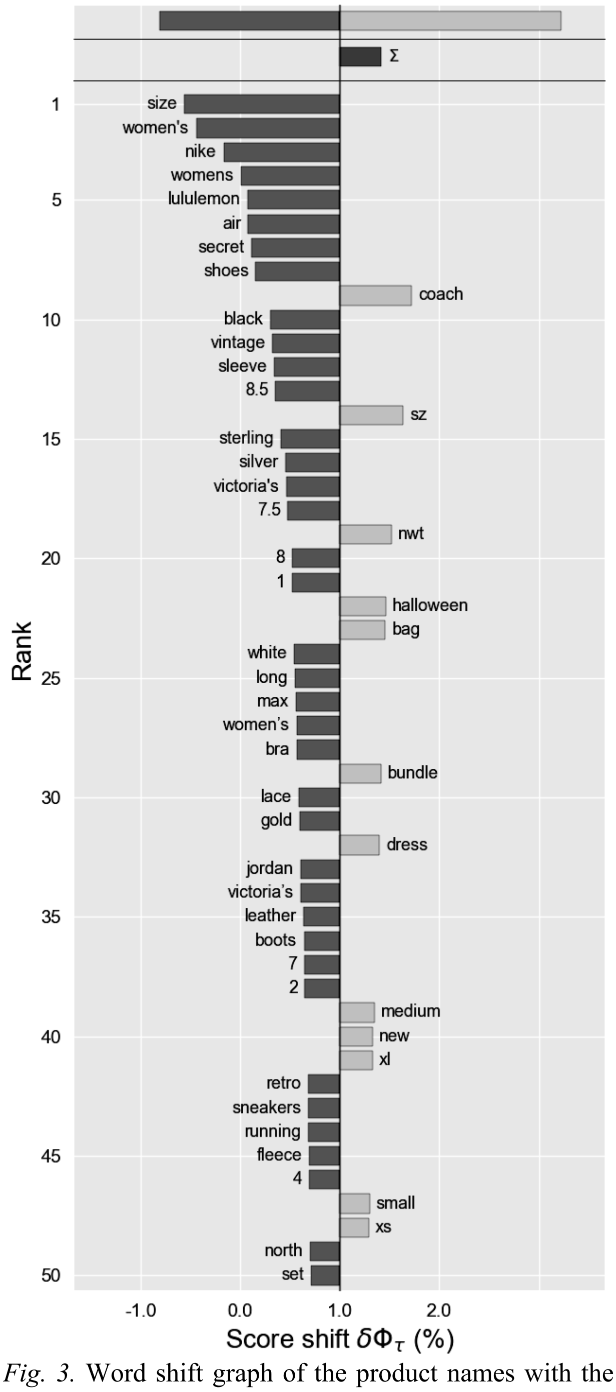

Fig. 3. Word shift graph of the product names with the lowest and highest binoculars scores

The examination using word shift graph (Fig. 3) revealed minimal significant differences in word usage between the two groups (the first group contains product names with the binoculars scores from first quartile (Q1) and the second group contains product names with the binoculars scores from fourth quartile (Q4). However, several subtle distinctions were noted. Primarily, descriptive words for sizes (e.g., small, medium, xs, xl) were used in “human-generated” names. In contrast, numerical representations (e.g., 7.5, 8, 8.5) were employed in “AI-generated” names, enhancing the accuracy and standardization of size descriptions. Additionally, abbreviations (e.g., sz, nwt) were often included in “human-generated” names. Thus, the example of the Women’s product category demonstrates how product names with different binocular scores differ from each other.

LLM models. As described in section 3, the binoculars method is used as an AI-generated text detector, which requires 2 LLM models. Moreover, these models should provide access to the raw logits of all tokens in the given sequence to calculate the binoculars score. Unfortunately, most state-of-the-art LLM models (GPT-4, Claude-3, etc.) do not provide access to such logits. Therefore, the open-source LLM models are considered, and the Falcon-7b model and the Falcon-7b-instruct model are chosen, which are pretrained generative text models with 7 billion parameters and demonstrate high performance.

It was carried out on the remote resources of Google Colab and consumed approximately 10 hours of A100 GPU to generate the scores for nearly 300000 unique product names.

Evaluation metric. To investigate the impact of LLM on e-commerce (namely, on product names), and based on the features of the selected MerRec dataset (unique user actions), the conversation rate is used as one of the central business metrics in e-commerce that indicates product performance. It is defined as the ratio of the total number of customers who purchased the product compared to the total number of customers who interacted with it.

Results. The proposed methodology's performance is evaluated on the real-data MerRec dataset. Overall, a positive correlation between binoculars score and conversation rate is found, which differs depending on the product category. These results are inspected in more detail in the following.

RQ1 Do text perplexity-based statistical indicators and e-commerce product metrics correlate?

First, the conversation rate scores are calculated for all products sold in May 2023; then, for the same products, the binoculars scores of their names are calculated. After that, the Spearman correlation coefficient between these indicators is calculated, and it found that it differs significantly for products of different categories (Fig. 4). For example, for the Men and Kids categories, the correlation is the highest at 0.54 and 0.53, respectively, which indicates a moderate correlation degree. The correlation is somewhat lower, but also significant, for products in the Women category (0.28). There is a group of categories for which the correlation is positive but very weak (Sports & outdoors, Pet Supplies, Toys & Collectibles, Vintage & collectibles). There are also categories for which the correlation is practically absent, but what is important to note is that there are no products for which the Spearman correlation is negative (except Garden & outdoor).

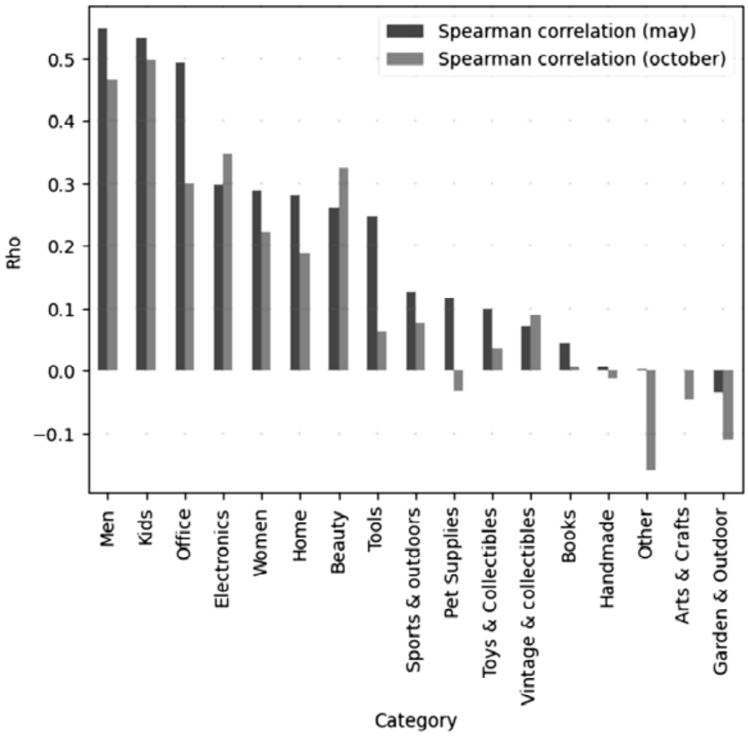

Fig. 4. Spearman correlation coefficients between binoculars score and conversation rate of products from the MerRec dataset

It is to noticed that the correlation is the largest for the most general groups (Men, Women, Kids, Office, Electronics), which are characterized by a wide range of products and their diversity, while products that can be attributed to a specific field of activity (Sports & outdoors, Pet Supplies, Vintage & collectibles, Handmade, etc.) have very weak or zero correlation. It can be assumed that for general categories, it is not easy to come up with an original product name that will distinguish it from others and interest customers; at the same time, for specific domain categories, the names of products may contain certain specifications, which will interpret them as original, which, however, is typical for them, and in no way distinguishes them from other products.

A similar analysis was conducted for products sold in October 2023. Similarly, a positive correlation between binoculars score and conversation rate is observed. However, for most categories, the correlation decreased; for a few, such as Other and Garden & outdoor, the correlation became negative, albeit very weak.

Thus, a moderate positive correlation between binoculars score (a text perplexity-based statistical indicator) and conversation rate (an e-commerce product metric) is seen. It can be interpreted that a higher probability of the product name being generated by a human (higher binoculars score) correlates with better product performance.

RQ2 Does the relationship between text perplexity-based statistical indicators and e-commerce product metrics evolves over time?

A statistical comparison of the correlation coefficients of the data for May and October is performed. It is found that for most categories, there is a statistically significant change in correlation (Fig. 5). Thus, finding a boxplot with

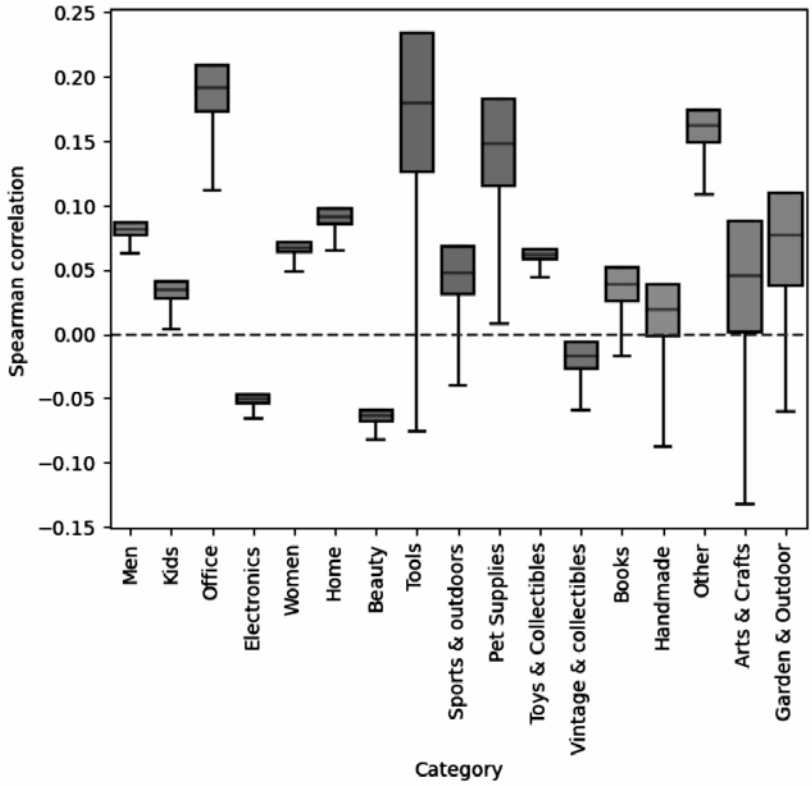

Fig. 5. Distribution of Spearman correlation coefficients across different product categories. Boxplots show median (red line) and 25- and 75-percentiles with whiskers ranging from 2.5- to 97.5-percentile

whiskers entirely above zero indicates a significant decrease in the correlation between binoculars score and conversation rate and placing it below zero, on the contrary, indicates an increase in correlation. It can be concluded that out of all 15 categories, for seven categories (including those with the highest correlation coefficients in May), the correlations decreased statistically; only for three categories increased, and for the rest of the categories, they remained unchanged (or their change is statistically insignificant).

So, for six months, from May to October, for most products, there is a trend to decrease the correlation coefficient between binoculars score (text perplexity-based statistical indicator) and conversation rate (e-commerce product metric). This may be due to the increased use of LLM technology to generate product names, but it is still small.

CONCLUSIONS

In this work, the methodology to determine the impact of AI-generated product names on e-commerce performance is proposed; namely, the relationship between the binoculars score of product names and the conversation rate of products is investigated. It examines in detail the current level of implementation of LLM technology in the field of e-commerce, considering a wide range of problems solved by language models. In addition, the existing state-of-the-art detection methods of machine-generated texts are described, and one of those methods that performs zero-shot and model-agnostic detection is used. Proposed approach is applied to real data for 2023 and a positive correlation between binoculars score (text perplexity-based statistical indicator) and conversation rate (e-commerce product metric) is found. This positive correlation tends to decrease, which is verified statistically. Thus, the impact of LLM technology on e-commerce is observed, and only an increase in this impact is expected in the future.

For future work, a semantic analysis of the comparison of product names over time on changing typical words in the product names triggered by the activity of LLM models can be conducted, which may be fascinating, but this is a question for further research.

REFERENCES

1. W.X. Zhao et al., “A survey of large language models,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2303.18223
2. A. Miaschi, D. Brunato, F. Dell’Orletta, and G. Venturi, “What makes my model perplexed? A linguistic investigation on neural language models perplexity,” in *Proceedings of Deep Learning Inside Out (DeeLIO): The 2nd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pp. 40–47, 2021. doi: 10.18653/v1/2021.deelio-1.5
3. D. Klakow, J. Peters, “Testing the correlation of word error rate and perplexity,” *Speech Communication*, vol. 38, no. 1–2, pp. 19–28, 2002. doi: 10.1016/S0167-6393(01)00041-3
4. S.F. Chen, D. Beeferman, and R. Rosenfeld, *Evaluation metrics for language models*. 1998. Available: https://kilthub.cmu.edu/articles/Evaluation_Metrics_For_Language_Models/6605324/files/12095765.pdf
5. A. Hans et al., “Spotting LLMs with binoculars: Zero-shot detection of machine-generated text,” *ArXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2401.12070

6. A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language models are unsupervised multitask learners,” *OpenAI blog*, vol. 1, no. 8, p. 9, 2019. Available: https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf
7. T.B. Brown, “Language models are few-shot learners,” *ArXiv*, 2020. doi: https://doi.org/10.48550/arXiv.2005.14165
8. D. Adiwardana et al., “Towards a human-like open-domain chatbot,” *ArXiv*, 2020. doi: https://doi.org/10.48550/arXiv.2001.09977
9. P. Lewis et al., “Retrieval-augmented generation for knowledge-intensive NLP tasks,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 9459–9474, 2020. Available: https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html
10. B. Miranda, A. Lee, S. Sundar, A. Casasola, and S. Koyejo, “Beyond Scale: The Diversity Coefficient as a Data Quality Metric for Variability in Natural Language Data,” *ArXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2306.13840
11. Gemini Team Google, “Gemini: a family of highly capable multimodal models,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2312.11805
12. H. Touvron et al., “Llama 2: Open foundation and fine-tuned chat models,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2307.09288
13. S.E. Spatharioti, D.M. Rothschild, D.G. Goldstein, and J.M. Hofman, “Comparing traditional and LLM-based search for consumer choice: A randomized experiment,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2307.03744
14. S. Frieder, J. Berner, P. Petersen, and T. Lukasiewicz, “Large language models for mathematicians,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2312.04556
15. F. Zeng, W. Gan, Y. Wang, N. Liu, and P.S. Yu, “Large language models for robotics: A survey,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2311.07226
16. G. Sousa, “Natural Language Processing and its applications in e-business,” *Cadernos de Investigação do Mestrado em Negócio Eletrónico*, vol. 2, 2022. doi: https://doi.org/10.56002/ceos.0070_cimne_1_2
17. Y. Huang, “Research on the Application of Natural Language Processing Technology in E-commerce,” in *ISCTT 2021; 6th International Conference on Information Science, Computer Technology and Transportation, 2021*, pp. 1–5. Available: https://ieeexplore.ieee.org/abstract/document/9738909
18. M. Chen, Q. Tang, S. Wiseman, and K. Gimpel, “Controllable paraphrase generation with a syntactic exemplar,” *ArXiv*, 2019. doi: https://doi.org/10.48550/arXiv.1906.00565
19. Q. Chen, J. Lin, Y. Zhang, H. Yang, J. Zhou, and J. Tang, “Towards knowledge-based personalized product description generation in e-commerce,” in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019*, pp. 3040–3050. doi: 10.1145/3292500.333072
20. Q. Ren et al., “A survey on fairness of large language models in e-commerce: progress, application, and challenge,” *ArXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2405.13025
21. J. Zhou, B. Liu, J.N.A.Y. Hong, K.-C. Lee, and M. Wen, “Leveraging Large Language Models for Enhanced Product Descriptions in eCommerce,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2310.18357
22. K.I. Roumeliotis, N.D. Tselikas, and D.K. Nasiopoulos, “LLMs in e-commerce: a comparative analysis of GPT and LLaMA models in product review evaluation,” *Natural Language Processing Journal*, vol. 6, p. 100056, 2024. doi: 10.1016/j.nlp.2024.100056
23. G. Chodak, K. Błażyczek, “Large Language Models for Search Engine Optimization in E-commerce,” in *International Advanced Computing Conference*, pp. 333–344, 2023. doi: 10.1007/978-3-031-56700-1_27
24. B. Peng, X. Ling, Z. Chen, H. Sun, and X. Ning, “eCeLLM: Generalizing Large Language Models for E-commerce from Large-scale, High-quality Instruction Data,” *ArXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2402.08831

25. C. Herold, M. Kozielski, L. Ekimov, P. Petrushkov, P.-Y. Vandenbussche, and S. Khadivi, "LiLiuM: eBay's Large Language Models for e-commerce," *ArXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2406.12023
26. A. Chowdhery et al., "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023. Available: http://jmlr.org/papers/v24/22-1144.html
27. K. Krishna, Y. Song, M. Karpinska, J. Wieting, and M. Iyyer, "Paraphrasing evades detectors of AI-generated text, but retrieval is an effective defense," *Advances in Neural Information Processing Systems*, vol. 36, 2024. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/575c450013d0e99e4b0ecf82bd1afaa4- Abstract-Conference.html
28. J. Kirchenbauer, J. Geiping, Y. Wen, J. Katz, I. Miers, and T. Goldstein, "A watermark for large language models," in *International Conference on Machine Learning*, 2023, pp. 17061–17084. Available: https://proceedings.mlr.press/v202/ kirchenbauer23a.html
29. I. Solaiman et al., "Release strategies and the social impacts of language models," *ArXiv*, 2019. doi: https://doi.org/10.48550/arXiv.1908.09203
30. R. Zellers et al., "Defending against neural fake news," *Advances in Neural Information Processing Systems*, vol. 32, 2019. Available: https://proceedings.neurips.cc/paper/2019/hash/3e9f0fc9b2f89e043bc6233994dfcf76-Abstract.html
31. X. Yu et al., "GPT paternity test: GPT generated text detection with GPT genetic inheritance," *CoRR*, 2023. Available: https://arxiv.org/pdf/2305.12519v2
32. X. Hu, P.-Y. Chen, and T.-Y. Ho, "Radar: Robust AI-text detection via adversarial learning," *Advances in Neural Information Processing Systems*, vol. 36, pp. 15077–15095, 2023. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/30e15e5941 ae0cdab7ef58cc8d59a4ca-Abstract-Conference.html
33. Y. Tian et al., "Multiscale positive-unlabeled detection of AI-generated texts," *ArXiv*, 2023. Available: https://arxiv.org/pdf/2305.18149
34. V. Verma, E. Fleisig, N. Tomlin, and D. Klein, "Ghostbuster: Detecting text ghostwritten by large language models," *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2305.15047
35. J. Pu, Z. Huang, Y. Xi, G. Chen, W. Chen, and R. Zhang, "Unraveling the mystery of artifacts in machine generated text," in *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6889–6898, 2022. Available: https://aclanthology.org/2022.lrec-1.744
36. C. Vasilatos, M. Alam, T. Rahwan, Y. Zaki, and M. Maniatakos, "HowkGPT: Investigating the detection of ChatGPT-generated university student homework through context-aware perplexity analysis," *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2305.18226
37. Y. Wang et al., "M4: Multi-generator, multi-domain, and multi-lingual black-box machine-generated text detection," *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2305.14902
38. E. Mitchell, Y. Lee, A. Khazatsky, C.D. Manning, and C. Finn, "DetectGPT: Zero-shot machine-generated text detection using probability curvature," in *International Conference on Machine Learning*, pp. 24950–24962, 2023. Available: https://proceedings.mlr.press/v202/mitchell23a.html
39. J. Su, T.Y. Zhuo, D. Wang, and P. Nakov, "DetectLLM: Leveraging log rank information for zero-shot detection of machine-generated text," *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2306.05540
40. E. Tulchinskii et al., "Intrinsic dimension estimation for robust detection of AI-generated texts," *Advances in Neural Information Processing Systems*, vol. 36, 2024. Available: https://proceedings.neurips.cc/paper_files/paper/2023/hash/7baa48bc166aa2013d78cbdc15010530-Abstract-Conference.html
41. X. Yang, W. Cheng, Y. Wu, L. Petzold, W.Y. Wang, and H. Chen, "DNA-GPT: Divergent N-Gram Analysis for Training-Free Detection of GPT-Generated Text," *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2305.17359

42. S.S. Ghosal, S. Chakraborty, J. Geiping, F. Huang, D. Manocha, and A.S. Bedi, “Towards possibilities & impossibilities of AI-generated text detection: a survey,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2310.15264
43. R. Tang, Y.-N. Chuang, and X. Hu, “The science of detecting LLM-generated text,” *Communications of the ACM*, vol. 67, issue 4, pp. 50–59, 2024. doi: 10.1145/3624725
44. M. Dhaini, W. Poelman, and E. Erdogan, “Detecting ChatGPT: A survey of the state of detecting ChatGPT-generated text,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2309.07689
45. B. Guo et al., “How close is ChatGPT to human experts? Comparison corpus, evaluation, and detection,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2301.07597
46. R. Varshney, N.S. Keskar, and R. Socher, “Limits of detecting text generated by large-scale language models,” in *2020 Information Theory and Applications Workshop (ITA), 2020*, pp. 1–5. doi: 10.1109/ITA50056.2020.9245012
47. H. Helm, C.E. Priebe, and W. Yang, “A Statistical Turing Test for Generative Models,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2309.08913
48. V.S. Sadasivan, A. Kumar, S. Balasubramanian, W. Wang, and S. Feizi, “Can AI-generated text be reliably detected?,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2303.11156
49. S. Chakraborty, A.S. Bedi, S. Zhu, B. An, D. Manocha, and F. Huang, “On the possibilities of AI-generated text detection,” *ArXiv*, 2023. doi: https://doi.org/10.48550/arXiv.2304.04736
50. L. Li, Z. A. Din, Z. Tan, S. London, T. Chen, and A. Daptardar, “MerRec: A Large-scale Multipurpose Mercari Dataset for Consumer-to-Consumer Recommendation Systems,” *ArXiv*, 2024. doi: https://doi.org/10.48550/arXiv.2402.14230
51. R. J. Gallagher et al., “Generalized word shift graphs: a method for visualizing and explaining pairwise comparisons between texts,” *EPJ Data Science*, vol. 10, no. 1, Jan. 2021. doi: 10.1140/epjds/s13688-021-00260-3

Received 02.09.2024

INFORMATION ON THE ARTICLE

Oleksandr S. Bratus, ORCID: 0009-0003-5004-1652, Educational and Research Institute for Applied System Analysis of the National Technical University of Ukraine “Igor Sikorsky Kyiv Polytechnic Institute”, Ukraine, e-mail: olexandr.bratus@gmail.com

ОЦІНЮВАННЯ ВПЛИВУ НАЗВ ПРОДУКТІВ, СТВОРЕНИХ ШТУЧНИМ ІНТЕЛЕКТОМ, НА ЕФЕКТИВНІСТЬ ЕЛЕКТРОННОЇ КОМЕРЦІЇ / О.С. Братусь

Анотація. Досліджено вплив великих мовних моделей (LLM) на електронну комерцію. Здійснено детальний огляд поточного рівня впровадження LLM у електронній комерції. Проаналізовано існуючі підходи до детекції текстів, згенерованих штучним інтелектом (ШІ), та визначено обмеження їх застосування. Запропоновано методологію визначення впливу LLM на електронну комерцію на основі порівняння індикаторів ШІ-згенерованих текстів та продуктових метрик. Продемонстровано застосування методології на реальних даних, що зібрані після релізу ChatGPT, і отримано результати статистичного аналізу, які показують додатну кореляцію між досліджуваними показниками. Підтверджено наявність динаміки цієї залежності та її зміни з часом. Отримані неявні індикатори вимірюють вплив LLM технології на сферу електронної комерції. Очікуємо, що вплив зростатиме, потребуючи подальших досліджень.

Ключові слова: великі мовні моделі, ШІ-детекція, електронна комерція, ефективність продукту.