تأثير تحليلات البيانات الكبيرة والنماذج اللغوية الكبيرة على الملكية الفكرية

محمد محمود عبدالله[1]، خالد محمد جابر[2]، موسى ابراهيم صلاح[3]، علاء أحمد اعمية[4]

DOI:10.15849/ZJJLS.240330.050

[1]قسم علم الحاسوب، كلية العلوم وتكنولوجيا المعلومات، جامعة الزيتونة الاردنية، الاردن
[2]قسم علم الحاسوب، كلية العلوم وتكنولوجيا المعلومات، جامعة الزيتونة الاردنية ، الاردن
[3]قسم نظم الوسائط المتعددة كلية العمارة والتصميم، جامعة الزيتونة الاردنية الاردن
[4]قسم القانون، كلية الحقوق، جامعة فلسطين الاهلية ، فلسطين
* للمراسلة m.abdallah@zuj.edu.jo

الملخص

ظهور تحليلات البيانات الكبيرة والنماذج اللغوية الكبيرة، كما يتجلى فيChatGPT ، قام بإعادة تعريف الصناعات من خلال كشف رؤى وإمكانيات إبداعية غير مسبوقة. يقوم هذا البحث بفحص تأثيرها على المشهد الخاص بالملكية الفكرية (IP). من خلال التفصيل في قضايا القابلية للبراءة، يتنقل البحث عبر التحديات الناشئة من الابتكارات المدعومة بالبيانات والاعتبارات الحقوقية المعقدة المتشابكة مع تحليل المحتوى. وعلاوة على ذلك، يفحص التفاعل بين ChatGPT وIP، مستكشفًا مخاوف حقوق الطبع والحفاظ على أسرار الصناعة لإخراج الذكاء الاصطناعي الناتج. تتضمن الأبعاد الأخلاقية والقانونية الخصوصية في البيانات، وتخفيف التحيز، والشفافية، والمساءلة. يتم توضيح التوصيات، التي تتعامل مع تقييم القابلية للبراءة، وتأمين حقوق الطبع، وممارسات الذكاء الاصطناعي الأخلاقية، والتكيف مع التطورات السياسية. في الختام، يبرز هذا الدراسة الإمكانيات التآزرية لتحليلات البيانات الكبيرة و ChatGPT مع التأكيد على ضرورة تحقيق توازن متناغم بين الابتكار وحماية الملكية الفكرية والاعتبارات الأخلاقية في المشهد التكنولوجي المعقد بشكل متزايد

الكلمات المفتاحية: تقنيات الذكاء الاصطناعي، التحديات الأخلاقية والقانونية، الرعاية الصحية، التعليم، النقل، الصناعة، التجارة الإلكترونية.

Impact of Big Data Analytics and Large Language Model on Intellectual Property

Mohammad Mahmoud Abdallah[1], Khalid Mohammad Jaber[2], Mousa Ibrahim Salah[3], Alaa Ahmad Omayah[4]

[1]Department of Computer Science collage of Science and IT, Al-Zaytoonah University of Jordan , Jordan.
[2]Department of Computer Science collage of Science and IT, Al-Zaytoonah University of Jordan , Jordan.
[3]Department of Multimedia collage of Architecture and Design Al-Zaytoonah University of Jordan , Jordan.
[4]Department of Law, collage of Law Palestine Ahliya University , Palestine
*** Crossponding author:** m.abdallah@zuj.edu.jo

Abstract

The synergistic emergence of big data analytics and Large Language Models, exemplified by ChatGPT, has redefined industries by unlocking unprecedented insights and creative capabilities. This paper investigates their impact on intellectual property (IP) landscapes. Delving into patentability issues, it navigates challenges arising from data-driven innovations and the intricate copyright considerations entwined with content analysis. Moreover, it scrutinizes the interplay between ChatGPT and IP, exploring copyright concerns and trade secret preservation for AI-generated outputs. Ethical and legal dimensions encompass data privacy, bias mitigation, transparency, and accountability. Recommendations are delineated, addressing patentability assessment, copyright safeguards, ethical AI practices, and evolving policy adaptations. In conclusion, this study emphasizes the symbiotic potential of big data analytics and ChatGPT while advocating for a harmonious balance between innovation, IP protection, and ethical considerations in an increasingly complex technological landscape.

Keywords: Dig Data Analytics, Intellectual Property, AI-generated content, Ethical Considerations, Large Language Model**.**

1 Introduction

Big data analytics and Large Language Model fusion have ushered in a transformative era of rapid technological progress. This paper explores the intricate relationship between these twin forces and their far-reaching impact on intellectual property (IP). As industries harness the power of big data to extract insights of unparalleled depth and scope, the subsequent emergence of ChatGPT and similar Large Language Models has provided them with creative capabilities that transcend traditional boundaries.

The convergence of big data analytics and Large Language Model holds immense potential, yet it also raises complex challenges at the intersection of innovation, ethics, and intellectual property rights (Moreno & Redondo, 2016). The overarching objective of this paper is to dissect the multifaceted ways in which these technological advancements influence established IP frameworks. This exploration begins by delving into the nexus between big data analytics and intellectual property, dissecting the patentability hurdles that arise from data-driven innovations, and unraveling the intricate labyrinth of copyright considerations entangled with data analysis.

Furthermore, the paper closely examines AI-generated content, focusing on Large Language Models such as ChatGPT. It investigates the implications of copyright on the outputs generated by these models, probing into the debate on ownership and originality in a domain characterized by human-AI collaboration. In parallel, the study navigates the uncharted territory of safeguarding proprietary AI components as trade secrets, underscoring the importance of striking a balance between open research and proprietary interests (Andanda, 2019).

Ethical and legal considerations punctuate the journey through this intricate landscape. As data privacy concerns, bias mitigation, transparency, and accountability take center stage, the paper identifies the pivotal role that responsible practices play in the convergence of these technologies (Lundqvist, 2018).

This research sheds light on the profound interplay between big data analytics, Large Language Model, and intellectual property. By traversing the realms of innovation, ethics, and law, it aspires to provide a holistic understanding of the implications, challenges, and opportunities arising from this dynamic interaction.

2 Big Data Analytics and Intellectual Property

The concept of big data analytics has revolutionized modern industries by leveraging large volumes of data to uncover patterns, correlations, and previously inaccessible insights. Organizations extract valuable information from diverse data sources through sophisticated algorithms and powerful computing resources, enabling informed decision-making, predictive analysis, and process optimization. This data-centric approach transcends traditional methods, offering a paradigm shift in various sectors, including finance, healthcare, marketing, and manufacturing (Staegemann, Volk, Abdallah, & Turowski, 2023; Taha, 2021).

However, this analytical prowess introduces intellectual property (IP) rights challenges. In the realm of data-driven innovations, patentability hurdles emerge due to the dynamic nature of data analysis.

2.1 Patentability Challenges in Data-Driven Innovations

Data-driven innovations often need help meeting the novelty and non-obvious criteria required for patent protection. While conventional inventions might be tangible products or specific processes, data analysis outcomes can be more abstract. Determining the novelty of insights derived from analyzing pre-existing data sets presents a unique challenge, as the data itself may be familiar. Still, the interpretations drawn from it might be innovative (Hamza & Pradana, 2022).

Moreover, the non-obviousness criterion is complex in data analytics. Identifying a technical advance that would have been obscure to someone skilled in the field might be challenging when dealing with data manipulation to extract valuable insights (Mattioli, 2014).

Distinguishing between technical innovation and data analysis outcomes further complicates the patentability assessment. The line between a genuinely inventive algorithm and a process involving novel data manipulation is often blurred. This distinction is critical for determining the eligibility of data analysis-based inventions for patent protection (Alzyadat et al., 2021; Hamza & Pradana, 2022).

2.2 Copyright Considerations in Big Data Analytics

In big data analytics, copyrighted content like text, images, and music is often included in the datasets under analysis. This integration raises questions about the intersection of copyright law and data analysis practices.

The principles of "fair use" and "transformative use" come into play when copyrighted content is used in data processing. Fair use allows for limited use of copyrighted material without seeking permission from the copyright holder, provided the usage is transformative and non-commercial. It does not negatively impact the original work's market value (Geiger, Frosio, & Bulayenko, 2018).

However, determining the transformative nature of data analysis is complex. While some studies may be genuinely transformative by revealing insights not present in the original content, others might merely reproduce or repurpose the copyrighted content, potentially leading to copyright infringement claims (Abbott, 2017; Hawashin, Althunibat, Kanan, AlZu'bi, & Sharrab, 2023).

The challenges associated with using copyrighted content in data analysis are further intensified by the sheer volume of data involved and the potential difficulty in obtaining accurate copyright information for each piece of content. This intricacy requires organizations to navigate copyright considerations meticulously to avoid legal complications (Kemp, 2014).

The marriage of big data analytics and intellectual property engenders a nuanced landscape. While data analytics revolutionizes industries, patentability, and copyright challenges necessitate a careful balance between innovation and IP protection (Kemp, 2014). This intersection underscores the importance of legal frameworks that can adequately accommodate the unique characteristics of data-driven innovations while respecting established IP rights.

3 Large Language Models and Intellectual Property

Large Language Models (LLM) like ChatGPT have emerged as transformative tools capable of generating coherent and contextually relevant text. These models, powered by deep learning techniques and extensive training on diverse datasets, exhibit capabilities that extend beyond mere information retrieval to creative content generation. ChatGPT, in particular, has garnered attention due to its ability to simulate human-like interactions, rendering it a valuable asset across industries.

The introduction of AI-generated content, facilitated by models like ChatGPT, introduces intricate challenges regarding copyright ownership. The conventional understanding of copyright, where human creators hold exclusive rights to their creations, becomes ambiguous in the context of AI-generated content (Abdikhakimov, 2023). Determining the rightful owner of content produced by AI poses a fundamental question. *Does the AI model's developer, the user who initiated the content generation, or the AI itself possess ownership over the resulting text?* This complex issue requires legal and ethical clarification to avoid potential disputes and establish a framework that aligns with copyright principles.

The difference between human-authored input and AI-generated output is essential in delineating copyright ownership. While human input initiates the process, the AI's ability to autonomously produce content adds complexity. Clear differentiation is crucial to establish the extent of human influence and delineate ownership rights accurately (Ma, Liu, & Yi, 2023). AI models, including ChatGPT, are underpinned by proprietary algorithms and training methodologies. These represent valuable intellectual property as they determine the model's performance and capabilities. They protect these components as trade secrets, safeguard developers' competitive advantage, and encourage continued innovation.

Trade secret protection must be balanced with the principles of open research. The collaborative nature of AI advancement benefits from knowledge sharing, enabling researchers to build upon existing models. Finding the equilibrium between safeguarding proprietary information and fostering open innovation is essential for the AI community's growth. The advent of Large Language Models like ChatGPT engenders novel discussions around intellectual property rights (Levine, 2023). Copyright issues stemming from AI-generated content necessitate a reexamination of traditional ownership paradigms. At the same time, trade secret protection for AI models prompts a careful balance between proprietary interests and collaborative progress. As these technologies evolve, shaping industries and human-AI interactions, addressing these intellectual property challenges will be instrumental in charting a responsible and innovative path forward.

4 Ethical and Legal Implications

In big data analytics, the concerns surrounding data privacy are paramount. Regulations such as the General Data Protection Regulation (GDPR) have imposed strict requirements on organizations collecting and processing personal data. The GDPR mandates transparent data practices, informed consent, and stringent data protection measures to uphold individuals' privacy rights (Regulation, 2018).

Collecting data for big data analytics and AI model training must involve informed consent from data subjects. This ensures individuals understand how their data will be used and provides them with the choice to participate. Anonymization, the process of removing personally identifiable information, is another critical aspect of data privacy. Proper anonymization safeguards individuals' identities while enabling meaningful analysis (Sorvisto, 2023).

The inherent biases present in data can translate into biased outcomes when utilized in decision-making processes. This is particularly concerning as AI systems, including ChatGPT, learn from historical data, potentially perpetuating and exacerbating biases (Adam, Balagopalan, Alsentzer, Christia, & Ghassemi, 2022).

AI-generated content, including text from ChatGPT, must be scrutinized for fairness and accountability. When AI systems generate content that could influence opinions, disseminate information, or interact with users, it's essential to uphold ethical standards. This includes identifying and rectifying biased or inappropriate content and holding developers accountable for their systems' outputs (Gevaert, Carman, Rosman, Georgiadou, & Soden, 2021).

Transparency is crucial in maintaining user trust. When AI systems like ChatGPT are involved in content creation, users should be informed about the AI's role. Transparency manages user expectations and ensures they are aware that the content is machine-generated, enabling them to assess its credibility critically (Walmsley, 2021).

Accountability becomes a crucial ethical consideration as AI-generated content proliferates across various domains. Establishing responsibility for AI-generated outputs ensures that any potential issues arising from the content can be addressed appropriately. This accountability extends to developers, organizations deploying AI models, and the users interacting with the generated content (Doshi-Velez et al., 2017).

The ethical and legal implications of integrating big data analytics and AI, represented by models like ChatGPT, necessitate careful considerations. Addressing data privacy concerns involves adhering to regulations like GDPR, obtaining informed consent, and ensuring anonymization. Recognizing and mitigating biases in data analysis and AI-generated content promotes fairness and ethical outcomes. Transparency and accountability emerge as central tenets, fostering user trust and upholding ethical standards in AI-driven interactions. As technological landscapes evolve, stakeholders must navigate these ethical and legal dimensions to ensure that the benefits of these technologies are harnessed while upholding individual rights and societal values.

5 Recommendations and Future Considerations

In data-driven innovations, establishing guidelines for patentability assessment requires focusing on the technical contribution and innovation that arise from data analysis. Patent offices should collaborate with domain experts to develop criteria that recognize the transformative nature of insights derived from data, emphasizing their impact on industries and technological progress.

1. Navigating Patent Office Challenges:

Given their abstract nature, data-driven inventions often present unique challenges for patent offices. To address this, patent offices should invest in specialized expertise to accurately evaluate the novelty and non-obviousness of data-driven innovations. Clear communication channels between inventors and patent examiners can foster a mutual understanding of the technical advancements involved.

2. Implementing Verification Mechanisms for Originality:

To manage copyright risks associated with AI-generated content, platforms, and organizations can implement mechanisms that verify the originality of content. This involves using algorithms to cross-reference generated content with existing works to ascertain its uniqueness and original contribution.

3. Differentiating AI-Generated and Human-Generated Content:

Guidelines should be developed to distinguish between AI-generated and human-generated content to address copyright ownership. Clear indicators, such as disclaimers or digital signatures, can help users discern content's origin, ensuring transparent ownership and mitigating potential legal disputes.

4. Incorporating Ethics in AI Development Life Cycle:

Embedding ethics within the entire AI development life cycle is crucial. Developers should undergo training to identify potential biases, understand data privacy regulations, and uphold responsible AI practices. This ensures that ethical considerations are woven into the fabric of AI model creation.

5. Legal and Ethical Considerations in AI Model Deployment:

Organizations deploying AI models, such as ChatGPT, should adopt policies that prioritize legal and ethical compliance. This includes continuously monitoring AI-generated content, implementing user-friendly mechanisms for reporting inappropriate content and fostering transparency about the AI's capabilities and limitations.

6. Evolving IP Laws and Regulations for AI-Generated Content:

Policy adaptations are necessary to accommodate the unique nature of AI-generated content. Evolving intellectual property laws to encompass AI-generated creations while clarifying ownership, usage rights, and duration of protection can foster innovation without compromising legal clarity.

7. International Cooperation in Addressing AI-Related IP Challenges:

Given the global nature of AI and its impact on intellectual property, international cooperation is essential. Collaborative efforts between nations can result in standardized regulations that facilitate cross-border AI deployments while ensuring consistent protection of IP rights.

These recommendations and future considerations aim to provide a roadmap for harmonizing intellectual property protection, ethical AI practices, and innovation in the era of big data analytics and AI-generated content. Stakeholders can navigate the evolving landscape by fostering cooperation, setting clear guidelines, and promoting responsible practices while upholding legal, ethical, and societal standards.

6 Conclusion

In the dynamic landscape shaped by big data analytics and Large Language Models like ChatGPT, the nexus of innovation, ethics, and intellectual property (IP) demands vigilant navigation. The synergistic potential of these technologies presents transformative opportunities across industries, accompanied by an array of intricate challenges.

This research paper has explored the multifaceted implications of this convergence, shedding light on the evolving dynamics of IP in the age of data-driven insights and AI-generated content. From patentability challenges in data analysis-based inventions to the ownership complexities surrounding AI-generated content, the boundaries of traditional IP frameworks are expanding. Data privacy concerns, bias mitigation, transparency, and accountability further underscore the ethical imperatives that stakeholders must prioritize.

A series of recommendations have been delineated in response, offering a roadmap for responsible innovation. Providing guidelines for patentability assessments, implementing mechanisms to manage copyright risks, and fostering ethical AI practices reflect the multifaceted approach required. The consideration of policy adaptations is equally crucial, as emerging AI-related challenges transcend national borders, necessitating international collaboration to shape effective regulatory frameworks.

In navigating these recommendations, stakeholders can achieve a harmonious balance that capitalizes on the transformative power of big data analytics and ChatGPT while safeguarding individual rights, promoting equitable innovation, and adhering to ethical principles. By fostering a holistic approach that marries technological advancement with ethical considerations and legal frameworks, the potential of these technologies can be harnessed for the collective benefit of society.

References

Abbott, R. (2017). Artificial intelligence, big data and intellectual property: protecting computer-generated works in the United Kingdom. *Research Handbook on Intellectual Property and Digital Technologies (Tanya Aplin, ed), Edward Elgar Publishing Ltd, Forthcoming*.

Abdikhakimov, I. (2023). *Legal aspects of AI generated content.* Paper presented at the International Conference on Legal Sciences.

Adam, H., Balagopalan, A., Alsentzer, E., Christia, F., & Ghassemi, M. (2022). Mitigating the impact of biased artificial intelligence in emergency decision-making. *Communications Medicine, 2*(1), 149.

Alzyadat, W., AlHroob, A., Almukahel, I. H., Muhairat, M., Abdallah, M., & Althunibat, A. (2021, 14-15 July 2021). *Big Data, Classification, Clustering and Generate Rules: An inevitably intertwined for Prediction.* Paper presented at the 2021 International Conference on Information Technology (ICIT).

Andanda, P. (2019). Towards a paradigm shift in governing data access and related intellectual property rights in big data and health-related research. *IIC-International Review of Intellectual Property and Competition Law, 50*(9), 1052-1081.

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., . . . Weinberger, D. (2017). Accountability of AI under the law: The role of explanation. *arXiv preprint arXiv:1711.01134*.

Geiger, C., Frosio, G., & Bulayenko, O. (2018). Text and data mining in the proposed copyright reform: Making the EU ready for an Age of Big Data? Legal Analysis and Policy Recommendations. *IIC-International Review of Intellectual Property and Competition Law, 49*, 814-844.

Gevaert, C. M., Carman, M., Rosman, B., Georgiadou, Y., & Soden, R. (2021). Fairness and accountability of AI in disaster risk management: Opportunities and challenges. *Patterns, 2*(11).

Hamza, R., & Pradana, H. (2022). A Survey of Intellectual Property Rights Protection in Big Data Applications. *Algorithms, 15*(11), 418.

Hawashin, B., Althunibat, A., Kanan, T., AlZu'bi, S., & Sharrab, Y. (2023). *Improving Arabic Fake News Detection Using Optimized Feature Selection.* Paper presented at the 2023 International Conference on Information Technology (ICIT).

Kemp, R. (2014). Legal aspects of managing Big Data. *Computer Law & Security Review, 30*(5), 482-491.

Levine, D. S. (2023). Generative Artificial Intelligence and Trade Secrecy.

Lundqvist, B. (2018). Big data, open data, privacy regulations, intellectual property and competition law in an internet-of-things world: The issue of accessing data. *Personal data in competition, consumer protection and intellectual property law: Towards a holistic approach?*, 191-214.

Ma, Y., Liu, J., & Yi, F. (2023). Is this abstract generated by ai? a research for the gap between ai-generated scientific text and human-written scientific text. *arXiv preprint arXiv:2301.10416*.

Mattioli, M. (2014). Disclosing big data. *Minn. L. Rev., 99*, 535.

Moreno, A., & Redondo, T. (2016). Text analytics: the convergence of big data and artificial intelligence. *IJIMAI, 3*(6), 57-64.

Regulation, G. D. P. (2018). General data protection regulation (GDPR). *Intersoft Consulting, Accessed in October*, *24*(1).

Sorvisto, D. (2023). Data Ethics. In *MLOps Lifecycle Toolkit: A Software Engineering Roadmap for Designing, Deploying, and Scaling Stochastic Systems* (pp. 217-236). Berkeley, CA: Apress.

Staegemann, D., Volk, M., Abdallah, M., & Turowski, K. (2023). *Towards the Application of Test Driven Development in Big Data Engineering.* Paper presented at the 2023 International Conference on Information Technology (ICIT).

Taha, O. (2021). Intellectual Property Rights in Jordanian Legislation A Comparative Study with Islamic Jurisprudence a New Endowment Formula *Al-Zaytoonah University of Jordan Journal for Legal studies*, *2*(3), 133-153. doi:10.15849/ZUJJLS.211130.06

Walmsley, J. (2021). Artificial intelligence and the value of transparency. *AI & SOCIETY*, *36*(2), 585-595.