sciendo

Management Systems in Production Engineering

2025, Volume 33, Issue 4, pp. 554-561

Date of submission of the article to the Editor: 02/2025
Date of acceptance of the article by the Editor: 10/2025

DOI 10.2478/mspe-2025-0057

AI-GENERATED HATE SPEECH IN INDUSTRIAL ENTERPRISES: A SYSTEMATIC REVIEW OF CHALLENGES AND MITIGATION STRATEGIES IN THE INDUSTRY 5.0 ERA

Justyna Żywiołek
Czestochowa University of Technology

Abstract:
This article presents a systematic review of the literature on AI-generated hate speech in industrial enterprises, considering the transition from Industry 4.0 to Industry 5.0. The study aims to identify key challenges related to the ethical, legal, and technological aspects of AI implementation in the industrial sector and proposes mitigation strategies to address the risks associated with AI-generated harmful content. The research highlights that AI systems, particularly those based on language models, can unintentionally propagate discriminatory or offensive content, posing risks to corporate reputation, organizational culture, and regulatory compliance. The study underscores the increasing role of regulations, such as the EU's AI Act, which classifies AI-generated content systems as high-risk applications requiring stringent oversight. The article outlines various strategies to mitigate the effects of AI-generated hate speech, including bias detection in AI models, real-time content moderation mechanisms, and human oversight in AI interactions. Transparency, interdisciplinary collaboration, and adherence to ethical guidelines are emphasized as essential components of responsible AI deployment in industrial settings. In conclusion, the study highlights the necessity for further exploration of AI's impact on digital communication within enterprises and the development of global regulatory standards to effectively manage AI-generated content in the Industry 5.0 era.

Key words: artificial intelligence, hate, hate speech, artificial intelligence in the manufacturing industry, industry 5.0

INTRODUCTION

The rapid advancement of artificial intelligence (AI) has reshaped industrial enterprises, driving automation, efficiency, and innovation across various sectors. As Industry 4.0 laid the foundation for data-driven, AI-powered smart manufacturing, Industry 5.0 is redefining the industrial landscape by fostering deeper collaboration between humans and intelligent systems. This new era emphasizes human-centric production, sustainability, and ethical considerations, underscoring the critical role of AI in augmenting human capabilities rather than merely replacing them [1]. However, alongside these technological advancements, AI has also introduced new challenges, particularly in the realm of digital communication and organizational reputation management [2].

One of the emerging concerns in industrial enterprises is the proliferation of AI-generated hate speech, which poses significant ethical, legal, and operational risks. AI-driven language models and content-generation tools, while designed to optimize communication and automate interactions, have inadvertently been exploited for malicious purposes, leading to the spread of harmful content within digital ecosystems [3]. Industrial enterprises, which increasingly rely on AI-powered platforms for marketing, customer engagement, and workforce management, are particularly vulnerable to the adverse effects of AI-generated hate speech. These risks extend beyond corporate image and public relations, potentially influencing workplace culture, employee well-being, and compliance with regulatory frameworks [4].

This systematic review aims to examine the challenges posed by AI-generated hate speech in industrial enterprises, analyzing its implications for corporate governance, ethical AI adoption, and regulatory compliance in the Industry 5.0 era [5]. By synthesizing insights from academic literature, industry reports, and empirical studies, this study seeks to provide a comprehensive overview of the mechanisms through which AI-generated hate speech emerges, its impact on industrial organizations, and potential mitigation strategies. The findings will contribute to the ongoing

discourse on responsible AI development and deployment [6, 7], emphasizing the need for proactive policies, robust monitoring frameworks, and ethical AI governance to ensure a balanced and sustainable integration of AI in industrial enterprises [8, 9].

LITERATURE REVIEW: AI-GENERATED HATE SPEECH IN INDUSTRIAL ENTERPRISES

Artificial intelligence (AI) has become a transformative force across various industries, including industrial enterprises, where it enhances efficiency, automation, and human-machine collaboration [10]. However, as AI systems advance, concerns over their ethical implications grow, particularly in relation to the generation and dissemination of hate speech. AI-generated hate speech refers to offensive, discriminatory, or harmful content produced by AI-driven systems, including chatbots, content generation tools, and automated response mechanisms. This phenomenon raises serious concerns for industrial enterprises as it can impact corporate reputation, workplace culture, regulatory compliance, and operational security [11].

Recent advancements in natural language processing (NLP) and generative AI models, such as large language models (LLMs), have enabled AI systems to generate human-like text. While these systems are designed to assist in communication, content creation, and customer interactions, they are susceptible to biases, misinformation, and manipulation, leading to the unintended or deliberate production of hate speech [12, 13]. Several studies highlight the risks posed by AI-generated hate speech in corporate and industrial settings, including biases in AI models where training data often reflects societal biases, resulting in AI-generated outputs that may reinforce discrimination against specific groups [14, 15]. Industrial enterprises using AI for customer service or marketing also face risks from adversarial attacks, where AI-generated responses can be weaponized for spreading harmful content. Additionally, companies found to be involved in AI-generated hate speech incidents suffer from consumer backlash, loss of trust, and legal ramifications.

As the industry transitions from Industry 4.0 to Industry 5.0, ethical AI governance is gaining traction. Industry 5.0 emphasizes human-centric AI, ensuring that AI technologies align with ethical standards and social responsibility. However, regulatory frameworks for AI-generated hate speech remain fragmented across regions, making compliance complex for global industrial enterprises. The European Union's AI Act (2022) categorizes AI-based content generation tools under high-risk AI applications, requiring stringent oversight and transparency measures [16]. At the corporate level, industrial enterprises are increasingly required to implement ethical AI policies, ensuring that AI-generated content adheres to corporate social responsibility (CSR) standards. Moreover, Industry 5.0 encourages a balance between AI automation and human oversight to prevent unintended consequences of AI-generated hate speech [17].

Addressing AI-generated hate speech requires a multifaceted approach that combines technological, organizational, and regulatory strategies [18]. The literature suggests key mitigation strategies, including bias detection and mitigation by regularly auditing AI models using fairness-aware algorithms and diverse training datasets. Implementing real-time content moderation tools and AI-powered hate speech detection mechanisms can help prevent harmful outputs, while ensuring human reviewers oversee AI-generated content in customer interactions and public-facing communications reduces risks [19]. Establishing clear ethical guidelines and accountability mechanisms within enterprises also helps maintain responsible AI usage. Furthermore, educating employees about AI risks, hate speech detection, and response protocols fosters a culture of responsible AI deployment.

While existing research provides valuable insights into AI-generated hate speech, several gaps remain, necessitating further exploration. More empirical studies are needed to evaluate the financial, reputational, and operational consequences of AI-generated hate speech in industrial settings. Additionally, different industries within the manufacturing sector may experience varying levels of exposure to AI-generated hate speech, requiring tailored mitigation strategies [20]. Advancements in AI explainability can help organizations better understand and mitigate risks associated with hate speech generation, while research into global AI governance frameworks will be critical to ensuring consistency and compliance across jurisdictions [21, 22, 23].

The integration of AI in industrial enterprises offers immense benefits but also presents new challenges, particularly in managing AI-generated hate speech. As Industry 5.0 emphasizes ethical AI and human-machine collaboration, industrial enterprises must proactively implement robust mitigation strategies to safeguard their reputation, regulatory compliance, and workplace culture. Future research should continue exploring innovative solutions to balance AI's transformative potential with ethical considerations, ensuring a responsible and sustainable digital future [24].

METHODOLOGY

This systematic literature review follows a structured approach to examine the challenges and mitigation strategies associated with AI-generated hate speech in industrial enterprises. The methodology comprises several key phases, including data collection, selection criteria, analysis, and synthesis of findings.

The first phase involves an extensive search of academic databases, including Scopus, Web of Science, IEEE Xplore, and Google Scholar, to gather relevant peer-reviewed articles, industry reports, and regulatory documents published between 2015 and 2024. Keywords such as "AI-generated hate speech," "industrial enterprises," "AI ethics," "Industry 5.0," "AI bias," and "AI governance" were used

to ensure comprehensive coverage of relevant literature. To maintain quality and credibility, only studies published in high-impact journals, conference proceedings, and official government or industry reports were considered.

The selection criteria included studies that explicitly address AI-generated hate speech within an industrial or corporate context, discuss the implications of AI bias and security risks, and propose mitigation strategies. Studies focusing solely on general AI ethics without a connection to industrial enterprises were excluded. A two-stage screening process was employed: first, titles and abstracts were reviewed to filter out irrelevant studies, followed by a full-text review of selected papers to confirm their relevance.

For data analysis, thematic synthesis was used to categorize findings into major themes, including AI bias detection, content moderation techniques, human-AI collaboration, regulatory frameworks, and ethical AI governance. Sentiment analysis and content coding techniques were applied to analyze industry reports and regulatory documents to identify emerging trends and gaps in the existing literature. Additionally, comparative analysis was conducted to evaluate how different industrial sectors approach AI-generated hate speech mitigation.

The methodological rigor of this review ensures a balanced and comprehensive examination of the subject, offering valuable insights into both the challenges posed by AI-generated hate speech and potential solutions for industrial enterprises navigating Industry 5.0. By systematically analyzing academic and industry perspectives, this study contributes to the ongoing discourse on AI ethics, security, and responsible implementation within corporate environments.

MATERIALS AND METHODS

The emergence of AI-generated hate speech as a research topic has gained scholarly attention in recent years. To understand the origins and early discussions on this issue, a bibliometric analysis was conducted using the Scopus database. This analysis focused on the years 2014 to 2025, with keywords such as “AI” and “hate speech” to identify relevant studies. The search revealed a total of 290 publications during this period, indicating that while AI-generated hate speech was not a dominant theme, it was already recognized as an emerging concern within academic discourse.

The bibliometric approach involved the preparation of an evolution-map, based on keyword co-occurrence analysis, citation analysis, and co-authorship network mapping. These methods allowed for the identification of key trends and the thematic evolution of research on AI-generated hate speech. Analyzing the development of topics over time made it possible to determine the major phases of this field's formation, highlighting periods of intensified research activity and the emergence of new perspectives. By tracking the dynamics of key concepts, it was possible to capture the shifting thematic scope and its impact on the advancement of interdisciplinary studies in this area. The use of Scopus as the primary database ensured that the analysis was based on high-quality, peer-reviewed sources, offering reliable insights into the development of this research field. The presence of 290 publications in this timeframe highlights that the phenomenon of AI-generated hate speech was in its formative stages, setting the groundwork for later, more extensive studies.

This bibliometric analysis underscores the novelty of AI-generated hate speech as a research topic and its gradual evolution into a significant issue in AI ethics and industrial applications. Understanding its early development provides a foundation for further exploration into mitigation strategies and regulatory considerations in contemporary industrial enterprises (Fig. 1).

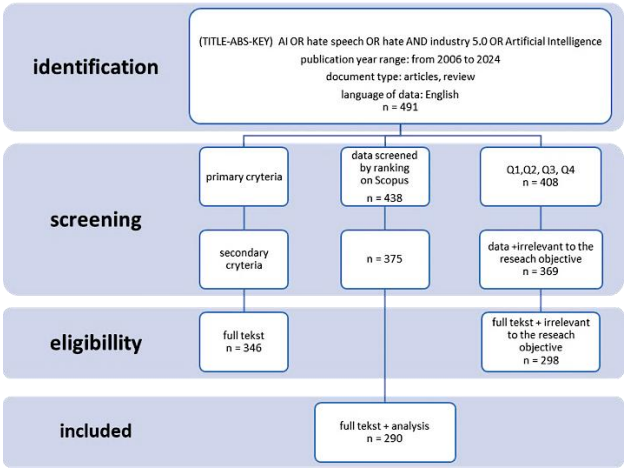

Fig. 1 Flow diagram of the article search and selection process

RESULTS

Artificial intelligence (AI) has become an essential tool in addressing the growing challenges of online communication, including the detection and moderation of hate speech. As concerns over digital safety, misinformation, and ethical AI deployment continue to rise, research in this area has expanded significantly. Understanding the trends, key developments, and gaps in existing studies is crucial for shaping future advancements in AI-driven hate speech detection. Research in this area has evolved over time, reflecting the rising concerns about ethical AI deployment, online safety, and regulatory challenges. Understanding the development of studies on AI and hate speech is essential for identifying key trends, technological advancements, and existing gaps. The following results provide an overview of research progress in this field, highlighting the dynamics of scholarly contributions and the role of AI in addressing hate speech in the context of Industry 5.0 (Fig. 2).

The Figure 2 illustrates the publication trends on AI and hate speech from 2014 to 2025. The data shows a minimal number of publications from 2014 to 2018, with only a slight increase in 2017. However, from 2019 onwards, there is a noticeable upward trend, which becomes more pronounced in 2020.

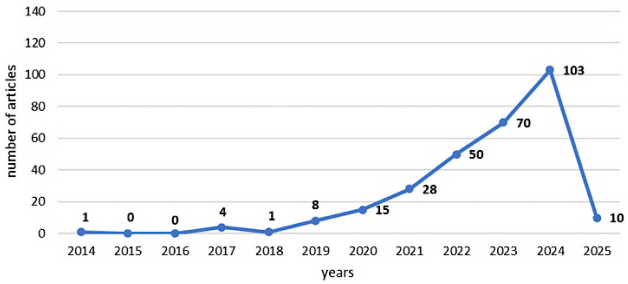

Fig. 2 Distribution of the publications by year

Between 2020 and 2023, the number of documents grew significantly, reflecting a growing academic interest in the topic. The highest number of publications is recorded in 2024, indicating peak engagement in this research area. However, in 2025, there is a sharp decline, which could be attributed to incomplete data collection for the year or a possible saturation of the topic.

The trend suggests that AI-generated hate speech is a relatively new research area that has gained substantial attention in recent years. The rapid increase in publications aligns with the broader discourse on AI ethics, misinformation, and content moderation. This pattern highlights the importance of continued research in this field, particularly in developing mitigation strategies and regulatory frameworks to address AI-generated hate speech.

The classification of research articles based on journal publications provides insights into the distribution of studies on artificial intelligence (AI) and hate speech across various academic outlets. The dataset comprises 290 articles, with notable contributions from a few leading journals and a highly fragmented distribution among numerous other sources.

The journal *Ceur Workshop Proceedings* is the most significant publication venue, contributing 25 articles, establishing itself as a central platform for discussions on AI and hate speech. Other notable contributors include *Lecture Notes in Computer Science, including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics* (14 articles), *Communications in Computer and Information Science* (12 articles), and *Lecture Notes in Networks and Systems* (11 articles). These journals highlight the interdisciplinary nature of the research field, integrating elements of computational intelligence, data analysis, and AI-driven content moderation.

Beyond these leading sources, several journals published between two and five articles, playing a crucial role in shaping research in AI and hate speech detection. Journals such as *IEEE Access* (4 articles), *ACM International Conference Proceeding Series* (3 articles), and *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (3 articles) reflect the increasing interest in the intersection of AI and content moderation strategies. Similarly, *Proceedings of the International Joint Conference on Neural Networks* (3 articles) and *Smart Innovation Systems and Technologies* (3 articles) emphasize the technological advancements and computational approaches relevant to addressing AI-generated hate speech.

The remaining 203 journals published only one or two articles each, demonstrating the fragmented nature of AI and hate speech research. Some key examples include *Big Data and Society* (2 articles), *Information Communication and Society* (2 articles), and *Computers and Education Artificial Intelligence* (2 articles). This wide dispersion of research suggests that AI and hate speech research spans multiple disciplines, including computer science, social sciences, ethics, and legal studies.

While the broad scope of journal publications indicates a growing academic interest in AI and hate speech, the field remains highly fragmented without a dominant central research platform. Given the increasing societal concerns regarding AI-generated hate speech, there is a need for more consolidated research efforts. A dedicated journal or special issues focused on AI's role in content moderation and ethical AI governance could enhance interdisciplinary collaboration and provide a structured approach to tackling the challenges associated with AI and hate speech detection (Table 1).

Table 1
The number of publications in each journal

Journal Name	Number of Publications
Ceur Workshop Proceedings	25
Lecture Notes In Computer Science Including Subseries Lecture Notes In Artificial Intelligence And Lecture Notes In Bioinformatics	14
Communications In Computer And Information Science	12
Lecture Notes In Networks And Systems	11
IEEE Access	4
ACM International Conference Proceeding Series3	3
Electronics	3
Proceedings Of The ACM SIGKDD International Conference On Knowledge Discovery And Data Mining	3
Proceedings Of The International Joint Conference On Neural Networks	3
Smart Innovation Systems And Technologies	3
Big Data And Society	2
Information Communication And Society	2
Computers And Education Artificial Intelligence	2
Other journals	203

The rapid advancement of artificial intelligence (AI) has significantly influenced various aspects of digital communication, content moderation, and societal interactions. AI technologies, such as natural language processing (NLP) and machine learning, offer substantial benefits in automating tasks, improving efficiency, and enhancing decision-making. However, they also introduce challenges related to ethics, misinformation, and content regulation. One of the most pressing concerns in this domain is the proliferation of AI-generated hate speech, which poses risks to individuals, communities, and organizations across digital platforms. This issue extends beyond online

interactions, impacting industries that rely on AI for value-chain operations and corporate sustainability.

The increasing reliance on AI-driven systems for monitoring and moderating online discourse has led to extensive research on detecting and mitigating hate speech. However, industries such as media, technology, and governance have encountered challenges in balancing the use of AI for content regulation while maintaining freedom of expression and ethical AI deployment. Research trends indicate a growing interest in AI's role in detecting hate speech, with a significant increase in scholarly publications from 2018 onward. This surge aligns with global discussions on digital ethics, misinformation management, and the need for regulatory frameworks to mitigate AI-driven risks. Additionally, businesses across various sectors, including the energy and resource industries, are recognizing the role of AI in value-chain processes, from operational efficiency to ethical compliance and risk mitigation.

Building a robust and ethical **value-chain** is crucial for organizations integrating AI into their operations. A well-structured value-chain ensures that AI applications align with corporate responsibility, regulatory standards, and stakeholder expectations. The **result of effectively making a value-chain** is a system that optimizes AI capabilities while mitigating risks associated with bias, misinformation, and unethical AI applications. This structured approach not only enhances operational efficiency but also strengthens brand reputation, fosters consumer trust, and ensures compliance with evolving digital governance frameworks.

Despite advancements in AI-powered content moderation, gaps remain in understanding the biases embedded within AI models, the unintended consequences of automated hate speech detection, and the regulatory implications of AI deployment in public and private sectors. Moreover, the impact of AI-driven content moderation on corporate value-chain operations, brand reputation, and consumer trust remains an underexplored area. This study aims to analyze the evolution of AI and hate speech research while emphasizing the necessity of **constructing a strong value-chain** that integrates AI responsibly. By exploring key technological developments, their implications, and the pressing issues that warrant further investigation, this analysis contributes to the discourse on ethical AI deployment, digital transformation, and corporate responsibility in the Industry 5.0 era (Fig. 3).

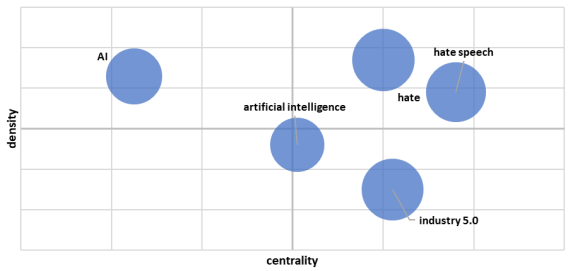

Fig. 3 Evolving Competencies and Skillsets for from 2014 to 2025

The strategic diagram illustrates the thematic evolution of AI and hate speech research from 2014 to 2025, highlighting their relevance, interconnections, and developmental trends. The two key dimensions represented-density and centrality-offer insights into the significance and maturity of these research themes. Density indicates the internal robustness of a theme, reflecting its development and depth, while centrality signifies its integration within the broader research landscape.

The figure reveals AI as an established and well-developed research theme, positioned in the upper-left quadrant with high density but moderate centrality. This suggests that AI research is deeply rooted within its domain, with extensive studies contributing to its theoretical and practical advancements. However, its moderate centrality indicates that AI's integration with other research topics, such as digital ethics, governance, and content regulation, still has room for expansion. Hate speech and AI-related discussions appear in the upper-right quadrant, demonstrating both high density and high centrality. This positioning underscores the growing importance of hate speech detection and mitigation in AI research, reflecting the increased focus on algorithmic biases, ethical considerations, and regulatory frameworks. The concentration of research in this area aligns with the surge in AI-driven moderation technologies, including natural language processing (NLP) applications and machine learning-based detection systems.

The central position of artificial intelligence within the diagram, connected to various other themes, reinforces its foundational role in multiple disciplines. The notable intersection between AI and hate speech suggests a significant body of work focused on addressing online toxicity, misinformation, and algorithmic fairness. Research themes such as Industry 5.0 also appear in the diagram, indicating an increasing interest in leveraging AI for ethical content governance within industrial and corporate applications. Moreover, the diagram highlights the evolution of AI's role in content regulation, showcasing the necessity for more sophisticated models that balance free speech with online safety. As AI-powered hate speech detection continues to evolve, challenges persist regarding biases in training data, over-policing of marginalized communities, and the unintended consequences of automated content moderation.

The visual representation suggests that AI-driven hate speech moderation is progressing towards greater integration with ethical AI governance, cybersecurity, and human oversight models. To enhance the effectiveness of AI in combating hate speech, future research must focus on developing bias-free AI models that account for linguistic and cultural variations, strengthening regulatory frameworks that define AI's role in content moderation while preserving freedom of expression, enhancing explainability and transparency in AI decision-making to increase trust among users and policymakers, and exploring AI's role in industrial applications, particularly

within Industry 5.0, to mitigate reputational and operational risks linked to hate speech propagation. The findings of the strategic diagram underscore AI's growing role in addressing online hate speech, with an emphasis on ethical AI implementation and regulatory alignment. As digital platforms increasingly rely on automated moderation tools, the need for interdisciplinary collaboration in AI governance, policy development, and technological refinement becomes more pressing. Moving forward, fostering responsible AI applications will be crucial to ensuring that advancements in AI-driven hate speech detection are both effective and ethically sound.

EVOLUTION MAP

In this section, we explore the thematic evolution of artificial intelligence (AI), hate speech detection, and Industry 5.0 between 2014 and 2025. Using bibliometric analysis tools, we identified key research topics emerging in distinct sub-periods, revealing a dynamic shift in focus from general AI applications to the ethical, regulatory, and technological challenges associated with hate speech moderation and AI integration within Industry 5.0. Prior to the research window of 2016-2018, discussions on AI were primarily centered around automation, machine learning algorithms, and the enhancement of computational capabilities, with limited focus on ethical concerns or societal impacts (Russell and Norvig, 2010; Bostrom, 2014). Subsequently, the rise of social media platforms and digital communication channels accelerated the need for AI-driven content moderation, leading to an increased focus on hate speech detection by 2019. During this period, key topics included natural language processing (NLP), sentiment analysis, and bias mitigation in AI models, highlighting concerns about fairness and accountability in automated decision-making. By 2020, research trends began emphasizing the role of Industry 5.0 in fostering human-centric AI applications, promoting collaboration between humans and intelligent systems to enhance ethical AI governance. From 2021 onwards, new research topics such as explainable AI (XAI), regulatory compliance, and digital ethics gained prominence, reflecting the growing need to balance innovation with responsible AI deployment. These themes are now recognized as critical components for ensuring transparency, reducing algorithmic bias, and improving public trust in AI systems. As seen in Figure 4, this evolution underscores the importance of interdisciplinary research in addressing the challenges posed by AI in content moderation and industrial applications, emphasizing the convergence of technological, ethical, and regulatory considerations.

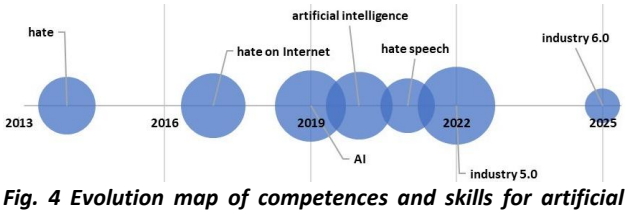

Fig. 4 Evolution map of competences and skills for artificial intelligence application themes over periods

The provided Figure 4 illustrates the thematic evolution of AI, hate speech, and industrial transformations over time, capturing key research developments from 2014 to 2025. The placement and size of the thematic clusters reflect the density and centrality of research topics, providing insights into their growth and interconnections.

The figure indicates that early discussions on hate speech and AI were relatively isolated, with the term "hate" appearing independently around 2013. This suggests that initial research focused on general hate-related content without direct links to AI-driven moderation technologies. By 2016, the concept of "hate on the Internet" emerged, signifying the increasing recognition of online hate speech as a distinct challenge within digital communication environments. This transition aligns with the growing influence of social media platforms and the need for regulatory and technological interventions.

A significant expansion occurs around 2019, where AI-related topics, such as "artificial intelligence" and "AI", become more prominent and start intersecting with hate speech research. This convergence highlights a crucial moment in the development of AI-driven hate speech detection and moderation strategies. The clustering of these themes indicates a shift towards automation in digital content regulation, reflecting advancements in natural language processing, machine learning, and bias detection methodologies.

The peak of research activity is observed in 2022, where "hate speech", "AI", and "Industry 5.0" form a dense network, emphasizing their integration within industrial and governance frameworks. Industry 5.0's appearance in this period signifies a shift towards more human-centric AI applications, where ethical considerations and collaborative AI-human interactions become key focal points. This suggests that AI is not only being used for hate speech detection but also integrated into broader industrial strategies emphasizing responsibility, transparency, and regulatory compliance.

Looking ahead, the emergence of "Industry 6.0" in 2025 suggests an evolving discourse on the next generation of industrial automation and AI applications. However, the smaller size of this cluster implies that research in this area is still in its early stages, requiring further exploration. The trajectory presented in the figure underscores the progressive evolution of AI applications, from isolated technological innovations to integrated systems influencing both digital communication and industrial governance.

The analysis of this figure highlights the growing interconnection between AI, hate speech moderation, and industrial digitalization. It suggests that future research should focus on refining AI governance frameworks, ensuring algorithmic transparency, and addressing ethical challenges posed by automated moderation systems. As AI continues to evolve, its implications for industry, online discourse, and regulatory landscapes will require continued interdisciplinary collaboration to strike a balance between innovation and societal responsibility.

DISCUSSION AND CONCLUSION

The intersection of AI, hate speech detection, and Industry 5.0 highlights the increasing importance of integrating ethical considerations into AI-driven industrial applications. Industry 5.0 emphasizes a human-centric approach, ensuring that AI technologies not only optimize efficiency but also uphold ethical standards and social responsibility. Hate speech detection serves as a key application where AI can contribute to digital governance, yet challenges such as bias, fairness, and contextual understanding remain critical barriers to overcome.

A major issue in AI-driven hate speech detection is the accuracy and impartiality of automated systems. While natural language processing and machine learning models have significantly advanced in identifying hate speech, they often struggle with linguistic nuances, cultural differences, and contextual ambiguity. This has implications for both online platforms and industrial environments where AI is used for compliance monitoring, workplace communication analysis, and reputational management.

Industry 5.0 further broadens the application of AI beyond digital platforms, incorporating it into organizational decision-making, human-machine collaboration, and regulatory compliance. The challenge lies in developing AI solutions that align with corporate ethics while mitigating risks associated with biased decision-making. Transparent and explainable AI (XAI) methodologies are critical in fostering trust in AI-driven hate speech detection, particularly in regulatory and legal frameworks.

Future research should focus on refining AI governance frameworks to ensure accountability and algorithmic fairness. Ethical AI deployment in Industry 5.0 requires interdisciplinary collaboration among policymakers, industry leaders, and AI researchers to establish standards that balance technological progress with societal responsibility. The continued evolution of AI in industrial applications presents an opportunity to enhance workplace inclusivity, safety, and ethical governance while addressing digital hate speech more effectively.

Overall, the findings from this analysis emphasize the need for a structured approach to integrating AI into Industry 5.0, ensuring that innovations in automation, human-AI collaboration, and digital content moderation align with ethical standards and regulatory requirements. As AI continues to develop, it is essential to prioritize transparency, accountability, and fairness to mitigate unintended consequences and foster responsible AI integration across industries.

REFERENCES

[1] B. Azgin and S. Kiralp, "Surveillance, Disinformation, and Legislative Measures in the 21st Century: AI, Social Media, and the Future of Democracies," *Social Sciences*, vol. 13, no. 10, p. 510, 2024, doi: 10.3390/socsci13100510.

[2] M.E. Cortés-Cediel, A. Segura-Tinoco, I. Cantador, and M.P. Rodríguez Bolívar, "Trends and challenges of e-government chatbots: Advances in exploring open government data and citizen participation content," *Government Information Quarterly*, vol. 40, no. 4, p. 101877, 2023, doi: 10.1016/j.giq.2023.101877.

[3] R.Ó. Fathaigh, T. Dobber, F. Zuiderveen Borgesius, and J. Shires, "Microtargeted propaganda by foreign actors: An interdisciplinary exploration," Maastricht *Journal of European and Comparative Law*, vol. 28, no. 6, pp. 856-877, 2021, doi: 10.1177/1023263X211042471.

[4] K. Kertysova, "Artificial Intelligence and Disinformation," *Secur. Hum. Rights*, vol. 29, 1-4, pp. 55-81, 2018, doi: 10.1163/18750230-02901005.

[5] S. Saniuk, S. Grabowska, and A. Thibbotuwawa, "Challenges of industrial systems in terms of the crucial role of humans in the Industry 5.0 environment," *Production Engineering Archives*, vol. 30, no. 1, pp. 94-104, 2024, doi: 10.30657/pea.2024.30.9.

[6] H. Iftikhar, M. Qureshi, J. Zywiołek, J.L. López-Gonzales, and O. Albalawi, "Short-term PM2.5 forecasting using a unique ensemble technique for proactive environmental management initiatives," *Front. Environ. Sci*., vol. 12, 2024, doi: 10.3389/fenvs.2024.1442644.

[7] L.Y. Hunter, C.D. Albert, J. Rutland, K. Topping, and C. Hennigan, "Artificial intelligence and information warfare in major power states: how the US, China, and Russia are using artificial intelligence in their information warfare and influence operations," *Defense & Security Analysis*, vol. 40, no. 2, pp. 235-269, 2024, doi: 10.1080/14751798.2024.2321736.

[8] J.A. Goldstein, J. Chao, S. Grossman, A. Stamos, and M. Tomz, "How persuasive is AI-generated propaganda?," PNAS Nexus, vol. 3, no. 2, pgae034, 2024, doi: 10.1093/pnasnexus/pgae034.

[9] J. Gonçalves, I. Weber, G.M. Masullo, M. Da Torres Silva, and J. Hofhuis, "Common sense or censorship: How algorithmic moderators and message type influence perceptions of online content deletion," *New Media & Society*, vol. 25, no. 10, pp. 2595-2617, 2023, doi: 10.1177/14614448211032310.

[10] Wenzelburger, "Between Technochauvinism and Human-Centrism: Can Algorithms Improve Decision-Making in Democratic Politics?," *European Political Science*, vol. 2022, p. 1, 2022.

[11] M.A. Khan et al., "Security and Privacy Issues and Solutions for UAVs in B5G Networks: A Review," *IEEE Trans. Netw. Serv. Manage*., p. 1, 2024, doi: 10.1109/TNSM.2024.3487265.

[12] M. Husnain, Q. Zhang, M. Usman, K. Hayat, K. Shahzad, and M. W. Akhtar, "How Chatbot negative experiences damage consumer-brand relationships in hospitality and tourism? A mixed-method examination," *International Journal of Hospitality Management*, vol. 126, p. 104076, 2025, doi: 10.1016/j.ijhm.2024.104076.

[13] Galantino, "How Will the EU Digital Services Act Affect the Regulation of Disinformation?," SCRIPTed, vol. 20, p. 89, 2023.

[14] T. Baviera, L. Cano-Orón, and D. Calvo, "Tailored Messages in the Feed? Political Microtargeting on Facebook during the 2019 General Elections in Spain," *Journal of Political Marketing*, pp. 1-20, 2023, doi: 10.1080/15377857.2023.2168832.

[15] T. Ahmad, E.A. Aliaga Lazarte, and S. Mirjalili, "A Systematic Literature Review on Fake News in the COVID-19 Pandemic: Can AI Propose a Solution?," *Applied Sciences*, vol. 12, no. 24, p. 12727, 2022, doi: 10.3390/app122412727.

[16] F. Filgueiras, "The politics of AI: democracy and authoritarianism in developing countries," *Journal of Information Technology & Politics*, vol. 19, no. 4, pp. 449-464, 2022, doi: 10.1080/19331681.2021.2016543.

[17] I. Hussain, M. Qureshi, M. Ismail, H. Iftikhar, J. Zywiołek, and J.L. López-Gonzales, "Optimal features selection in the high dimensional data based on robust technique: Application to different health database," *Heliyon*, vol. 10, no. 17, e37241, 2024, doi: 10.1016/j.heliyon.2024.e37241.
[18] D. Helbing, *Towards Digital Enlightenment*. Cham: Springer International Publishing.
[19] J. Żywiołek, "Knowledge-Driven Sustainability: Leveraging Technology for Resource Management in Household Operations," *ECKM*, vol. 25, no. 1, pp. 974-982, 2024, doi: 10.34190/eckm.25.1.2375.
[20] J. Żywiołek, "Empirical examination of AI-powered decision support systems: ensuring trust and transparency in information and knowledge security," *SPSUTOM*, vol. 2024, no. 197, pp. 679-695, 2024, doi: 10.29119/1641-3466.2024.197.37.
[21] J. Żywiołek, A. Szymonik, T. Smal, *Navigating Supply Chain Turbulence*. New York: Productivity Press, 2025, https://doi.org/10.4324/9781003561736.
[22] H. Tumber and S. Waisbord, The Routledge Companion to Media Disinformation and Populism: Routledge.
[23] E. Fetahi, M. Hamiti, A. Susuri, J. Ajdari, and X. Zenuni, "AI-Based Hate Speech Detection in Albanian Social Media: New Dataset and Mobile Web Application Integration," *Int. J. Interact. Mob. Technol*., vol. 18, no. 24, pp. 190-208, 2024, doi: 10.3991/ijim.v18i24.50851.
[24] J. Żywiołek, K. Mathiyazhagan, U. Shahzad, X. Zhao, and T. Saikouk, "Enhancing cognitive metrics in supply chain management through information and knowledge exchange," *IJLM, 2025*, doi: 10.1108/IJLM-04-2024-0243.

Justyna Żywiołek
ORCID ID: 0000-0003-0407-0826
Czestochowa University of Technology
Faculty of Management
ul. Dabrowskiego 69, 42-201 Czestochowa, Poland
e-mail: justyna.zywiolek@pcz.pl