IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 8 February 2024, accepted 12 March 2024, date of publication 18 March 2024, date of current version 28 March 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3379010

RESEARCH ARTICLE

How Generative AI Was Mentioned in Social Media and Academic Field? A Text Mining Based on Internet Text Data

WENCHAO ZHANG[1,2], RUONAN YAN[1], AND LEI YUAN[1,2]

[1]Faculty of Education, Guangxi Normal University, Guilin 541004, China
[2]Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, Guilin 541004, China

Corresponding authors: Lei Yuan (9761541@qq.com) and Wenchao Zhang (410677657@qq.com)

This work was supported in part by the National Social Science Foundation of China (NSSFC) under Grant BCA230277, in part by Guangxi Education Science "14th Five-Year Plan" 2022 Commissioned Key Project under Grant 2022AA12, and in part by Guangxi Education Science Planning Project under Grant 2022JD12.

ABSTRACT As ChatGPT has evolved, generative AI (Artificial Intelligence) has gone viral on the internet since 2022. Heated discussions on generative AI have appeared in both social media and academic field, generating massive textual data. Overwhelming media coverage of generative AI may lead to biased conception. To date, there has been no systematic analysis of how generative AI is mentioned on the internet. Moreover, little attention has been paid to demonstrating the gap in perceptions of generative AI between social media and academic field. This study seeks to focus on the following specific research questions: What are the key terms related to generative AI, what are the key term differences in social media and academic field on generative AI, and what are the topic differences of generative AI in social media and academic field? A text-mining approach supported by KH-coder was employed. The research data were drawn from two main text sources: the Sina Weibo platform and the CNKI periodical database. The results revealed statistically significant differences in key terms and topics related to generative AI between the social media and academic field. Our findings enhance the understanding of public ideas and the trend of generative AI on the internet, and provide supportive information for future studies on generative AI applications.

INDEX TERMS Generative AI, AIGC, KH-coder, text mining.

I. INTRODUCTION

At the end of 2020, as OpenAI released the public version of ChatGPT, information regarding generative AI (Artificial Intelligence) has headlined the Internet [1], [2], generating extensive interest from scholars and the general public. Generative AI, which conceptually resembles AIGC, has powerful capabilities for automatically analyzing and generating text, images, videos, etc. [1]. However, although overwhelming social media coverage provides individuals with numerous opportunities to have a glimpse of generative AI, the content on social media tends to be biased or sometimes misleading [3]. In the academic community, the issue of generative AI applications in mainstream industries [4], [5], [6], [7], copyright related to originality [8], technical limitations [9], [10], and security and safety problems [11], [12] have received considerable critical attention. Until now, far too little attention has been paid to revealing how generative AI was mentioned on the Internet. In addition, there is a lack of research on the gap between social media and the academic field in discussing generative AI. That is to say, academic circles do not quite understand what the public is discussing about generative AI and vice versa.

The extensive discussions on generative AI on the Internet have generated substantial text data, offering a unique opportunity to investigate the gap between social media and academic field. This study focuses on Chinese-language sources to provide insights into the cultural and regional

The associate editor coordinating the review of this manuscript and approving it for publication was Vivek Kumar Sehgal.

nuances of generative AI's perception. It situates its findings within a broader theoretical framework by drawing on media representation, public perception of AI, and the intersection of technology with societal issues. The work on the role of social media in shaping public discourse provides a valuable lens through which to view the portrayal of generative AI in social media. Additionally, the critical analysis of AI's societal impact offers insights into the broader context of AI's role in society. Our analysis is informed by these theoretical perspectives, aiming to provide a more comprehensive understanding of the portrayal of generative AI in both social media and academic field.

Sina Weibo, one of the most popular social media platforms in China, has seen a surge in information about generative AI since the release of ChatGPT. This platform's user-generated content provides a rich source for understanding the public's perception of generative AI within a specific cultural context.The Sina Weibo platform allows users to post and share short text-based messages on the Internet, thereby creating rich user-generated content. Many users post their daily activities and seek or share information. Although each post is limited to 140 characters, within numerous users, it can generate large-scale text data. Big text data can be collected and utilized in text mining for additional implicit, previously unknown, and potentially useful information for academic, educational, financial, and other purposes. Along with the heated discussion on generative AI in social media, academic databases have presented a large number of articles on generative AI. This content is undoubtedly in textual form and is beneficial for text mining studies. Recently, an increasing number of scholars have used automatic text analysis tools to analyze large volumes of text data to avoid manual reading and to find useful information for business [13], [14], [15], education [16], [17], [18], [19], health care [20], [21], [22], etc.

Despite the widespread use of generative AI and its impact on social media and academia, little is known about what has been mentioned in these two domains. To better understand this, we focused on the following research questions(RQs).

RQ1: What are the key terms related to generative AI?

RQ2: What are the differences in key terms related to generative AI in the social media and academic field?

RQ3: What are the topic differences between generative AI in social media and academic field?

To answer these research questions, a comprehensive text-mining study on generative AI is required. Language plays an important role in social networks [23]; users tend to participate in the same social network if they speak the same language. Therefore, in this study, we mainly focused on Chinese texts on the Internet. Our work provides a unique perspective on the gap in perceptions related to generative AI between social media and academic field, focusing on the Chinese language and cultural context. This study's findings can serve as a foundation for future research that includes a broader range of languages and platforms.

II. METHODOLOGY

A. TEXT MINING

If all the data in the world are equivalent to the water on Earth, then textual data are like the ocean, making up a majority of the volume [24]. However, analyzing massive amounts of multidimensional raw text data is complicated and time-consuming. Tata mining has capabilities in this regard [25]. Text mining is a methodology for discovering novel, valuable, and useful information, knowledge, and hidden patterns from large datasets using various statistical approaches [26]. Fuzzy Set Theory plays an important role in text mining [27] and is particularly useful for modeling the semantics of text, including fuzzy concepts and certain types of linguistic expressions [26]. This theory enables text mining to handle uncertain, imprecise, or vague information. Thus, through text mining, researchers can automatically analyze large-scale texts and generate implicit, previously unknown, and potentially useful insights from textual data [19].

Text mining is based on various advanced techniques stemming from statistics, machine learning, and linguistics [28]. Examples include word-level analysis (e.g., frequency analysis), word association analysis (e.g., network analysis), and advanced techniques (e.g., text classification, text clustering, topic modeling, information retrieval, and sentiment analysis). Some of the first applications of text mining came about when people were trying to organize documents [29] and have been implemented on text data stored in structured databases [30]. Owing to easy access to digital text and convenient data gathering techniques, data sources applied in text mining studies have changed progressively from existing databases to customized datasets conducted by individuals. Text mining has been widely applied to analyze transcripts and speeches, meeting transcripts, academic journal articles, websites, emails, blogs, microblogs, and social media networking sites across a broad range of application areas [31].

Text mining approaches are quicker, more favorable, time-saving, and objective compared with traditional models for transforming data to knowledge with some manual analysis and interpretation [25]. The workflow in texting mining consists of three processes: text data extraction, text data preprocessing, and text data analysis. In the first step, text data are extracted and gathered from various sources such as web pages, databases, and blogs. In the second step, data cleaning and reduction were implemented to filter or remove irrelevant text segments. In the final step, researchers use text mining tools to complete a specific analysis (word-level analysis, word association analysis, advanced techniques) and refine the discovered information to solve specific research questions. When conducting data-mining research, it is advisable to formulate clear research questions in advance. However, this can also be accomplished with vague research questions. This is because data mining often yields interesting and creative information beyond the researcher's expectations during the research process.

This study, guided by the aforementioned research questions, adheres to a problem-solving paradigm predicated on

TABLE 1. Text data source and dataset.

Database	Sina Weibo platform	CNKI periodical database
Keywords	生成式人工智能 (generative artificial intelligence); 生成式 AI(generative AI); AIGC(AIGC); ChatGPT	
Dataset	Sina Weibo posts 215,777 words	Academic articles 111,137 words

Note. Words in ( ) are the English translation.

text mining methodologies delineated in [19]. Our inquiry encompasses a text mining analysis of generative AI discourse across social media and academic domains, with a specific focus on Chinese-language content.

B. DATA ACQUISITION AND PREPROCESSING

The Internet is a rich source of text data, where researchers can collect text data from social media platforms, online databases, etc. The text data acquired in this study were from two different sources: Sina Weibo posts as social media text data sources and CNKI periodical databases as academic field text data sources.

Sina Weibo users can choose to post their opinions, arguments, experiences, and so forth publicly to everyone on public timelines or privately to selected users on private homepages. It is difficult to collect private posts available only to users' friends or followers. Collecting private Sina Weibo posts may violate user privacy. Hence, the text data collected from the Sina Weibo platform were all posts that appeared in the public timeline. We used text data crawler software named "Bazhuayu" (https://www.bazhuayu.com) to automatically fetch Sina Weibo posts related to generative AI. The keywords searched on the Sina Weibo platform are summarized in Table 1. There were 215,777 text data points in this collection. We tagged them as "Sina Weibo posts" and saved them in a txt file.

CNKI is a comprehensive academic database that contains a large amount of Chinese academic literature, dissertations, patent documents, and so on. It is a resourceful database for collecting academic textual data. We established several criteria to ensure that we fetched academic text data related to the generative AI. The core and CSSCI periodical databases in CNKI were selected as academic article text data sources. We searched several keywords in these databases, including "generative artificial intelligence", "generative AI", "AIGC", and "ChatGPT", to increase the likelihood of capturing academic articles related to generative AI. In the selection step, we read the titles and abstracts of the collected articles to check whether each of them was related to our aims and removed duplicate and irrelevant articles. Finally, we extracted the titles and abstracts of 632 articles from the periodic database in CNKI. Text data of 111,137 Chinese characters were collected and tagged as "Academic articles." We saved them in the same txt file as mentioned above.

Data preprocessing is a fundamental step in text mining, which normalizes words in unstructured data. This is a crucial process for the noise removal and tokenization of text data. For instance, if data preprocessing is skipped before advanced data analysis in this research, "生成式人工智能" (generative artificial intelligence)in the text data will be tokenized as two separate words: "生成式"(generative) and "人工智能" (artificial intelligence). However, since "生成式人工智能" (generative artificial intelligence) is the keyword in this research, it should be unified and counted as one single word for further analysis. Thus, we used "Select Words to Analyze" function in KH-coder to force pick up following strings as individual word: 人 工 智 能 (artificial intelligence), 生 成 式 人 工 智 能 (generative artificial intelligence), 生成式AI (generative AI).

C. TEXT DATA ANALYSIS

The dataset was analyzed using KH-coder Version 3.0, a versatile text-mining software that leverages Natural Language Processing (NLP) to extract and analyze information from multilingual text data. KH-coder's compatibility with Windows, Linux, and Macintosh, along with its support for over 10 languages—spanning English, Chinese, Japanese, Korean, German, and Spanish—positions it as a robust tool for cross-cultural text analysis. Although our study captures a static snapshot of generative AI discourse, we acknowledge the potential for KH-coder to facilitate longitudinal studies in future research, thereby tracking the evolution of discourse over time. The software's user-friendly interface and comprehensive functionality made it an ideal choice for our data mining endeavors, particularly during the text data analysis phase.

First, we employed the text parsing function to enumerate word frequencies, meticulously categorizing each parsed term into a frequency list. This process, while effective for our current analysis, could be expanded in future studies to include temporal dimensions, allowing for the examination of term frequency trends over time. We only preserved nouns in the word frequency list and excluded verbs, adjectives, adverbs, auxiliary words, prepositions, interjections, and conjunctions from the word frequency list. Because they contribute little to this study, they may even cause confusion in advanced analyses. Second, feature word extraction and co-occurrence network analysis were executed using KH-coder's functions, revealing the thematic landscape of generative AI discussions. These analyses, while insightful for our current study, could be extended in future research to include temporal analysis, providing a more dynamic understanding of the evolution of key terms and concepts. Feature word extraction and co-occurrence network analysis can help us discover the differences in key terms in different text domains, both specifically and visually. Third, topic analysis was conducted using the crosstab function in the KH-coder. We artificially extracted four main topics from the word frequency list: *Technology*, *Education*, *Society*, *Economy*. We used the KH-coder to investigate the proportion of mentions of these topics in different text domains.

TABLE 2. Top 50 noun key terms related to generative AI.

Key terms	Freq.[a]	Key terms	Freq.
人工智能 (artificial intelligence)	3870	工具(tool)	486
AI(AI)	2798	社会(society)	470
ChatGPT(ChatGPT)	2528	能力(ability)	464
技术(technology)	2247	知识(knowledge)	455
模型(model)	1105	中国(China)	426
生成式 AI (generative AI)	1099	安全(safety)	424
数据(data)	1096	监管(regulation)	411
内容(content)	908	影响(influence)	410
应用(application)	878	办法(measure)	402
发展(development)	867	用户(user)	402
教育(education)	836	创新(innovation)	386
智能(intelligence)	802	网络(net)	381
服务(service)	735	苹果(Apple)	381
信息(information)	691	行业(industry)	381
人类(human)	662	企业(enterprise)	379
领域(domain)	658	全球(global)	375
公司(company)	645	管理(management)	371
人(people)	634	产品(product)	363
问题(issue)	625	算法(algorithm)	349
AIGC(AIGC)	605	法律(law)	344
风险(risk)	594	视频(video)	340
科技 (science and technology)	583	时代(era)	338
语言(language)	575	方面(aspect)	337
研究(research)	552	治理(govern)	333
数字(digit)	493	市场(market)	331

[a]"Freq." means "Frequency". Words in ( ) are the English translation.

III. RESULT

A. KEY TERMS RELATED TO GENERATIVE AI

Word frequency analysis is the most popular technique in text mining [32], which is based on Natural Language Processing to automatically count and list word frequencies from text data. Using word-frequency analysis, researchers can provide an overview of the key terms mentioned in the dataset. Therefore, we applied the word frequency function of the KH-coder to solve RQ1 in this study. The top 50 noun key terms with their word frequencies are listed in Table 2. By looking at the key terms related to generative AI, we can see that the top 10 key terms all belong to the technical vocabulary. "ChatGPT" as a new technology was mentioned enormously. In addition, the key terms "model," "data," "content," "application" is the main concerns of generative AI related text data.

B. KEY TERM DIFFERENCES IN SOCIAL MEDIA AND ACADEMIC FIELD ON GENERATIVE AI

1) FEATURE WORDS EXTRACTION

To answer RQ2, we applied the feature word extraction function of the KH-coder. By utilizing this function, key terms in different text domains can be automatically extracted as feature words. Table 3 displays the results obtained from the feature word extraction in Sina Weibo posts and academic articles related to generative AI. This result shows that the feature words in Sina Weibo posts and academic articles on generative AI are different. In Sina Weibo posts, feature words "company," "service," "enterprise" are all relevant to business. And "education" is not extracted as a feature word in Sina Weibo posts but in academic articles. On the other hand, to our surprise "ChatGPT," "technology" is not the feature word in Sina Weibo post, but in academic articles. These results suggest that the generative AI mentioned in social media is mostly related to the economic domain. In contrast, in academic articles, generative AI is mentioned in diverse fields such as technology, education, and academic research.

TABLE 3. Feature words extraction.

Sina Weibo posts		Academic articles	
AI (AI)	.247	ChatGPT (ChatGPT)	.290
生成式 (generation)	.153	人工智能 (artificial intelligence)	.273
生成式 AI (generative AI)	.122	技术 (technology)	.213
公司 (company)	.075	教育 (education)	.125
服务 (service)	.065	内容 (content)	.117
科技 (science and technology)	.052	发展 (development)	.106
人 (people)	.051	应用 (application)	.098
办法 (measure)	.045	人类 (human)	.098
工具 (tool)	.041	智能 (intelligence)	.093
企业 (enterprise)	.039	研究 (research)	.091

Note. The values in this table represent Jaccard similarity coefficients. The words in ( ) are the English translation.

2) CO-OCCURRENCE NETWORK OF KEY TERMS

Co-occurrence networks can visualize the relationships between key terms in text data. It can also represent similarities and differences in key terms in various text domains. To visually observe the coherence and distinction of the key terms in Sina Weibo posts and academic articles on generative AI, we used a KH-coder to generate the co-occurrence network of the key terms. Fig.1 displays the co-occurrence network of the key terms in the collected datasets. As shown in Fig.1, the green circles represent the key terms numerously mentioned in Sina Weibo posts and academic articles, which indicate the coherence of the key terms in these two different text domains. Yellow circles display key feature terms related to specific text domains. As can be seeing in Fig.1, technical vocabulary, such as "ChatGPT," "technology," "artificial intelligence," "data," etc. are all connected with Sina Weibo posts and academic articles. This result indicates that both social media and academia are concerned with the technical aspects of generative AI. On the other hand, in Fig.1, key feature terms with different domain characteristics are

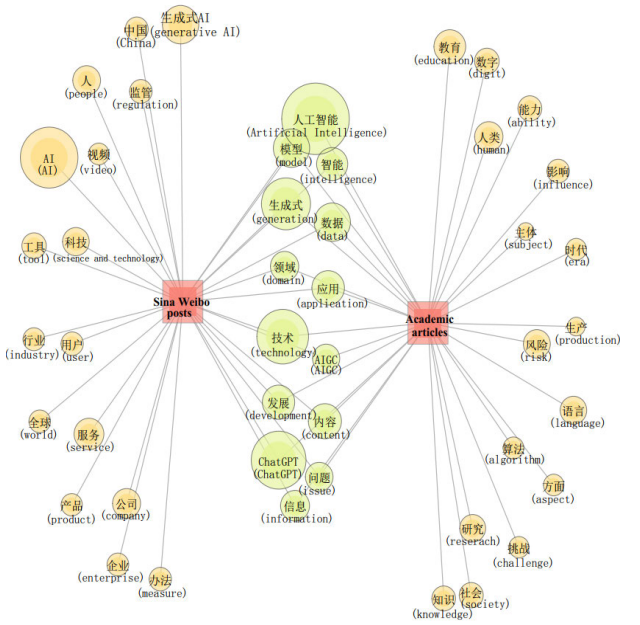

FIGURE 1. Co-occurrence network of key terms.

associated with Sina Weibo posts and academic articles. This result supports previous results, as social media focuses more on the economic aspect of generative AI, and the academic field has a broader focus on generative AI.

Note. Squares represent text domains. The size of each circle represents the frequency of the key terms. Line indicates the relationship. The words in ( ) are the English translation.

C. TOPICS MENTIONED IN SOCIAL MEDIA AND ACADEMIC FIELD ON GENERATIVE AI

1) CODING RULE MAKING

The coding function in KH-coder can automatically code text data according to a coding rule developed by researchers. It is typically applied to analyze the ratio of specific topics mentioned in different text domains. Thus, this function is suitable for solving RQ3. Coding rule making is a fundamental and essential procedure when utilizing the coding function in a KH-coder. Therefore, we referenced the word frequency list and artificially categorized it into four main topics–*Technology*, *Society*, *Education*, *Economy*–and created a coding rule text file. The KH-coder can read coding rules from the text file and code the text data using a topic with multiple words. Finally, the ratio of the four main topics mentioned in the Sina Weibo posts and academic articles on generative AI was extracted separately.

2) CROSSTAB ANALYSIS

To distinguish the differences in topics mentioned in Sina Weibo posts and academic articles on generative AI, we employed a crosstab analysis in KH-Coder. Table 4 and Fig.2 present summary statistics of the results. Table 4 illustrates the proportions of the different topics mentioned in the generative AI-related text data. One issue that emerged from

TABLE 4. Proportions of different topics mentioned In text data.

	Technology	Education	Society	Economy	N of Documents
Sina Weibo posts	4697 (66.37%)	738 (10.43%)	2975 (42.04%)	2586 (36.54%)	7077
Academic articles	2674 (81.65%)	1231 (37.59%)	2230 (68.09%)	780 (23.82%)	3275
Total	7371 (71.20%)	1969 (19.02%)	5205 (50.28%)	3366 (32.52%)	
chi-square	254.162**	1070.466**	606.896**	164.616**	

** $P < 0.01$

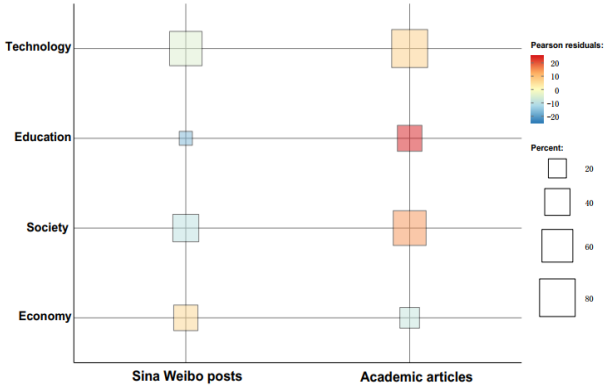

FIGURE 2. Bubble plot of tabulation results.

the result was in text data related to generative AI, *Technology* (71.20%) as a main topic was mentioned in the top spot, followed by *Society* (50.28%), *Economy* (32.52%), and *Education* (19.02%). The same sequence was also presented in the Sina Weibo posts. This result implicitly suggests that education is not yet an attractive area of focus in generative AI discourse compared with technology, society, and economy.

What stand out in Table 4 presents the proportions of the four main topics in Sina Weibo posts and academic articles. With the exception of Economy, the proportions of Technology, Education, and Society were lower in Sina Weibo posts but higher in academic articles. Furthermore, the proportions of these four topics were significantly different (P <0.01). This outcome statistically revealed that compared to the academic field, discussions on generative AI in social media were more concerned with the economic realm.

The KH-coder can generate tabulation results as a bubble plot and provide better visibility. Fig.2 shows the cross-tabulation data by bubble plot with four topics and two text domains. In this bubble plot, the square sizes were used to represent the ratio of texts to which each topic code was applied. Fig.2 shows significant differences in generative AI-related topics between Sina Weibo posts and academic articles. Especially on the topic of education, the square of academic articles was double the square of Sina Weibo posts.

IV. DISCUSSION, LIMITATION, AND FUTURE STUDY

To the best of our knowledge, this is the first study to use a text mining approach to analyze generative AI related text data from different text domains. In this study, Sina Weibo posts

and academic articles on generative AI were collected from the Sina Weibo and CNKI periodic databases, respectively.

The primary objective of this study was to discern salient terminologies associated with generative AI within the digital textual landscape, with a particular emphasis on the Chinese-language milieu. Our engagement with the extant literature on media representation and public perception of AI has contextualized our findings, and we advocate for future research to incorporate sentiment analysis and qualitative methodologies to augment our quantitative insights. Through analyzing the words frequency list generated by text mining software KH-coder, this study found that words belong to technology were mentioned massively in generative AI text data, such as ''ChatGPT,'' ''model,'' ''data,'' etc. Generative AI, as a newly emerging and rapidly evolving technology, still contains considerable uncertainties. These uncertainties and powerful capabilities of generative AI have attracted people from different fields, including the general public and the academic community, to discuss technologies related to generative AI. In terms of lexical categorization, generative AI belongs to the technology vocabulary and carries technological attributes from the very beginning of its existence. Thus, it is not surprising that a large number of technology terms appeared in generative AI-relevant texts on the Internet.

Addressing the second research question, our analysis delineated distinct terminological patterns in Sina Weibo posts and academic publications, underscoring the divergent priorities and perspectives that characterize the discourse on generative AI within social media and academic spheres. This revelation accentuates the imperative for a more holistic comprehension of generative AI's multifaceted portrayal across diverse domains. The most obvious finding to emerge from the feature word extraction and co-occurrence network analysis is that, compared with the academic field, generative AI mentioned in social media is relatively focused on economy and has a weak connection with education. The educational application of generative AI is still being discussed and confined to a relatively small community. A possible explanation for this might be that technologies represented by generative AI were initially associated with economy and were not created for educational purposes. Moreover, technologies have a stronger impact on the economy than education. Therefore, the public is inclined to pursue the economic outcomes of new technologies rather than educational applications. This inclination will be reflected implicitly in social media dialogues.

On the question of what are the topic differences of generative AI in social media and academic articles, this study found the proportions of different topics mentioned in generative AI related text data ranked as *Technology*, *Society*, Economy, and *Education*. Except *Economy*, the proportion of *Technology*, *Society* and *Education* was lower in Sina Weibo posts, but higher in academic articles. Furthermore, significant differences (P <0.01) of these four topics in the two text domains were confirmed through crosstab analysis and visualized in a bubble plot. These results revealed that distinctive differences exist in the social media and academic field of generative AI. Texts displayed on social media represent an inclined and biased perception of generative AI. It is possible to hypothesize that issues relevant to generative AI massively discussed in the academic field might not be acknowledged in a timely manner by the public, and the public's perception and discussion on generative AI is still not comprehensive.

The preliminary findings of this research contribute to a more nuanced comprehension of the prevailing concerns within social media and academic discourses on generative AI, particularly within the confines of the Chinese language ecosystem. These insights may guide future inquiries by underscoring the significance of linguistic and cultural considerations in the analysis of generative AI's societal impact. Both domains should mutually inform each other of each other. To avoid group polarization [33] resulting from an excessively homogeneous perception, social media must be more engaged with diverse academic perspectives on generative AI. Correspondingly, in the academic field, there must be a thorough consideration of the public's perception of generative AI, along with exploring various ways to present a diverse range of academic discussions and results on generative AI via social media.

Surprisingly, this study showed that there were fewer posts on generative AI related to education and more posts on economy. This reflects the pursuit of generative AI to achieve economic benefits. This public discussion of generative AI presented on social media will also have an implicit impact on the perception of generative AI in the education sector. Educators may overly excite generative AI's potential to bring about transformative changes in education, such as increasing efficiency and facilitating personalized learning. However, incomplete and unsystematic understanding of generative AI, compounded by the over-promotion of AI's capabilities on social media, may lead to abuse and misuse of generative AI in education. Ultimately, technopoly [34], as pointed out by Neil Bozeman, which results in social institutions, such as schools forfeiting their sovereignty to technology, reemerges in the realm of education without awareness. All technologies have a life cycle and generative AI is no exception as a new technology. Thus, it is impossible to prevent generative AI from affecting the authority of educational institutions. Nevertheless, it is possible to share diverse academic findings related to generative AI on social media and to improve educators' understanding of generative AI. Educators with a more thorough understanding of generative AI via social media will be able to utilize generative AI more appropriately in education.

This study has three limitations that are integral to its scope and methodology. Firstly, our focus on Chinese-language sources, while providing a culturally nuanced perspective, may not fully represent the global spectrum of generative AI perceptions, as cultural and regional idiosyncrasies profoundly influence discourse. Secondly, the dynamic and rapidly changing nature of Internet content means that our analysis captures a static moment in an ever-evolving

discourse, potentially obscuring temporal nuances. Lastly, the technical limitations prevented us from collecting the text data of comments corresponding to the posts on generative AI, which would have allowed us to gauge public sentiment and engagement with the topics more comprehensively. Despite these constraints, this study offers valuable insights into the Chinese-language discourse on generative AI and proposes avenues for future research to encompass a broader linguistic and cultural milieu, including the integration of user-generated comments and longitudinal analysis to capture the full breadth of generative AI's evolving discourse.

Despite the need for cautious interpretation of the results, this study significantly advances our comprehension of the pivotal terminology, terminological disparities, and thematic variations pertaining to generative AI within Chinese language social media and academic discourse. Subsequent research endeavors could extend the scope of textual data analysis to encompass a multilingual array of platforms, thereby capturing the dynamic dimensions of generative AI-related content. Furthermore, it would be insightful to explore the public's response to generative AI-related posts and scholarly insights across interdisciplinary domains, offering a richer tapestry of perspectives on the societal impact and reception of this burgeoning technology.

ACKNOWLEDGMENT

(Ruonan Yan is co-first author.)

REFERENCES

[1] F.-Y. Wang, J. Yang, X. Wang, J. Li, and Q.-L. Han, "Chat with ChatGPT on industry 5.0: Learning and decision-making for intelligent industries," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 4, pp. 831–834, Apr. 2023.

[2] T. Wu, S. He, J. Liu, S. Sun, K. Liu, Q.-L. Han, and Y. Tang, "A brief overview of ChatGPT: The history, status quo and potential future development," *IEEE/CAA J. Autom. Sinica*, vol. 10, no. 5, pp. 1122–1136, May 2023, doi: 10.1109/JAS.2023.123618.

[3] C. Zhang, C. Zhang, S. Zheng, Y. Qiao, C. Li, M. Zhang, S. K. Dam, C. M. Thwal, Y. L. Tun, L. L. Huy, D. kim, S.-H. Bae, L.-H. Lee, Y. Yang, H. Tao Shen, I. S. Kweon, and C. S. Hong, "A complete survey on generative AI (AIGC): Is ChatGPT from GPT-4 to GPT-5 all you need?" 2023, *arXiv:2303.11717*.

[4] W. Zhang, "Application and development of robot sports news writing by artificial intelligence," in *Proc. IEEE 2nd Int. Conf. Data Sci. Comput. Appl. (ICDSCA)*. Dalian, China: IEEE, Oct. 2022, pp. 869–872, doi: 10.1109/ICDSCA56264.2022.9988077.

[5] D. Baidoo-Anu and L. O. Ansah, "Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning," *J. AI*, vol. 7, pp. 52–62, Dec. 2023, doi: 10.61969/jai.1337500.

[6] J. Kietzmann, J. Paschen, and E. Treen, "Artificial intelligence in advertising: How marketers can leverage artificial intelligence along the consumer journey," *J. Advertising Res.*, vol. 58, no. 3, pp. 263–267, Sep. 2018, doi: 10.2501/jar-2018-035.

[7] M. Farina and A. Lavazza, "ChatGPT in society: Emerging issues," *Frontiers Artif. Intell.*, vol. 6, Jun. 2023, Art. no. 1130913, doi: 10.3389/frai.2023.1130913.

[8] T. He, "The sentimental fools and the fictitious authors: Rethinking the copyright issues of AI-generated contents in China," *Asia Pacific Law Rev.*, vol. 27, no. 2, pp. 218–238, Jul. 2019, doi: 10.1080/10192557.2019.1703520.

[9] A. Borji, "A categorical archive of ChatGPT failures," 2023, *arXiv:2302.03494*.

[10] C. Chen, J. Fu, and L. Lyu, "A pathway towards responsible AI generated content," 2023, *arXiv:2303.01325*.

[11] M. Gupta, C. Akiri, K. Aryal, E. Parker, and L. Praharaj, "From ChatGPT to ThreatGPT: Impact of generative AI in cybersecurity and privacy," *IEEE Access*, vol. 11, pp. 80218–80245, 2023, doi: 10.1109/ACCESS.2023.3300381.

[12] Y. Yuan, W. Jiao, W. Wang, J.-T. Huang, P. He, S. Shi, and Z. Tu, "GPT-4 is too smart to be safe: Stealthy chat with LLMs via cipher," 2023, *arXiv:2308.06463*.

[13] J. Li, H. J. Wang, Z. Zhang, and J. L. Zhao, "A policy-based process mining framework: Mining business policy texts for discovering process models," *Inf. Syst. e-Bus. Manage.*, vol. 8, no. 2, pp. 169–188, Mar. 2010, doi: 10.1007/s10257-009-0112-x.

[14] Saura and Bennett, "A three-stage method for data text mining: Using UGC in business intelligence analysis," *Symmetry*, vol. 11, no. 4, p. 519, Apr. 2019, doi: 10.3390/sym11040519.

[15] S. Bhattacharjee, D. Delen, M. Ghasemaghaei, A. Kumar, and E. W. T. Ngai, "Business and government applications of text mining & natural language processing (NLP) for societal benefit: Introduction to the special issue on 'text mining & NLP,'" *Decis. Support Syst.*, vol. 162, Nov. 2022, Art. no. 113867, doi: 10.1016/j.dss.2022.113867.

[16] Q. Liu, S. Zhang, Q. Wang, and W. Chen, "Mining online discussion data for understanding teachers reflective thinking," *IEEE Trans. Learn. Technol.*, vol. 11, no. 2, pp. 243–254, Apr. 2018, doi: 10.1109/TLT.2017.2708115.

[17] J. Martí-Parreño, E. Méndez-Ibáñez, and A. Alonso-Arroyo, "The use of gamification in education: A bibliometric and text mining analysis," *J. Comput. Assist. Learn.*, vol. 32, no. 6, pp. 663–676, Dec. 2016, doi: 10.1111/jcal.12161.

[18] R. Ferreira-Mello, M. André, A. Pinheiro, E. Costa, and C. Romero, "Text mining in education," *WIREs Data Mining Knowl. Discovery*, vol. 9, no. 6, pp. 1–49, Nov. 2019, doi: 10.1002/widm.1332.

[19] J. Yang, Kinshuk, and Y. An, "A survey of the literature: How scholars use text mining in educational studies?" *Educ. Inf. Technol.*, vol. 28, no. 2, pp. 2071–2090, Feb. 2023, doi: 10.1007/s10639-022-11193-3.

[20] C. Dreisbach, T. A. Koleck, P. E. Bourne, and S. Bakken, "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data," *Int. J. Med. Informat.*, vol. 125, pp. 37–46, May 2019, doi: 10.1016/j.ijmedinf.2019.02.008.

[21] S. Hyun and C. Cooper, "Application of text mining to nursing texts: Exploratory topic analysis," *CIN, Comput., Informat., Nursing*, vol. 38, no. 10, pp. 475–482, Oct. 2020, doi: 10.1097/cin.0000000000000681.

[22] X. Li, A. Yang, and H. Yan, "Priorities and instruments of local elderly care policies in China: Text mining and comparative analysis," *Frontiers Public Health*, vol. 9, Jul. 2021, Art. no. 647670, doi: 10.3389/fpubh.2021.647670.

[23] A. Java, X. Song, T. Finin, and B. Tseng, "Why we Twitter: Understanding microblogging usage and communities," in *Proc. 9th WebKDD 1st SNA-KDD workshop Web mining social Netw. Anal.*, San Jose, CA, USA, Aug. 2007, pp. 56–65, doi: 10.1145/1348549.1348556.

[24] E. Siegel, *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die, Revised and Updated*. Hoboken, NJ, USA: Wiley, 2015, doi: 10.1002/9781119172536.

[25] W. Xiao, L. Jing, Y. Xu, S. Zheng, Y. Gan, and C. Wen, "Different data mining approaches based medical text data," *J. Healthcare Eng.*, vol. 2021, pp. 1–11, Dec. 2021, doi: 10.1155/2021/1285167.

[26] C. Justicia de la Torre, D. Sánchez, I. Blanco, and M. J. Martín-Bautista, "Text mining: Techniques, applications, and challenges," *Int. J. Uncertainty, Fuzziness Knowl.-Based Syst.*, vol. 26, no. 4, pp. 553–582, Aug. 2018, doi: 10.1142/s0218488518500265.

[27] R. Kruse, C. Borgelt, and D. Nauck, "Fuzzy data analysis: Challenges and perspectives," in *Proc. FUZZ-IEEE IEEE Int. Fuzzy Systems. Conf.*, Seoul, South Korea: IEEE, 1999, pp. 1211–1216, doi: 10.1109/FUZZY.1999.790074.

[28] M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, "Chapter 1—Analytics defined," in *Information Security Analytics*, M. R. M. Talabis, R. McPherson, I. Miyamoto, J. L. Martin, and D. Kaye, Eds. Boston, MA, USA: Syngress, 2015, pp. 1–12, doi: 10.1016/B978-0-12-800207-0.00001-0.

[29] D. R. Cutting, D. R. Karger, J. O. Pedersen, and J. W. Tukey, "Scatter/gather: A cluster-based approach to browsing large document collections," in *Proc. 15th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retr. SIGIR*. Copenhagen, Denmark: ACM Press, 1992, pp. 318–329, doi: 10.1145/133160.133214.

[30] M. Goebel and L. Gruenwald, "A survey of data mining and knowledge discovery software tools," *ACM SIGKDD Explorations Newslett.*, vol. 1, no. 1, pp. 20–33, Jun. 1999, doi: 10.1145/846170.846172.
[31] H. Hassani, C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi, "Text mining in big data analytics," *Big Data Cognit. Comput.*, vol. 4, no. 1, p. 1, Jan. 2020, doi: 10.3390/bdcc4010001.
[32] C. Zhai and S. Massung, *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. New York, NY, USA: Morgan & Claypool, 2016, doi: 10.1145/2915031.
[33] L. Iandoli, S. Primario, and G. Zollo, "The impact of group polarization on the quality of online debate in social media: A systematic literature review," *Technological Forecasting Social Change*, vol. 170, Sep. 2021, Art. no. 120924, doi: 10.1016/j.techfore.2021.120924.
[34] N. Postman, *Technopoly: The Surrender of Culture to Technology*, 1st Vintage Books ed. New York, NY, USA: Vintage Books, 1993.

WENCHAO ZHANG received the Ph.D. degree in language and culture from Osaka University, Osaka, Japan. He is currently an Associate Professor with the Faculty of Education, Guangxi Normal University, Guilin, China. He is also a Teacher with the Department of Educational Technology, Faculty of Education, Guangxi Normal University. He is a member of the Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University, and the Director of the Institute of Basic Education Informatization, Guangxi Education Informatization Development Research Centre. He has been involved in the development of information technology textbooks for primary school. His research interests include artificial intelligence in education, natural language processing, m-learning, and steam.

RUONAN YAN is currently an Associate Professor with the Faculty of Education, Guangxi Normal University, Guilin, China. She is the Deputy Director of Guangxi Ethnic Education Development Research Centre, which is a key research base for humanities and social sciences in universities, Guangxi. She is also the Director of the Educational Anthropology Professional Committee, China Union of Anthropological and Ethnological Science (CUAES). She is dedicated to the study of ethnic education and the integration of art education and ICT. Her research interests include ethnic culture and education, aesthetic education, and music.

LEI YUAN is currently a Professor with the Faculty of Education, Guangxi Normal University, Guilin, China. He is the Deputy Director of the Key Laboratory of Education Blockchain and Intelligent Technology, Ministry of Education, Guangxi Normal University. He is also the Executive Deputy Director of Guangxi Education Informatization Development Research Center. He devoted himself to research in ICT educational applications and steam education. He has published more than ten books, along with more than 50 CSSCI articles. His research interests include the educational applications of ICT, steam education, and elementary education.