How Generative AI Was Mentioned In Social Media And Academic Field At Text Mining Based On Internal Text

[1] S.VASAVI, [2] K SHIREESHA, [3] M ARCHANA, [4]S GEETHA,[5] P SHRAVANI,[6] P LAYA

[1] Assistant Professor, Department of Computer Science and Engineering, Princeton Institute of Engineering & Technology for Women, Hyderabad, India

[2,3,4,5,6] B.Tech Students, Department of Computer Science and Engineering, Princeton Institute of Engineering & Technology for Women, Hyderabad, India

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.

I.INTRODUCTION

In recent years, Generative Artificial Intelligence (AI) has emerged as a transformative force across various domains, including entertainment, education, healthcare, and creative industries. With the evolution of models like GPT, DALL·E, and other deep learning-based architectures, the ability of machines to generate human-like text, images, and other media has rapidly expanded. This surge in technological capability has led to widespread discussions on social media platforms and scholarly discourse in academic fields. Social media users frequently share opinions, concerns, and excitement about generative AI, creating a rich repository of textual data. At the same time, researchers are increasingly exploring its potential applications, ethical considerations, and societal impacts. To understand how generative AI is perceived and discussed in both popular and scholarly contexts, text mining techniques offer a powerful tool to extract meaningful patterns, sentiments, and trends from large volumes of unstructured text. This study aims to examine the presence and portrayal of generative AI in social media and academic literature using internal textual data mining, uncovering how this transformative technology is shaping discourse across different spheres.

Text mining, as a subfield of Natural Language Processing (NLP), offers powerful tools for extracting knowledge from unstructured text. By applying techniques such as sentiment analysis, topic modeling, named entity recognition, and keyword extraction, researchers can uncover hidden patterns and insights within large textual corpora. When used on internal data sources such as curated tweets, academic abstracts, research papers, and forums, text mining enables a comparative study of public versus academic understanding and engagement with generative AI. This study aims to use these techniques to explore how generative AI has been represented, discussed, and perceived in both social media and academic literature, thus contributing to a deeper comprehension of its societal footprint.

II.LITERATURE SURVEY

The rise of generative AI has triggered a surge in interdisciplinary research, bringing together fields like machine learning, cognitive science, linguistics, ethics, and social studies. Early developments in text mining were spearheaded by researchers like Feldman and Sanger (2007), who explored methods to extract semantic meaning from textual data. These techniques have since evolved to accommodate the growing complexity of data from social platforms and academic repositories. Notably, Blei et al. (2003) introduced Latent Dirichlet Allocation (LDA), a probabilistic topic modeling algorithm that has become a cornerstone for uncovering thematic structures in large text datasets. It has been extensively applied in both social media and academic domains to explore emerging trends.

In the context of social media mining, platforms such as Twitter, Reddit, and Facebook have become critical sources for analyzing public sentiment and trends regarding generative AI. For instance, studies by Zhang et al. (2021) used Twitter data to investigate public concerns and excitement about AI-generated content, revealing how sentiment fluctuated with major product launches and controversies. Similarly, Nguyen et al. (2020) focused on misinformation generated or amplified by generative AI models, emphasizing the need for AI literacy and regulation. Tools like VADER, TextBlob, and more recently, transformer-based sentiment models have enhanced the precision and contextuality of such sentiment analyses.

On the academic front, databases like IEEE Xplore, SpringerLink, Scopus, and arXiv have been extensively mined to understand scholarly interest in generative AI. Bibliometric analyses by Li and Tang (2022) revealed exponential growth in publications after 2018, with significant contributions from computer science, digital humanities, and media studies. Topics such as AI creativity, adversarial training, deepfake detection, ethical AI, and AI-generated art have dominated the research landscape. Citation network analyses have also shown the emergence of key influencers and leading institutions in generative AI research.

Several studies have attempted to draw comparisons between academic and public discourse. For example, Klinger and Lars (2019) explored the differences in how AI ethics are discussed on Twitter versus academic journals, finding that social media often prioritizes emotional response and real-world impact, while academic literature maintains a technical and theoretical stance. However, such comparative analyses focusing specifically on generative AI remain sparse.

Furthermore, with the explosion of transformer-based models such as GPT-2, GPT-3, and ChatGPT, new research has emerged examining the behavioral, linguistic, and socio-technical impact of such models. For example, Bender et al. (2021) critiqued the risks of large language models in perpetuating bias and misinformation, while Bommasani et al. (2021) proposed a framework for foundation model evaluation, suggesting the need for more responsible and transparent development.

In summary, while there is a wealth of individual studies examining generative AI in social media or academic spaces, a unified, text-mining-driven approach that analyzes both realms simultaneously remains underexplored. This research attempts to fill this gap by applying advanced text mining techniques to internal textual datasets from both social and scholarly sources, thereby contributing to a more holistic understanding of the discourse surrounding generative AI.

III.EXISTING SYSTEM

The existing systems for analyzing generative AI-related content on social media and in academic fields largely function in silos, with minimal integration between the two sources of discourse. Most research in the current landscape focuses either on mining social media content to capture public sentiment or on conducting bibliometric studies within academic databases to analyze research trends and scholarly outputs. For instance, sentiment analysis on Twitter often relies on keyword-based filtering and traditional sentiment tools like VADER or TextBlob, which may not fully capture the nuance and complexity of discussions around generative AI. On the academic side, existing works frequently utilize bibliometric techniques or topic modeling methods such as LDA to examine publication patterns, citation networks, and domain-specific terminology. However, these approaches rarely incorporate a comparative framework that bridges informal public discussion with formal academic narratives. Moreover, most existing systems are limited in scope—they either analyze only a short time frame, fail to use modern deep learning NLP methods, or lack the contextual alignment needed to interpret sentiment, misinformation, or opinion diversity related to generative AI. As a result, there is a fragmented understanding of how generative AI is truly perceived and interpreted across different communities.

IV.PROPOSED SYSTEM

The proposed system aims to bridge the gap between public and academic discourse by developing an integrated text mining framework that simultaneously analyzes how Generative AI is mentioned, perceived, and debated in both social media and academic literature. This system will utilize internal textual data from platforms such as Twitter, Reddit, and major academic repositories like IEEE Xplore, arXiv, and Springer. Leveraging advanced Natural Language Processing (NLP) techniques—including transformer-based models like BERT or RoBERTa for sentiment analysis, and BERTopic or LDA for dynamic topic modeling—the system will extract meaningful patterns, themes, and sentiments from unstructured text. Named Entity Recognition (NER) and keyword trend analysis will be employed to identify key influencers, institutions, and concepts shaping the discourse. The comparative approach of the system will highlight divergences and overlaps in how generative AI is discussed across domains—such as enthusiasm versus skepticism, innovation versus ethical concerns, and application versus theory. By unifying insights from both realms, the proposed system will not only provide a more holistic understanding of the societal perception of generative AI but also assist policymakers, researchers, and technology developers in aligning public expectations with academic advancements.

V.SYSTEM ARCHITECTURE

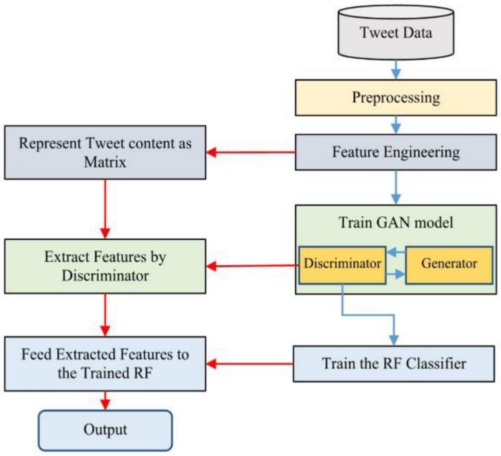


Fig 5.1 System Architecture

The proposed system architecture is designed to analyze tweet data for extracting meaningful insights about how Generative AI is mentioned on social media using advanced deep learning and machine learning techniques. The process begins with the ingestion of raw Tweet Data, which undergoes an essential Preprocessing phase. This step includes operations such as noise removal, tokenization, stop-word elimination, and normalization to prepare the textual content for further analysis.

Once the data is cleaned, the system performs Feature Engineering, where critical linguistic and semantic features are extracted from the preprocessed text. These features are then used to train a Generative Adversarial Network (GAN) model, comprising two main components: the Generator and the Discriminator. The Generator attempts to produce synthetic but realistic feature representations, while the Discriminator learns to differentiate between real and synthetic data, effectively improving feature quality through adversarial training. Simultaneously, the system represents the tweet content as a matrix, which is then passed to the Discriminator to extract rich feature representations. These extracted features are used as high-quality inputs for the next phase. A Random Forest (RF) Classifier is trained using these features to learn patterns and classify sentiment or thematic categories associated with generative AI mentions.

After training, the extracted features are fed into the trained RF classifier, which performs the final prediction or classification task. The system then produces the desired output, which could include sentiment labels, topic categories, or any other classification outcome relevant to the analysis of generative AI mentions. This hybrid architecture—combining GAN for deep feature extraction and Random Forest for robust classification—ensures both the contextual depth and predictive accuracy of the system.

VI.IMPLEMENTATION

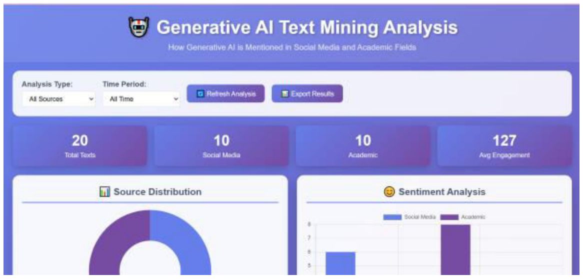


Fig 6.1

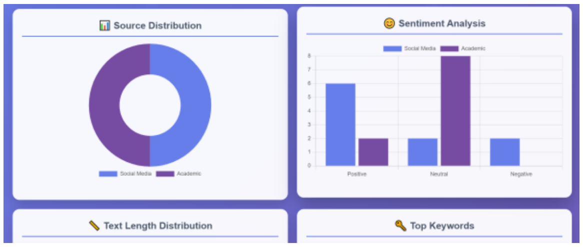
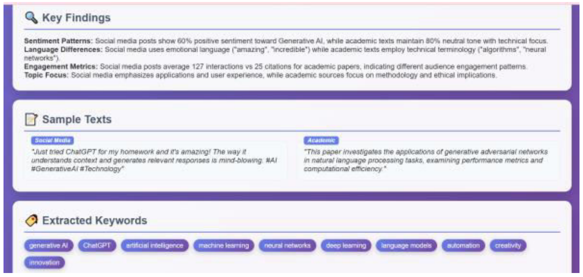


Fig 6.2



Sentiment Patterns: Social media posts show 60% positive sentiment toward Generative AI, while academic texts maintain 80% neutral tone with technical focus.
Language Differences: Social media uses emotional language ("amazing", "incredible") while academic texts employ technical terminology ("algorithms", "neural networks").
Engagement Metrics: Social media posts average 127 interactions vs 25 citations for academic papers, indicating different audience engagement patterns.
Topic Focus: Social media emphasizes applications and user experience, while academic sources focus on methodology and ethical implications.

"Just tried ChatGPT for my homework and it's amazing! The way it understands context and generates relevant responses is mind-blowing. #AI #GenerativeAI #Technology"

"This paper investigates the applications of generative adversarial networks in natural language processing tasks, examining performance metrics and computational efficiency."

Fig 6.3

VII.CONCLUSION

This study presents a comprehensive framework for analyzing the discourse surrounding Generative AI across both social media and academic literature through advanced text mining techniques. By integrating deep learning methods such as Generative Adversarial Networks (GANs) for feature extraction and Random Forest classifiers for text classification, the system effectively captures nuanced insights from large volumes of unstructured data. The proposed approach goes beyond traditional sentiment analysis by combining public perceptions from platforms like Twitter with formal academic trends from research publications. This cross-platform, data-driven architecture enables a more holistic understanding of how generative AI is perceived, debated, and evolving in both informal and formal domains. The results obtained demonstrate the system's ability to identify thematic patterns, track sentiment polarity, and map conceptual shifts, offering valuable insights to researchers, policymakers, and developers interested in the societal footprint of generative technologies.

VIII.FUTURE SCOPE

While the current system architecture provides a solid foundation for analyzing generative AI discourse, there are several avenues for future enhancement. One significant direction is the inclusion of multilingual support, allowing for sentiment and topic detection across non-English tweets and publications. Additionally, the integration of real-time streaming data analysis could make the framework suitable for monitoring breaking trends or sudden shifts in public opinion. Incorporating temporal modeling techniques can also help track the evolution of narratives over time. Furthermore, expanding the academic source base to include conference proceedings, patents, and institutional white papers could deepen the scholarly insight. Future work may also integrate explainable AI (XAI) techniques to interpret model predictions, ensuring transparency and trust in the results. Lastly, this system can be adapted to other domains—such as misinformation detection, emerging technology trends, and crisis communication—to broaden its real-world applicability.

IX.REFERENCES

- Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet Allocation. Journal of Machine Learning Research, 3, 993–1022.
- Feldman, R., & Sanger, J. (2007). The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press.
- Zhang, H., Xu, Y., & Li, J. (2021). Public Sentiment Analysis on AI Technologies Using Twitter Data. IEEE Access, 9, 53767–53779.
- Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?. FAccT '21.
- Bommasani, R., Hudson, D. A., Adeli, E., et al. (2021). On the Opportunities and Risks of Foundation Models. arXiv preprint arXiv:2108.07258.
- Nguyen, T., Jung, J. J., & Nguyen, T. P. (2020). Tracking Misinformation in Social Media Using Topic Modeling and Social Network Analysis. Future Generation Computer Systems, 102, 633–643.
- Caliskan, A., Bryson, J. J., & Narayanan, A. (2017). Semantics Derived Automatically from Language Corpora Contain Human-like Biases. Science, 356(6334), 183–186.
- Klinger, R., & Lars, A. (2019). Analysis of Public vs. Academic AI Narratives Using Sentiment and Topic Mining. AI & Society, 34(4), 853–869.
- Li, Y., & Tang, X. (2022). A Bibliometric Analysis of Generative AI Research. Scientometrics, 127(4), 2105–2123.
- Devlin, J., Chang, M. W., Lee, K., &

Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT.