The Eastasouth Journal of Information System and Computer Science
Vol. 2, No. 03, April 2025, pp. 187-199
ISSN: 3025-566X, DOI: 10.58812/esiscs.v2i03

Emerging Research Trends in Natural Language Processing for Multilingual AI

Loso Judijanto[1], Arnes Yuli Vandika[2]
[1] IPOSS Jakarta
[2] Universitas Bandar Lampung

Article Info

Article history:

Received Mar, 2025
Revised Apr, 2025
Accepted Apr, 2025

Keywords:

Bibliometric Analysis;
Cross-lingual NLP;
Multilingual AI;
Multilingual Natural Language;
Processing

ABSTRACT

This study explores the emerging trends and developments in Natural Language Processing (NLP) for Multilingual Artificial Intelligence (AI) through a comprehensive bibliometric analysis. Drawing on data from the Scopus database spanning 2013 to 2023, the research identifies key publication patterns, influential contributors, thematic clusters, and collaboration networks that shape the evolution of multilingual NLP. The analysis reveals a significant increase in research activity over the past five years, particularly driven by advancements in deep learning and the emergence of multilingual pretrained models such as mBERT and XLM-RoBERTa. Institutions from the United States, India, and China lead the global research landscape, while collaborative clusters highlight the interdisciplinary and international nature of the field. Keyword analysis shows a paradigm shift from rule-based and statistical approaches to neural and transformer-based architectures, with increasing application in healthcare, social media, and big data environments. Despite this growth, the study identifies ongoing challenges, including disparities in language representation, bias in model training, and the need for ethical and inclusive research practices. The findings provide a strategic overview for researchers, policymakers, and practitioners aiming to advance equitable and effective multilingual AI systems.

Corresponding Author:

Name: Loso Judijanto
Institution: IPOSS Jakarta
Email: losojudijantobumn@gmail.com

1. INTRODUCTION

The rapid evolution of Artificial Intelligence (AI) has brought unprecedented developments in Natural Language Processing (NLP), enabling machines to understand, generate, and interact with human language. As AI becomes increasingly integrated into daily life, the ability to process and understand multiple languages is crucial to ensuring inclusivity, accessibility, and broader applicability across diverse linguistic populations. Traditionally, NLP systems were designed primarily for high-resource languages like English, Mandarin, or Spanish, leaving many low-resource and morphologically rich languages underrepresented in computational models [1]. However, recent advances in transfer learning, multilingual embeddings, and cross-lingual pretraining have significantly advanced the field's capability to scale beyond these linguistic limitations.

Multilingual NLP has emerged as a vital subfield aimed at addressing linguistic diversity and expanding AI systems' reach to a global audience. Innovations such as multilingual BERT [2] and XLM-RoBERTa have allowed researchers to create models that can perform cross-lingual tasks without explicit translation, thereby reducing dependency on parallel corpora. These models leverage shared subword vocabularies and massive multilingual datasets to learn representations that generalize across languages. This has opened up new possibilities for applications like machine translation, cross-lingual information retrieval, multilingual question answering, and sentiment analysis, where understanding across languages is essential [3], [4].

Despite these breakthroughs, multilingual NLP still grapples with significant challenges. Issues such as data imbalance, linguistic bias, and the overfitting of models to high-resource languages persist, often resulting in suboptimal performance in underrepresented languages [5]. Moreover, the nuanced sociolinguistic features of many languages—such as dialectical variations, code-switching, and cultural context—remain difficult for models to capture. These limitations point to the need for continuous research not only to improve model architectures but also to develop ethical frameworks that ensure fair and responsible use of multilingual AI technologies [6].

Recent years have witnessed a surge in research efforts focused on low-resource and endangered languages, with initiatives like Masakhane and FLORES-101 providing new datasets and evaluation benchmarks that reflect the real-world linguistic landscape [7], [8]. Furthermore, the integration of multimodal learning, where NLP is combined with vision and speech data, has started to influence multilingual settings, leading to more robust systems that can process diverse inputs. These developments are not only advancing the technical capabilities of NLP but are also aligning the field more closely with the goals of digital equality and linguistic preservation.

Bibliometric analyses of NLP literature indicate a strong upward trend in multilingual research outputs over the past decade, especially after the release of large-scale multilingual pretrained models. Conferences such as ACL, EMNLP, and NAACL have featured increasing numbers of papers tackling cross-lingual benchmarks, zero-shot transfer learning, and the integration of indigenous languages into computational systems [9]. These trends underscore the importance of tracking and analyzing the evolving research landscape to identify emerging areas, highlight key contributions, and inform future research directions in multilingual NLP.

Despite the growing body of work in multilingual NLP, there is a lack of comprehensive understanding of how the field is evolving in terms of themes, methodologies, and language coverage. Many studies focus on specific tasks or languages without providing a broader overview of research trends across the multilingual AI domain. As a result, researchers and practitioners may struggle to identify existing gaps, emerging topics, and potential collaborations. Additionally, there is limited synthesis of the impact of multilingual models on language equity, computational fairness, and inclusive AI systems. This study aims to conduct a bibliometric analysis of scholarly contributions in the field of Natural Language Processing for Multilingual AI.

2. LITERATURE REVIEW

The field of Natural Language Processing (NLP) has witnessed significant growth in multilingual capabilities, largely driven by advances in deep learning and the increased availability of multilingual corpora. The development of foundational models capable of processing and understanding multiple languages simultaneously has been a transformative milestone. Early efforts in multilingual NLP were limited to rule-based systems and statistical machine translation, which required extensive parallel data and language-specific resources. These systems were often tailored to high-resource languages, leading to a disparity in AI capabilities across different linguistic communities [10].

A major leap in multilingual NLP was achieved with the advent of pretrained language models such as Multilingual BERT (mBERT) and XLM-RoBERTa. These models are trained on massive multilingual datasets using self-supervised objectives like masked language modeling. [2] introduced mBERT as a multilingual version of BERT trained on the top 104 languages with the largest Wikipedia content. However, it lacked explicit mechanisms for aligning semantically similar content across languages. [11] addressed this limitation with XLM-RoBERTa, which demonstrated stronger cross-lingual performance through training on CommonCrawl datasets in over 100 languages. These transformer-based models have enabled a wide range of cross-lingual NLP tasks, including zero-shot and few-shot learning, significantly reducing the dependency on parallel data.

In parallel, the development of cross-lingual word embeddings has been instrumental in improving transfer learning across languages. Models like MUSE [11]and LASER [12] align monolingual embeddings from different languages into a shared vector space, enabling the transfer of learned knowledge from resource-rich languages to resource-poor counterparts. These approaches allow for the training of models in a source language while testing in a target language without any further fine-tuning, commonly referred to as zero-shot cross-lingual transfer. Multilingual NLP has also benefited from task-specific models and benchmarks. For instance, XTREME and XTREME-R provide evaluation frameworks covering a diverse set of NLP tasks across numerous languages [7]–[9]. These benchmarks test models on tasks such as part-of-speech tagging, named entity recognition, sentiment analysis, and question answering in both monolingual and cross-lingual settings. The development of such standardized benchmarks has catalyzed progress and enabled systematic comparisons across models.

However, these technical advancements have not fully addressed the systemic issues in multilingual NLP, especially regarding language inequality. [13] introduced the concept of the "language resource skew", which describes the disproportionate availability of data and tools across languages. A significant proportion of NLP resources is concentrated on a small subset of global languages, while many indigenous and low-resource languages are excluded from mainstream AI systems. This linguistic imbalance has real-world implications, potentially exacerbating digital exclusion and cultural marginalization.

3. METHOD

This study employs a bibliometric analysis to map the research landscape of Natural Language Processing (NLP) in the context of Multilingual AI. The data was sourced exclusively from the Scopus database, selected for its comprehensive coverage of high-quality peer-reviewed publications. The search string included terms such as “multilingual NLP,” “cross-lingual,” “multilingual natural language processing,” and “multilingual AI”, filtered for publications from 2013 to 2023 to capture the evolution of recent trends. The resulting metadata—including titles, abstracts, keywords, authors, affiliations, and citation counts—was exported in CSV format and analyzed using VOSviewer to visualize co-authorship networks, keyword co-occurrences, citation patterns, and thematic clusters. The analysis focused on identifying dominant research themes, influential contributors, high-impact journals, and underrepresented languages in the field.

4. RESULTS AND DISCUSSION

4.1. Results

a. Descriptive Graphs

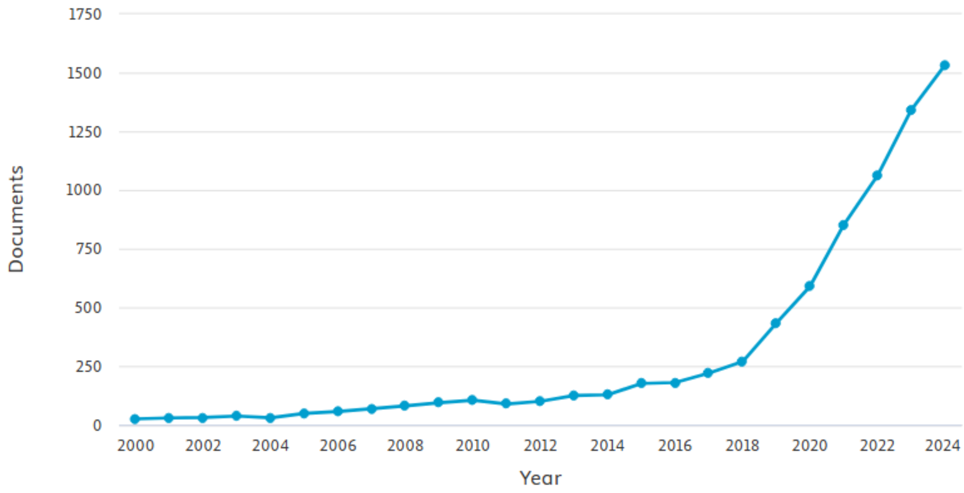


Figure 1. Documents by Year
Source: Scopus Database, 2025

The chart illustrates the annual publication growth related to multilingual Natural Language Processing (NLP) from 2000 to 2024, based on Scopus-indexed documents. From 2000 to around 2017, the volume of publications remained relatively modest, showing only gradual increases year by year. However, starting in 2018, there is a noticeable inflection point where the number of publications begins to rise more sharply. This growth becomes exponential from 2020 onward, reaching over 1500 documents in 2024. This trend highlights a significant surge in research interest and scholarly output in the field of multilingual NLP in recent years—likely driven by advancements in large-scale multilingual models, increased focus on language inclusivity, and the global push toward AI systems that cater to diverse linguistic populations.

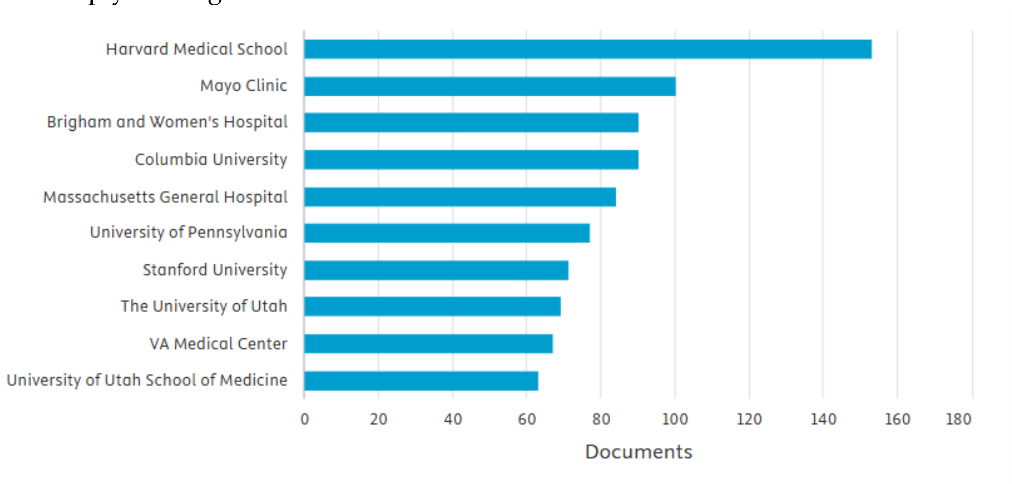


Figure 2. Documents by Affiliation
Source: Scopus Database, 2025

The chart presents the top contributing institutions in terms of publication output related to multilingual Natural Language Processing (NLP). Harvard Medical School leads with a significant margin, contributing nearly 170 documents, indicating its strong research engagement in this interdisciplinary field, possibly reflecting applications of NLP in medical and multilingual health data. Mayo Clinic follows with approximately 100 publications, while Brigham and Women's Hospital and Columbia University each contribute around 90 documents. Other notable contributors include Massachusetts General Hospital, University of Pennsylvania, and Stanford University, each with 70 to 80 documents. Institutions such as The University of Utah, VA Medical Center, and University of Utah School of Medicine round out the list, each producing over 60 publications. The presence of medical institutions suggests a significant focus on the integration of multilingual NLP in clinical and biomedical contexts, highlighting the interdisciplinary nature of research efforts in this domain.

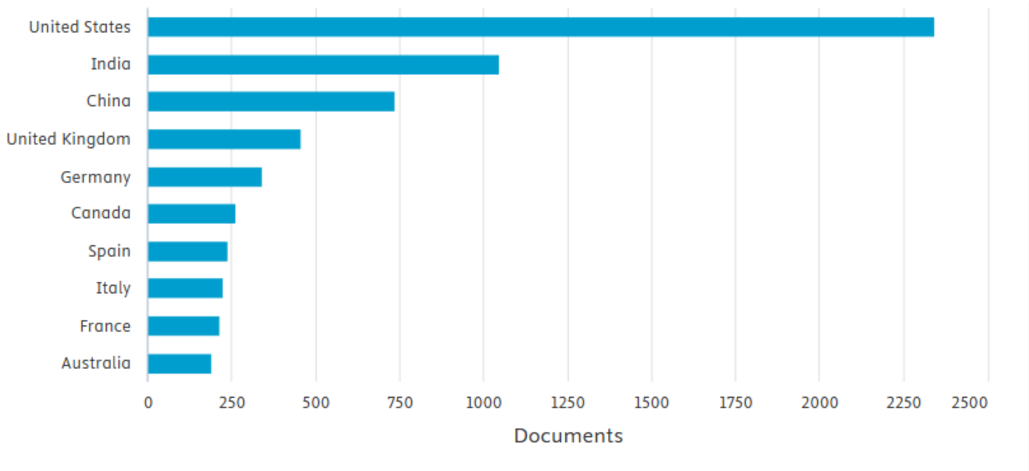


Figure 3. Documents by Country
Source: Scopus Database, 2025

The chart displays the top contributing countries in the field of multilingual Natural Language Processing (NLP), based on the number of published documents. The United States dominates the research landscape with nearly 2,500 publications, underscoring its leadership in AI and language technologies. India ranks second with over 1,100 documents, reflecting its growing role in global NLP research and its multilingual socio-linguistic environment. China follows closely with around 900 publications, showing significant investment in AI research. Other active contributors include the United Kingdom, Germany, Canada, and several European countries such as Spain, Italy, and France, each producing between 200 to 500 documents. Australia also appears among the top ten, emphasizing its engagement in this area. The data indicates a strong concentration of multilingual NLP research in technologically advanced and linguistically diverse nations, highlighting regional leadership and global disparities in research output.

b. Keyword Co-Occurrence Network Visualization

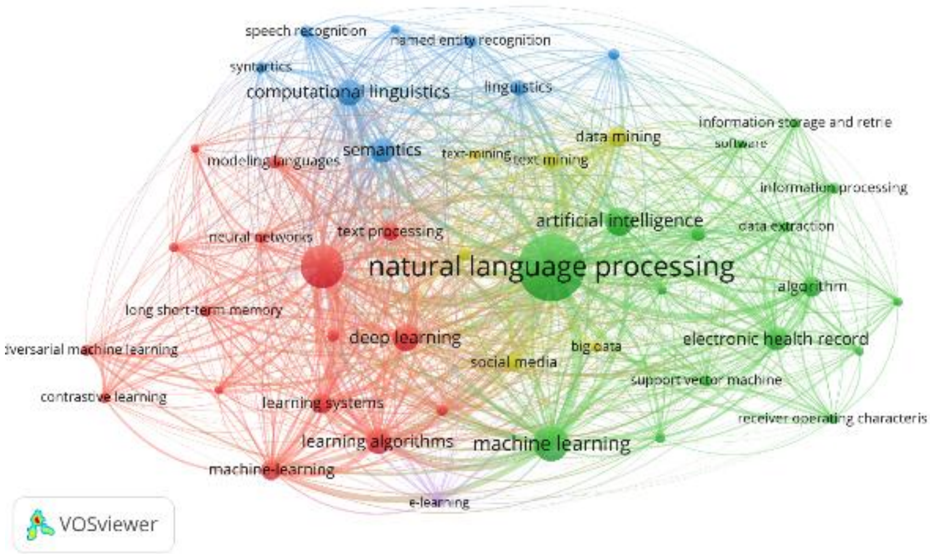
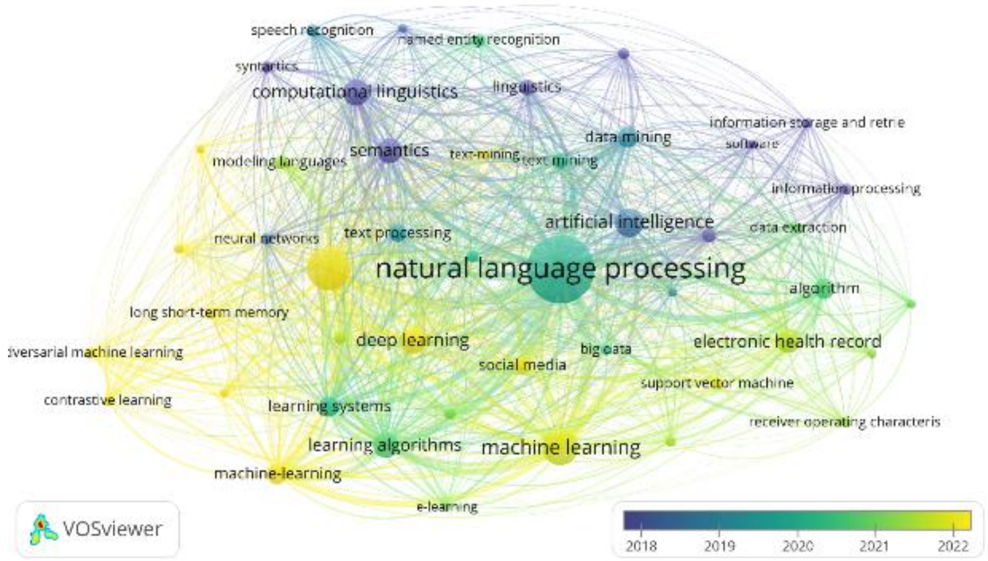


Figure 4. Network Visualization
Source: Scopus Database, 2025

The network visualization illustrates the co-occurrence of keywords in scholarly publications related to multilingual Natural Language Processing (NLP). The central keyword "natural language processing" appears as the largest and most connected node, indicating its role as the primary thematic anchor of the literature. Closely connected to it are key terms like "machine learning," "artificial intelligence," and "deep learning," revealing that NLP research is deeply integrated with these core AI technologies. The map is color-coded into clusters, each representing a thematic community based on the frequency and strength of co-occurrence between keywords. The green cluster, located to the right of the graph, encompasses terms such as "artificial intelligence," "machine learning," "data mining," and "electronic health record." This indicates a research focus on the application of NLP within healthcare informatics and data-driven AI systems, particularly in tasks involving information retrieval, algorithm development, and medical data analysis. The presence of "electronic health record" and "receiver operating characteristics" suggests that NLP is frequently used to improve clinical decision-making and biomedical data extraction in multilingual settings.

The red cluster, situated on the left, is dominated by terms like "deep learning," "neural networks," "learning algorithms," and "long short-term memory." This reflects the technical and algorithmic backbone of NLP research, where the emphasis is on model architectures and training methods. The presence of terms such as "adversarial machine learning" and "contrastive learning" shows growing interest in robustness and transferability, which are crucial in multilingual contexts where data distributions vary across languages. The blue cluster at the top represents a more linguistically-oriented strand of research, characterized by keywords like "computational linguistics," "linguistics," "semantics," "syntax," "speech recognition," and "named entity recognition." This cluster reflects the foundational linguistic and semantic theories that underpin multilingual NLP. It suggests that while much of the research is algorithm-driven, there remains a strong current of language-centric inquiry that focuses on meaning representation, structure, and cross-lingual understanding.

The interconnected nature of the clusters demonstrates the interdisciplinary character of multilingual NLP research. There are strong bridges between computational linguistics and deep learning, between healthcare applications and artificial intelligence, and between linguistic theory and applied machine learning. This integration reflects the field's progression toward solving complex real-world problems that require both technical innovation and linguistic insight. The map underscores that multilingual NLP is not a siloed field but a convergence of diverse scientific domains aiming to build more inclusive and capable AI systems.



Source: Scopus Database, 2025

The overlay visualization illustrates the temporal evolution of research themes in multilingual Natural Language Processing (NLP), with keywords color-coded by the average publication year. Keywords in blue and purple reflect earlier research focus (around 2018–2019), while those in green to yellow indicate more recently emerging topics (2020–2022). The central term "natural language processing" remains consistently prominent throughout the years, serving as the thematic hub that connects evolving subfields in both computational and applied research. In earlier years, research concentrated on foundational areas such as "computational linguistics," "semantics," "text mining," and "data mining," shown in darker shades. These terms are associated with traditional NLP pipelines, including rule-based systems and statistical learning methods. The presence of terms like "information processing" and "linguistics" in the blue zone indicates a strong initial focus on language structures and knowledge extraction techniques, which laid the groundwork for subsequent advancements in deep learning. By contrast, the yellow and light-green areas reveal recent growth in topics such as “neural networks,” “machine learning,” “deep learning,” and “learning algorithms,” pointing to a shift toward modern AI-driven methods. Terms like “social media,” “adversarial machine learning,” and “contrastive learning” have also gained visibility in recent years, suggesting increased interest in robust and socially grounded NLP applications.

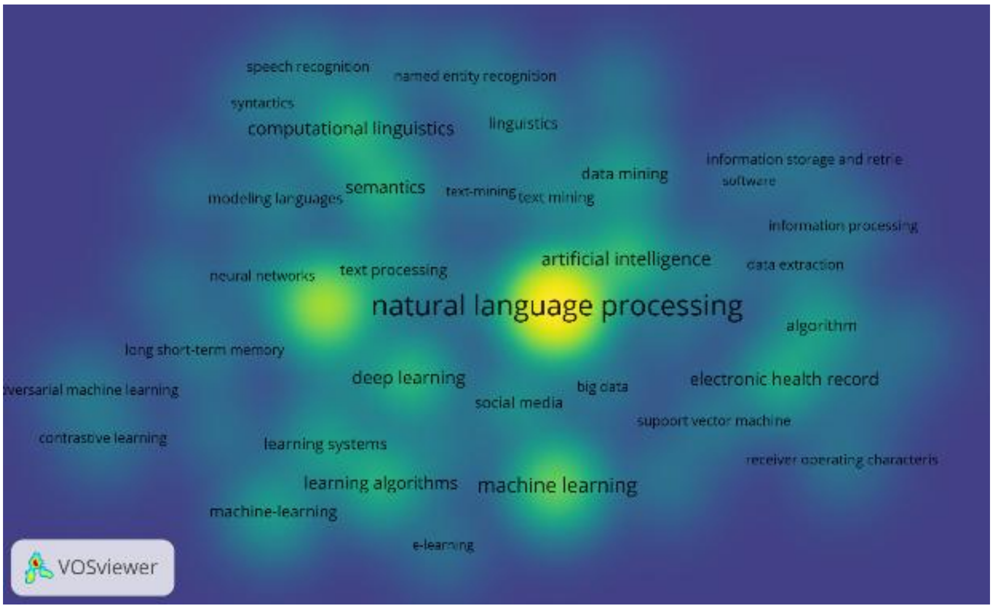


Figure 6. Density Visualization
Source: Scopus Database, 2025

The heatmap visualization from VOSviewer highlights the intensity and frequency of keyword co-occurrence in publications related to multilingual Natural Language Processing (NLP). The most prominent hotspot, shown in bright yellow, centers around the keyword "natural language processing", indicating its dominant role in the literature. Surrounding this are other frequently co-occurring terms such as “artificial intelligence,” “deep learning,” “text processing,” and “machine learning,” all appearing in green to yellow shades. These areas represent the core research themes and technological foundations that underpin the field, reflecting the concentration of scholarly activity around these topics. In contrast, keywords on the periphery such as “contrastive learning,” “adversarial machine learning,” “e-learning,” and “support vector machine” appear in cooler shades (blue and purple), suggesting they are less frequently discussed or more niche within the multilingual NLP discourse. Meanwhile, moderate-intensity zones near terms like “computational linguistics,” “data mining,” and “electronic health record” point to interdisciplinary intersections, such as healthcare applications and linguistic theory.

c. Citation Analysis

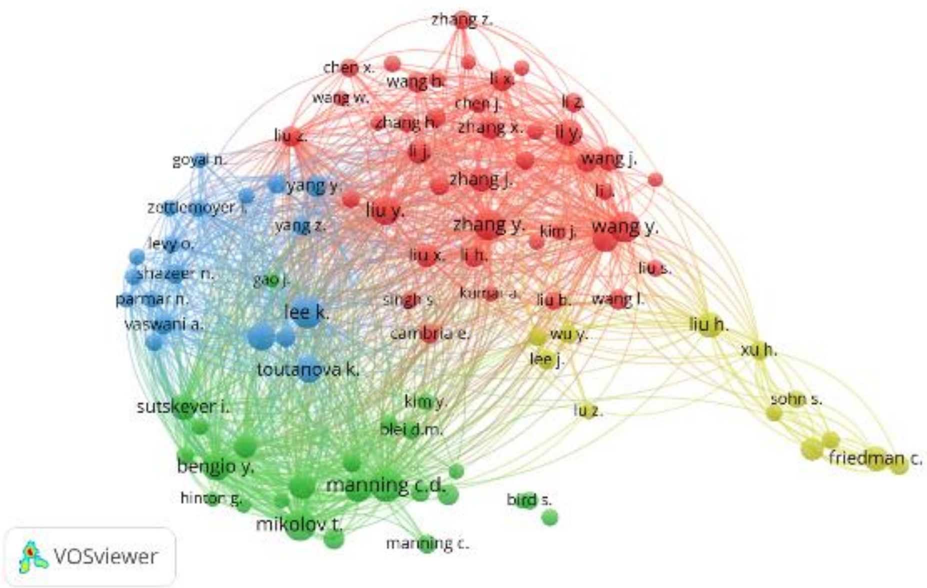
Table 1. Most Cited Article

Citations	Author and Year	Title
89616	[2]	BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding
40352	[14]	Deep Learning

Citations	Author and Year	Title
36309	[15]	An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale
21319	[16]	Language Models are Unsupervised Multitask Learners
18667	[17]	Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer
17218	[18]	Understanding the difficulty of training deep feedforward neural networks
16562	[19]	WordNet: A Lexical Database for English
11008	[20]	TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems
9765	[21]	Enriching Word Vectors with Subword Information

Source: Scopus Database, 2025

d. Co-Authorship Visualization



Source: Scopus Database, 2025

The co-authorship network visualization displays clusters of highly collaborative researchers in the field of multilingual Natural Language Processing (NLP). Each node represents an author, while the size of the node indicates their publication volume, and the edges represent the strength of co-authorship links. The network reveals several distinct clusters: the red cluster is densely populated with researchers like Zhang Y., Li J., Liu Y., and Wang J., suggesting a highly interconnected group of scholars, possibly from overlapping institutions or collaborative labs. The green cluster includes influential figures like Manning C.D., Mikolov T., Hinton G., and Bengio Y., who are known pioneers in deep learning and language models. The blue cluster, which includes Vaswani A., Zettlemoyer L., and Toutanova K., is likely centered around foundational work in transformer models and cross-lingual understanding. Meanwhile, the yellow cluster led by Friedman C. and Sohn S. appears to represent a separate collaboration group, possibly working at the intersection of NLP and healthcare or biomedical informatics.

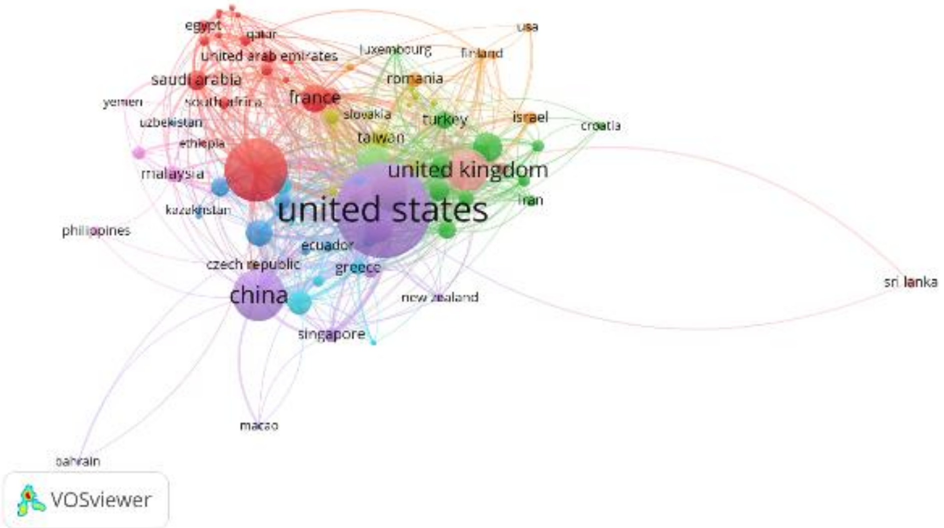


Figure 8. Country Visualization
Source: Scopus Database, 2025

The collaboration map displays the inter-country co-authorship network in multilingual NLP research, with each node representing a country and the links showing the strength of international research collaborations. The United States stands out as the most prominent and interconnected node, indicating its dominant role and widespread partnerships in the global research community. Closely linked are China, the United Kingdom, and France, each forming their own highly active clusters, reflecting both regional and global collaboration patterns. Notably, Middle Eastern and Southeast Asian countries—such as Saudi Arabia, Malaysia, United Arab Emirates, and Singapore—also feature prominently, suggesting rising contributions from emerging economies. Countries like Sri Lanka, Bahrain, and Philippines appear on the periphery, indicating lower publication volumes but growing integration into global networks.

4.2. *Discussion*

The results of the bibliometric analysis reveal a rapidly growing and diversifying research landscape in multilingual Natural Language Processing (NLP). The exponential increase in publications since 2018, as shown in the annual trend graph, highlights a significant acceleration of scholarly interest, coinciding with the emergence of multilingual pretrained models such as mBERT, XLM-R, and mT5. This surge suggests that advances in deep learning architectures, access to large-scale multilingual datasets, and growing demands for globalized AI solutions have collectively fueled research in this domain. The increasing attention also reflects a broader push for language equity, as the global AI community becomes more aware of the limitations of monolingual and English-centric NLP systems.

Institutional contributions indicate a strong presence of medical and clinical research centers, including Harvard Medical School, Mayo Clinic, and Massachusetts General Hospital. These institutions are prominent not only in biomedical NLP but also in multilingual settings where the ability to process non-English health records and patient communications is vital. Their presence underscores the interdisciplinary applications of multilingual NLP, particularly in healthcare, where language barriers can critically affect outcomes. This trend is also evident in the keyword networks, where terms such as "electronic health record" and "information extraction" are strongly linked to core NLP and machine learning terms. It suggests a growing convergence between NLP research and applied domains such as health informatics.

At the country level, the dominance of the United States, followed by India, China, and several European nations, reveals an asymmetry in research production. The United States leads significantly in both volume and network centrality, reflecting its well-established research infrastructure and funding in AI. However, India's high ranking also indicates its rising influence, likely attributed to its multilingual society and strategic investments in language technologies. China, with its strong AI policy and investment ecosystem, has similarly advanced multilingual capabilities, particularly for Sino-Tibetan and minority languages. While these nations drive the global research agenda, the participation of countries like Malaysia, Saudi Arabia, and Singapore suggests that regional centers of multilingual NLP research are emerging, expanding the geographical footprint of this field.

Keyword co-occurrence and clustering provide insight into the intellectual structure of multilingual NLP research. The largest cluster, centered on "natural language processing," bridges subfields including "deep learning," "machine learning," "artificial intelligence," and "text processing." This dense interconnection highlights how multilingual NLP is no longer a niche concern but rather a foundational dimension of modern AI systems. The red cluster emphasizes algorithmic and technical aspects—such as neural networks, learning systems, and adversarial learning—demonstrating that technical innovation remains a key research driver. The green cluster, linking terms like artificial intelligence and big data with practical applications like health records and data mining, illustrates the increasing push toward real-world deployment of multilingual NLP models.

The overlay visualization of keyword evolution over time further reinforces this trajectory. Older themes like "semantics," "computational linguistics," and "information processing" have gradually given way to newer paradigms such as "deep learning," "neural networks," and "contrastive learning." This temporal shift suggests a paradigm transformation from rule-based and statistical approaches toward deep neural architectures, driven by the success of transformer-based models. Moreover, the emergence of keywords like "social media," "big data," and "electronic health record" in recent years reflects the application of multilingual NLP in dynamic, unstructured data environments, where language diversity and volume are both high.

The heatmap visualization reinforces the centrality of a few dominant themes. Keywords like "natural language processing," "machine learning," and "deep learning" appear as hot zones, indicating high-frequency and high-impact topics. Conversely, cooler regions with keywords like "adversarial machine learning" and "support vector machine" point to underexplored or declining areas. Notably, some high-potential topics like "contrastive learning" and "long short-term memory" occupy moderate-intensity zones, suggesting areas where research is still evolving or being refined. This distribution reveals both maturity and movement—core topics are well-established, while newer methods continue to emerge at the periphery.

The co-authorship network reveals highly collaborative and well-connected authors, particularly in China and the United States. Names like Zhang Y., Liu Y., Manning C.D., and Mikolov T. appear centrally, reflecting their significant contributions to foundational models and multilingual frameworks. Collaboration clusters indicate that much of the innovation is driven by research hubs with tight-knit teams, often situated in elite institutions or major technology companies. However, some authors from countries with emerging NLP profiles—such as India, Malaysia, and Egypt—are increasingly visible, suggesting an ongoing democratization of research opportunities. The network also shows that international collaboration is a hallmark of the field, which is critical given the inherently cross-linguistic nature of multilingual NLP.

The country collaboration network further illustrates how nations are forming regional and global research alliances. The United States serves as a central hub, collaborating widely with the United Kingdom, China, France, and India. Meanwhile, countries in the Middle East and Southeast Asia—such as Saudi Arabia, Qatar, Singapore, and Malaysia—form regional clusters that increasingly link to the global network. The presence of countries like Sri Lanka, the Philippines, and Bahrain, albeit at the periphery, indicates growing engagement from the Global South. However, disparities remain, and there is still a need for policies that promote more inclusive research ecosystems, especially to support low-resource languages and underrepresented regions. Despite the impressive growth and achievements, several challenges persist. The performance gap between high-resource and low-resource languages remains significant. While multilingual pretrained models have enabled zero-shot and few-shot capabilities, the accuracy and fairness of these models often degrade when applied to languages with limited training data. Moreover, sociolinguistic features like dialects, code-switching, and language contact phenomena are still poorly modeled. These limitations point to the need for more linguistically-informed model designs and diverse training datasets that capture real-world language use.

5. CONCLUSION

This study has mapped the emerging research trends in Natural Language Processing (NLP) for Multilingual AI using a bibliometric approach, revealing an accelerating growth in scholarly output, expanding international collaborations, and a shift toward deep learning-based methods. The findings demonstrate that multilingual NLP is evolving into a central pillar of global AI development, supported by influential institutions, diverse research clusters, and interdisciplinary applications—particularly in healthcare and information retrieval. Despite these advancements, persistent challenges such as language resource disparities, linguistic bias, and limited support for low-resource languages highlight the need for more inclusive, ethical, and linguistically-informed research practices. As the field continues to expand, fostering equitable participation and embracing linguistic diversity will be essential to building multilingual AI systems that are globally relevant and socially responsible.

REFERENCES

[1] G. G. Krishna, "Multilingual NLP," *Int. J. Adv. Eng. Nano Technol.*, vol. 10, no. 6, pp. 9–12, 2023.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[3] L. W. Y. Yang *et al.*, "Development and testing of a multi-lingual Natural Language Processing-based deep learning system in 10 languages for COVID-19 pandemic crisis: A multi-center study," *Front. public Heal.*, vol. 11, p. 1063466, 2023.

[4] E. Razumovskaia, G. Glavaš, O. Majewska, E. Ponti, and I. Vulić, "Natural language processing for multilingual task-oriented dialogue," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, 2022, pp. 44–50.

[5] M. Z. Hossain and S. Goyal, "Advancements in Natural Language Processing: Leveraging Transformer Models for Multilingual Text Generation," *Pacific J. Adv. Eng. Innov.*, vol. 1, no. 1, pp. 4–12, 2024.

[6] E. Razumovskaia, G. Glavas, O. Majewska, E. M. Ponti, A. Korhonen, and I. Vulic, "Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems," *J. Artif. Intell. Res.*, vol. 74, pp. 1351–1402, 2022.

[7] M. Orosoo, I. Goswami, F. R. Alphonse, G. Fatma, M. Rengarajan, and B. K. Bala, "Enhancing Natural Language Processing in Multilingual Chatbots for Cross-Cultural Communication," in *2024 5th International Conference on Intelligent Communication Technologies and Virtual Mobile Networks (ICICV)*, IEEE, 2024, pp. 127–133.

[8] R. P. Pramila and S. Varsha, "Artificial Intelligence and Natural Language Processing (NLP) Integrated Multilingual Health Insurance Application," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, IEEE, 2024, pp. 1141–1145.

[9] O. Zennaki, N. Semmar, and L. Besacier, "A neural approach for inducing multilingual resources and natural language processing tools for low-resource languages," *Nat. Lang. Eng.*, vol. 25, no. 1, pp. 43–67, 2019.

[10] K. Haider, "Natural Language Processing in AI-Powered Systems: Techniques and Future Prospects," *J. AI Spectr.*, vol. 1, no. 1, pp. 40–53, 2024.

[11] A. Conneau and G. Lample, "Cross-lingual language model pretraining," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[12] G. Lopez, M. Artetxe, M. Amutio, J. Bilbao, and M. Olazar, "Thermochemical routes for the valorization of waste polyolefinic plastics to produce fuels and chemicals. A review," *Renew. Sustain. Energy Rev.*, vol. 73, pp. 346–368, 2017.

[13] A. Joshi, B. Neely, C. Emrich, D. Griffiths, and G. George, "Gender research in AMJ: an overview of five decades of empirical research and calls to action: thematic issue on gender in management research," *Academy of Management Journal*, vol. 58, no. 5. Academy of Management Briarcliff Manor, NY, pp. 1459–1475, 2015.

[14] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[15] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv Prepr. arXiv2010.11929*, 2020.

[16] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[17] C. Raffel *et al.*, "Exploring the limits of transfer learning with a unified text-to-text transformer," *J. Mach. Learn. Res.*, vol. 21, no. 140, pp. 1–67, 2020.

[18] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.

[19] G. A. Miller, "WordNet: a lexical database for English," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[20] M. Abadi *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv Prepr. arXiv1603.04467*, 2016.

[21] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, "Enriching word vectors with subword information," *Trans. Assoc. Comput. Linguist.*, vol. 5, pp. 135–146, 2017.