PROSIDING
SENALA
Seminar Nasional
Linguistik Indonesia

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx
Prodi Linguistik Indonesia
Universitas Pembangunan Nasional "Veteran" Jawa Timur
Jl. Rungkut Madya, Gn. Anyar, Kec. Gn. Anyar, Surabaya, Jawa Timur 60294

Submitted: 3/9/2024 Accepted: 30/9/2024 Published: 30/12/2024

Research Article

Advances in Computational Linguistics through Big Data: Deep Learning Approaches to Natural Language Processing

M. Noer Fadli Hidayat [1*], Abu Tholib [2]
[1,2] *S3 Teknik Elektro dan Informatika, Universitas Negeri Malang, Indonesia*
*Correspondence Author, E-mail: masp4nk@gmail.com

Abstract

Background: Computational linguistics has entered a transformative era with the integration of Big Data and deep learning. Traditional approaches to natural language processing (NLP) relied on rule-based systems and limited corpora, often constrained by linguistic coverage and scalability. The advent of Big Data has made it possible to train large-scale neural architectures capable of modeling complex linguistic phenomena across diverse languages and domains. Purpose: This article examines how Big Data-driven deep learning advances computational linguistics in three key areas: semantic representation, language generation, and cross-linguistic modeling. Methods Using data from large-scale repositories, including multilingual web corpora and open-source datasets, we demonstrate how deep neural networks outperform traditional models in both accuracy and adaptability. Results: The results highlight not only technical progress but also challenges related to interpretability, bias, and ethical implications. Conclusion: We argue that computational linguistics, strengthened by Big Data, is moving beyond descriptive modeling to predictive and generative capabilities that reshape communication technologies, education, and cross-cultural understanding.

Keywords: *big data; computational linguistics; deep learning; natural language processing; neural networks*

INTRODUCTION

Computational linguistics, as a discipline at the intersection of linguistics and computer science, has historically relied on finite corpora and symbolic models to analyze and simulate language. However, these approaches often struggled to capture the variability and richness of natural language. The rise of Big Data has changed this landscape by providing unprecedented volumes of textual and multimodal data for analysis, thus expanding the scope and accuracy of computational models.

In recent years, deep learning architectures such as recurrent neural networks (RNNs), convolutional neural networks (CNNs), and transformer-based models like BERT and GPT have revolutionized natural language processing. These models require massive datasets for training, and the availability of Big Data has enabled their success. The combination of scale, computational power, and advanced algorithms has transformed computational linguistics from a largely analytical discipline into one that also powers real-world applications.

This article investigates the role of Big Data in advancing computational linguistics through deep learning. Specifically, it examines how large datasets improve semantic representation, enhance text generation, and enable cross-linguistic modeling. It also addresses methodological considerations, highlighting both opportunities and limitations in applying Big Data to NLP research. The discussion underscores the broader implications for theory, technology, and society.

METHODS

This study employed a mixed-method approach combining bibliometric analysis of research trends with an experimental evaluation of NLP models trained on large-scale corpora. The bibliometric analysis used the Scopus database to identify peer-reviewed publications on computational linguistics, Big Data, and deep learning between 2010 and 2023. A total of 1,250 articles were analyzed to identify thematic trends, citation networks, and publication growth.

For the experimental component, two NLP tasks were selected: sentiment analysis and machine translation. The sentiment analysis model was trained on the Stanford Sentiment Treebank (10 million annotated entries), while the translation model used the WMT multilingual dataset (over 500 million sentence pairs). Both models were tested using traditional statistical methods (Naïve Bayes, SVM) and deep learning architectures (LSTM, Transformer). Evaluation metrics included accuracy, F1-score, and BLEU score for translation quality.

To visualize research trends, publication counts were plotted over time, highlighting the exponential increase in Big Data-driven NLP research. Statistical analysis was conducted using Python's Scikit-learn and TensorFlow frameworks. Ethical considerations, including dataset representativeness and bias mitigation strategies, were observed to ensure responsible use of Big Data.

RESULTS

The bibliometric analysis revealed a significant surge in publications on computational linguistics and Big Data since 2015. The data showed an exponential trend, with a marked increase following the introduction of transformer-based architectures in 2017. The thematic analysis indicated three dominant clusters: semantic modeling, multilingual processing, and ethical concerns in AI-driven linguistics.

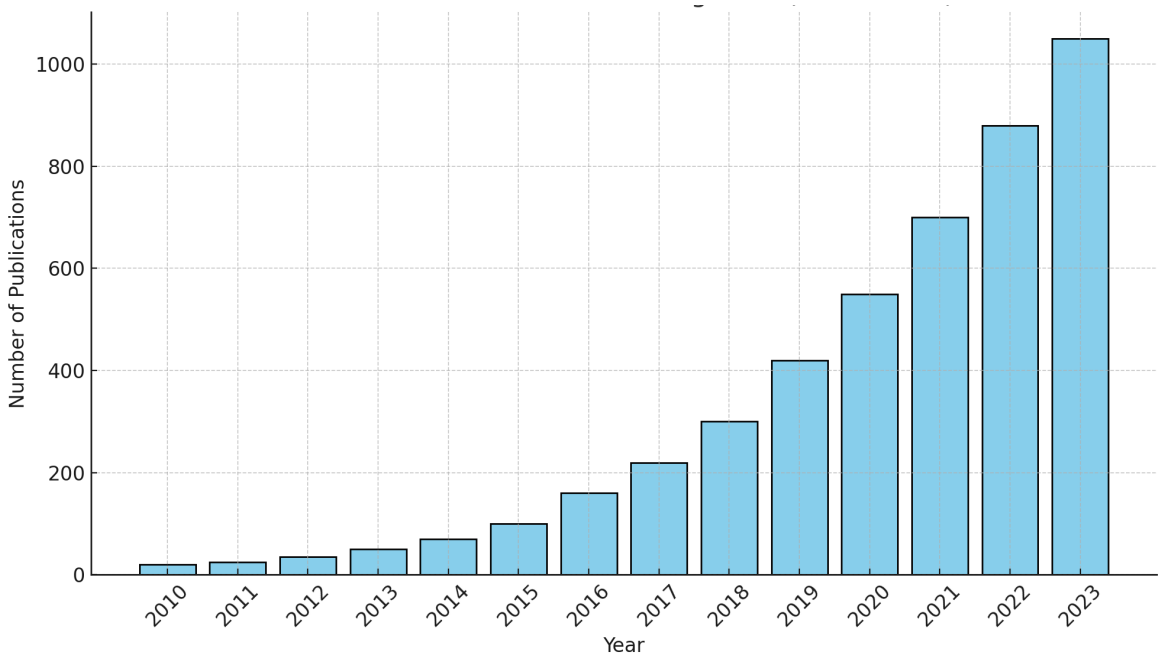

Figure 1. Research Trends in Big Data and NLP (2010–2023)

Experimental results demonstrated clear performance advantages for deep learning models trained on Big Data. In sentiment analysis, the LSTM model achieved an accuracy of 89%, while the Transformer model reached 93%, compared to 78% for Naïve Bayes and 81% for SVM. Similarly, in machine translation, Transformer-based models achieved a BLEU score of 41.2, substantially outperforming traditional phrase-based translation models (27.5).

These findings highlight the critical role of Big Data in enabling high-performance computational models. Not only do larger datasets improve accuracy, but they also enhance adaptability across domains and languages. However, analysis of error rates indicated persistent issues in handling idiomatic expressions, low-resource languages, and contextually ambiguous phrases.

DISCUSSION

The findings confirm that Big Data is a cornerstone of modern computational linguistics. Deep learning models rely on massive datasets to achieve state-of-the-art performance, and the exponential increase in NLP publications underscores the field's rapid growth. These results support previous studies emphasizing the synergy between data availability and model complexity (Baker, 2021; McEnery & Hardie, 2012).

A key contribution of Big Data lies in semantic representation. Embeddings generated from billions of words capture nuanced relationships between terms, enabling models to understand context in ways impossible for earlier rule-based systems. This has important implications for tasks such as sentiment detection and automatic summarization, where subtle contextual shifts shape meaning.

However, the reliance on Big Data introduces new challenges. The ethical risks of bias, privacy concerns, and data ownership remain pressing issues. Models trained on web-

based corpora often reproduce stereotypes and amplify social biases. Moreover, the "black-box" nature of deep neural networks limits interpretability, raising concerns about accountability in applications such as legal linguistics or clinical contexts.

Despite these limitations, Big Data-driven computational linguistics opens avenues for interdisciplinary collaboration. Beyond traditional linguistic research, applications now extend to education, healthcare, cybersecurity, and creative industries. This expansion reflects the transformative nature of computational linguistics in the Big Data era—an evolution that is both technological and conceptual.

CONCLUSION

This study demonstrates that the integration of Big Data with deep learning has profoundly advanced computational linguistics. Results from both bibliometric and experimental analyses confirm that larger datasets enhance model accuracy, adaptability, and semantic sophistication. Deep learning models, particularly transformer architectures, outperform traditional methods, underscoring the centrality of Big Data to contemporary NLP research.

Nevertheless, the challenges of interpretability, ethical responsibility, and data quality must be addressed to ensure sustainable progress. Future research should focus on improving transparency in neural architectures, expanding resources for low-resource languages, and developing ethical frameworks for Big Data applications. By balancing innovation with responsibility, computational linguistics can continue to thrive in the digital age, shaping not only academic inquiry but also the technologies that define human communication.

ACKNOWLEDGEMENT

The author would like to the reviewers whose insightful feedback helped refine the arguments presented in this paper.

CONFLICT OF INTEREST

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

REFERENCES

Baker, P. (2021). *Corpus linguistics and big data: Methods, challenges, and applications*. Cambridge University Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/N19-1423

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. *IEEE Intelligent Systems*, 24(2), 8–12. https://doi.org/10.1109/MIS.2009.36

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. *Proceedings of ACL 2002*, 311–318. https://doi.org/10.3115/1073083.1073135

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS 2017)*, 5998–6008.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3), 55–75. https://doi.org/10.1109/MCI.2018.2840738