PROSIDING
SENALA
Seminar Nasional
Linguistik Indonesia

P-ISSN: xxxx-xxxx
E-ISSN: xxxx-xxxx

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

Research Article

Big Data and the Transformation of Media Literacy: Linguistic Insights into Digital Communication Practices

Farhan [1], Muallim [2]
[1] *Prodi Komunikasi dan Penyiaran Islam, Universitas Nurul Jadid, Indonesia*
[2] *Prodi Pendidikan Bahasa Arab, Universitas Nurul Jadid, Indonesia*
*Correspondence Author, E-mail: farhan@unuja.ac.id

Abstract

Background: The digital revolution has reshaped how media is produced, consumed, and interpreted, presenting new challenges and opportunities for media literacy. In the era of Big Data, vast amounts of linguistic and multimodal data are generated through social media platforms, news portals, and online interactions. Purpose: This study investigates how Big Data transforms media literacy by analyzing digital communication practices from a linguistic perspective. Methods Drawing upon large-scale corpora of online discourse, the research employs computational and discourse-analytic methods to explore patterns of language use, misinformation, and participatory communication. Results: Results reveal how Big Data enhances the capacity to detect narrative framing, identify misinformation, and understand audience engagement. At the same time, ethical concerns regarding privacy, data bias, and algorithmic influence raise critical questions about equitable access to digital knowledge. The findings suggest that media literacy, when supported by Big Data analytics, transcends traditional critical reading skills and evolves into a dynamic competence that integrates linguistic analysis, critical thinking, and digital ethics. Conclusion: This transformation has significant implications for education, public discourse, and democratic participation in an increasingly datafied world.

Keywords: *big data; digital communication; linguistic analysis; media literacy; misinformation*

INTRODUCTION

Media literacy has traditionally been defined as the ability to access, analyze, evaluate, and create media content. However, in the digital era, this concept requires substantial redefinition. The exponential growth of online communication, coupled with algorithm-driven information flows, has fundamentally altered the way individuals encounter and interpret

media (Livingstone, 2014). Traditional literacy skills, while still essential, are insufficient to navigate the complex, data-saturated digital environment.

The role of Big Data in this transformation is crucial. Every click, share, and post generates a digital footprint that contributes to massive datasets of human communication (boyd & Crawford, 2012). For linguists, this provides unprecedented opportunities to examine language use across diverse platforms, registers, and communities. The ability to analyze millions of posts allows researchers to map discursive trends, identify patterns of misinformation, and study how narratives spread in online ecosystems (Howard & Hussain, 2013).

Linguistic insights are particularly valuable in media literacy research because language is the primary medium through which media meaning is constructed. Whether in hashtags, headlines, or memes, linguistic patterns shape public understanding of social issues. Corpus linguistics and computational methods make it possible to uncover hidden ideologies and manipulative practices embedded in digital texts (Baker, 2021). At the same time, discourse analysis provides qualitative depth to interpret how audiences negotiate meaning in participatory online spaces.

Despite these advances, challenges remain. The vastness of Big Data raises questions about representativeness, algorithmic filtering, and ethical use of personal communication data (Hargittai et al., 2020). Furthermore, the prevalence of misinformation and disinformation complicates the ability of individuals to distinguish credible from false content. Media literacy, therefore, must evolve into a broader competence that integrates linguistic analysis, data literacy, and critical digital ethics.

This article explores the transformation of media literacy in the Big Data era, focusing on linguistic insights into digital communication practices. It examines methodological approaches for analyzing large-scale online discourse, presents empirical findings from corpora of social media and digital news, and discusses the implications for media literacy education and democratic engagement.

METHODS

This study employed a mixed-method design combining corpus linguistics, computational analytics, and discourse studies. Data were drawn from two large-scale corpora: (1) a Twitter dataset consisting of 200 million posts related to global news topics between 2018 and 2022, and (2) an online news corpus of 50 million articles from major international outlets over the same period.

Data preprocessing followed standard corpus procedures, including cleaning, tokenization, and removal of spam or non-linguistic entries. Metadata such as hashtags, retweets, and publication dates were preserved to enable the study of narrative spread and temporal dynamics (Pustejovsky & Stubbs, 2012). For the news corpus, headlines and leads were separately coded to examine framing strategies.

The analysis used computational methods including keyword analysis, collocation networks, sentiment analysis, and topic modeling (Grieve, 2021). Visualization tools such

as word clouds and frequency graphs were employed to map recurring linguistic themes. Additionally, discourse analysis was conducted on selected subsets of texts to interpret narrative framing, misinformation strategies, and rhetorical devices (van Dijk, 2013).

Ethical guidelines were strictly followed. Only publicly available posts were analyzed, with anonymization applied to protect user identities. The research adhered to FAIR (Findable, Accessible, Interoperable, Reusable) data principles, ensuring methodological transparency and reproducibility (Wilkinson et al., 2016).

RESULTS

The analysis of the Twitter corpus revealed strong lexical clustering around politically charged terms such as *fake news*, *vaccine*, and *climate crisis*. Hashtags functioned as both organizational and ideological markers, with collocation networks showing how certain terms became discursively linked to misinformation campaigns (Howard & Hussain, 2013).

The news corpus revealed shifts in narrative framing. Collocation analysis showed a decline in neutral terminology (e.g., *global warming*) and an increase in crisis-oriented terms (e.g., *climate emergency*). Keyword analysis highlighted how emotionally charged language correlated with audience engagement metrics, such as article shares and comments (Baker, 2021).

Figure 1. Frequency of “Fake News” Mentions
in Twitter vs News Media (2018–2022)

Year	Twitter (Millions)	News Articles (Thousands)
2018	1.2	4.5
2019	1.6	6.2
2020	3.8	12.0
2021	3.2	10.8
2022	2.9	9.5

Sentiment analysis revealed greater polarization in social media discourse compared to news articles. Twitter posts showed heightened negative sentiment peaks during major crises (COVID-19 outbreak, political elections), while news coverage maintained a more neutral tone. However, analysis also revealed subtle bias in journalistic language, particularly in framing politically sensitive issues.

Topic modeling identified three dominant themes across both corpora: (1) misinformation and conspiracy narratives, (2) climate and health crises, and (3) democratic participation and civic discourse. While news media tended to cluster around institutional authority, Twitter exhibited greater fragmentation and ideological divergence.

DISCUSSION

The results demonstrate that Big Data fundamentally reshapes media literacy by enabling fine-grained analysis of digital discourse. The frequency and collocation analysis confirm that linguistic markers such as hashtags and crisis-oriented terms shape public perception

of key issues (boyd & Crawford, 2012). This validates earlier research that media meaning is co-constructed through both institutional and participatory discourses (Livingstone, 2014).

Importantly, the findings highlight the role of linguistic practices in the spread of misinformation. The term *fake news* illustrates how labels are strategically deployed in political discourse, often as tools to delegitimize opposing viewpoints rather than as descriptors of factual inaccuracy (Howard & Hussain, 2013). Media literacy, therefore, must include the ability to critically interrogate the linguistic strategies behind such labels.

The comparative analysis between Twitter and news media illustrates the complementary yet contrasting functions of each domain. While news outlets maintain editorial norms, they increasingly adopt emotionally resonant language to capture attention. Social media, by contrast, amplifies immediacy and polarization, fostering echo chambers that undermine deliberative discourse (Hargittai et al., 2020).

Big Data analysis also raises ethical and pedagogical implications. From an ethical standpoint, the reliance on user-generated data necessitates robust frameworks for privacy protection and data governance (Wilkinson et al., 2016). Pedagogically, media literacy must evolve beyond evaluating source credibility to encompass understanding algorithmic curation, digital rhetoric, and the linguistic cues that shape meaning in online environments.

Thus, the transformation of media literacy is not merely about adapting to new media formats but about integrating linguistic analysis, computational methods, and ethical reflection. This redefinition equips individuals to navigate digital ecosystems more critically and responsibly.

CONCLUSION

This study shows that Big Data transforms media literacy by enabling the large-scale linguistic analysis of digital communication practices. Findings highlight how linguistic patterns, particularly in hashtags, framing strategies, and sentiment, shape public perception and engagement. The comparative analysis of social media and news corpora reveals both convergence and divergence in how narratives circulate across domains.

Nevertheless, challenges such as misinformation, polarization, and ethical concerns underscore the need for a redefined model of media literacy. Future efforts should prioritize integrating corpus-based linguistic insights with critical pedagogy and digital ethics. Such integration ensures that media literacy evolves as a dynamic, interdisciplinary competence essential for democratic participation in a datafied society.

ACKNOWLEDGEMENT

The author would like to thank the research assistants who supported data collection and preprocessing. Appreciation is also extended to peer reviewers for their constructive feedback that strengthened this article.

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.

boyd, d., & Crawford, K. (2012). Critical questions for Big Data. *Information, Communication & Society*, 15(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Grieve, J. (2021). *Corpus linguistics for online communication: A guide to the study of digital discourse*. Routledge.

Hargittai, E., Füchslin, T., & Schäfer, M. S. (2020). How do young adults engage with science and research on social media? Some preliminary findings and an agenda for future research. *Social Media + Society*, 6(3), 1–10. https://doi.org/10.1177/2056305120940704

Howard, P. N., & Hussain, M. M. (2013). *Democracy's fourth wave? Digital media and the Arab Spring*. Oxford University Press.

Livingstone, S. (2014). Media literacy and the challenge of new information and communication technologies. *The Communication Review*, 7(1), 3–14. https://doi.org/10.1080/10714420490280152

Pustejovsky, J., & Stubbs, A. (2012). *Natural language annotation for machine learning*. O'Reilly Media.

van Dijk, T. A. (2013). *Discourse and knowledge: A sociocognitive approach*. Cambridge University Press.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., … Mons, B. (2016). The FAIR guiding principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. https://doi.org/10.1038/sdata.2016.18