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

Big Data in Forensic Linguistics: Improving Authorship Attribution and Threat Detection in Cyber Contexts

Moh Atikurrahman [1]
[1] *Prodi Sastra Indonesia, UIN Sunan Ampel Surabaya, Indonesia*
*Correspondence Author, E-mail: atiqurrahmann@uinsa.ac.id

Abstract

Background: Forensic linguistics, the study of language in legal and investigative contexts, has gained increasing relevance in the digital era. The proliferation of online communication has created both challenges and opportunities for authorship attribution and threat detection. Purpose: This study explores how Big Data enhances forensic linguistic practices by enabling large-scale analysis of digital texts, such as emails, chat messages, and social media posts. Methods: Using natural language processing (NLP), stylometry, and machine learning techniques, we analyze millions of documents to identify linguistic fingerprints, detect threatening language, and attribute authorship in cybercrime cases. Results: Results demonstrate that Big Data improves accuracy in identifying authorship through stylistic markers and enhances the detection of threats by analyzing lexical, syntactic, and pragmatic patterns. However, ethical concerns—including privacy, consent, and the risk of algorithmic bias—pose significant challenges. This article argues that Big Data-driven forensic linguistics represents a powerful tool for law enforcement and legal proceedings, but its application must be guided by strict ethical frameworks. Conclusion: By combining linguistic theory, computational models, and Big Data analytics, forensic linguistics can significantly contribute to cybercrime prevention and the protection of digital communities.

Keywords: *authorship attribution; big data; cybercrime; forensic linguistics; threat detection*

INTRODUCTION

Forensic linguistics applies linguistic theory and methodology to legal and investigative contexts, focusing on issues such as authorship attribution, disputed meanings, and the interpretation of threatening communications (Coulthard & Johnson, 2017). Traditionally, the field relied on manual analysis of small samples, such as disputed letters or recorded testimony. While effective in some cases, these approaches were limited in scope and struggled to keep pace with the explosion of digital communication.

In the Big Data era, vast amounts of text are generated daily across platforms such as email, messaging apps, and social media. This deluge of data has created new possibilities for forensic linguistics. With computational tools, investigators can now analyze millions of words in seconds, detecting linguistic features and patterns that were previously inaccessible (Grant, 2022). Such capacity is especially critical in cybercrime contexts, where threats and fraudulent activities are often embedded in massive streams of communication.

Authorship attribution has long been a central concern of forensic linguistics. By identifying consistent stylistic markers—such as word frequency, punctuation use, and syntactic preferences—linguists can infer the likely author of a disputed text. Big Data enhances this process by enabling machine learning models to detect subtle linguistic fingerprints across large corpora (Juola, 2006).

Similarly, threat detection requires both linguistic sensitivity and computational scale. Threats may be direct ("I will harm you") or indirect ("Something bad will happen"), often concealed through metaphor or coded language. Big Data tools can analyze such patterns across multiple cases, improving the ability to distinguish genuine threats from non-threatening expressions (Holt & Bossler, 2021).

This article investigates how Big Data reshapes forensic linguistics in the areas of authorship attribution and threat detection. It presents methodological approaches, empirical findings, and ethical considerations, highlighting both the opportunities and challenges of applying Big Data to forensic contexts.

METHODS

This study employed a two-pronged approach focusing on (1) authorship attribution and (2) threat detection. For authorship attribution, a corpus of 1.5 million anonymized emails and online forum posts was compiled. Preprocessing included tokenization, removal of metadata, and feature extraction for stylistic markers such as average sentence length, function word frequency, and punctuation patterns (Stamatatos, 2009). Machine learning models, including Support Vector Machines (SVM) and Random Forests, were trained to classify texts by author.

For threat detection, a dataset of 500,000 police-reported threatening messages (emails, social media posts, and SMS) was analyzed. Each text was annotated by trained linguists according to threat typology (direct, indirect, conditional, hyperbolic). NLP techniques were applied to identify lexical cues, syntactic markers, and discourse strategies typical of threatening communication (Coulthard & Johnson, 2017).

Evaluation metrics included accuracy, precision, recall, and F1-score. Visualization tools such as confusion matrices and feature importance graphs were employed to assess model performance. Ethical safeguards included anonymization of sensitive texts and strict adherence to data protection regulations.

RESULTS

Authorship attribution models demonstrated significant improvements when trained on Big Data corpora. The SVM model achieved an accuracy of 87%, while Random Forests reached 90%, compared to 72% accuracy for traditional frequency-based methods. Function words, punctuation patterns, and collocation clusters emerged as the strongest predictors of authorship.

Threat detection analysis showed that lexical cues such as *kill*, *bomb*, and *attack* were strong indicators of direct threats, while discourse markers like *if you don't* and *unless you* flagged conditional threats. The machine learning classifier achieved an F1-score of 0.89 in distinguishing threats from non-threatening texts, surpassing earlier rule-based approaches.

Table 1. Performance of Authorship Attribution and Threat Detection Models

Task	Traditional Methods	Big Data + Machine Learning
Authorship Attribution	72% accuracy	87–90% accuracy
Threat Detection	68% accuracy	88–89% F1-score

Visualization of feature importance revealed that stylistic markers such as pronoun frequency and use of contractions were particularly influential in authorship attribution. In threat detection, the combination of lexical and syntactic cues provided higher predictive power than either alone.

The analysis also revealed challenges. For authorship attribution, accuracy dropped for very short texts (under 50 words), suggesting limitations in sparse data. For threat detection, the system occasionally flagged figurative language or satire as threatening, underscoring the importance of human oversight.

DISCUSSION

The findings confirm that Big Data significantly enhances forensic linguistics in cyber contexts. Authorship attribution benefits from larger datasets that reveal consistent stylistic markers across diverse texts (Juola, 2006). Similarly, threat detection improves when machine learning models are trained on large, annotated corpora, enabling more reliable identification of dangerous communications (Grant, 2022).

However, the results also highlight persistent limitations. Short texts remain difficult to attribute accurately, as they provide fewer stylistic features. This is particularly relevant in social media, where threats are often conveyed in brief messages. Similarly, the system's difficulty in distinguishing satire or metaphor from genuine threats underscores the need for hybrid approaches that combine computational scale with human expertise (Holt & Bossler, 2021).

Ethical considerations are paramount. The use of personal communication data raises concerns about privacy and consent. Even when texts are anonymized, algorithmic models may reinforce existing biases, disproportionately flagging certain linguistic groups as

suspicious (Barocas & Selbst, 2016). Forensic linguistics in the Big Data era must therefore adopt transparency, fairness, and accountability as guiding principles.

The integration of Big Data into forensic linguistics also has broader implications for legal systems. Automated authorship attribution and threat detection can support law enforcement and judicial processes, but they must not be viewed as infallible evidence. Instead, they should complement linguistic expertise, providing probabilistic support that informs but does not replace human judgment (Coulthard & Johnson, 2017).

In sum, Big Data expands the scope and precision of forensic linguistics, but its effectiveness depends on maintaining methodological rigor and ethical responsibility.

CONCLUSION

This study demonstrates that Big Data significantly improves authorship attribution and threat detection in forensic linguistics. Machine learning models trained on large corpora achieved higher accuracy and reliability than traditional methods, revealing stylistic and lexical patterns with strong predictive value.

At the same time, limitations such as short-text analysis, figurative language, and ethical concerns highlight the need for caution. Future research should focus on hybrid models combining computational efficiency with linguistic expertise, as well as the development of ethical frameworks for forensic applications. With such safeguards, Big Data can enhance forensic linguistics as a vital tool for cybercrime prevention and justice.

ACKNOWLEDGEMENT

The author would like to thank the annotators who contributed to the threat detection corpus and to the reviewers for their constructive feedback.

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

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, 104(3), 671–732. https://doi.org/10.2139/ssrn.2477899

Coulthard, M., & Johnson, A. (2017). *An introduction to forensic linguistics: Language in evidence* (2nd ed.). Routledge.

Grant, T. (2022). *Applied forensic linguistics: Problems and perspectives*. Palgrave Macmillan.

Holt, T. J., & Bossler, A. M. (2021). *The Palgrave handbook of international cybercrime and cyberdeviance*. Springer.

Juola, P. (2006). Authorship attribution. *Foundations and Trends in Information Retrieval*, 1(3), 233–334. https://doi.org/10.1561/1500000005

Stamatatos, E. (2009). A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3), 538–556. https://doi.org/10.1002/asi.21001