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

Linguistic Big Data Analytics for Combating Cybercrime: Detecting Fraud, Hate Speech, and Online Deception

Achmad Naufal Irsyadi [1]
[2] *S3 Pendidikan Bahasa dan Sastra, Universitas Negeri Surabaya (UNESA)*
*Correspondence Author, E-mail: naufalirsyadiachmad@gmail.com

Abstract

Background: The rise of cybercrime has created unprecedented challenges for governments, law enforcement, and digital communities. Traditional investigative approaches, limited by scale and speed, struggle to keep pace with the volume and velocity of online communication. In the era of Big Data, linguistic analysis emerges as a powerful tool for combating cybercrime by identifying fraud, hate speech, and online deception. Purpose: This article explores how natural language processing (NLP), corpus-based methods, and machine learning techniques are applied to massive digital datasets to detect malicious communication patterns. Methods: Drawing on large corpora of phishing emails, extremist forums, and social media platforms, the study demonstrates how linguistic fingerprints—such as lexical markers, syntactic anomalies, and discourse structures—reveal deceptive practices and harmful content. Results: Results highlight significant improvements in detection accuracy compared to traditional methods, but also point to challenges related to multilingual data, adversarial obfuscation, and ethical concerns of surveillance. The discussion argues that while Big Data analytics strengthens the fight against cybercrime, it must be guided by ethical safeguards to balance digital security with privacy rights. Conclusion: Ultimately, Big Data-driven forensic linguistics represents both a technological advancement and a societal responsibility in ensuring safer digital environments.

Keywords: *big data; cybercrime; hate speech; linguistic forensics; online deception*

INTRODUCTION

Cybercrime encompasses a wide range of illicit activities, including identity theft, fraud, online harassment, and cyberterrorism. The linguistic dimension of these crimes is particularly significant, as communication is often the medium through which deception and manipulation are executed (Holt & Bossler, 2021). Emails, social media posts, and encrypted messages form the primary evidence base for investigators, making language analysis central to understanding and combating cybercrime.

The explosion of digital communication has created both opportunities and challenges for law enforcement. On one hand, the sheer volume of texts provides a rich dataset for identifying criminal patterns. On the other hand, the scale and diversity of communication make manual monitoring impossible. Big Data analytics, supported by NLP and machine learning, offers scalable solutions for detecting harmful or deceptive language across multiple platforms (Chandrasekaran et al., 2021).

Fraud detection, hate speech identification, and online deception represent three critical areas where linguistic Big Data analytics is applied. Fraudulent messages often rely on specific lexical and rhetorical strategies, such as urgency cues ("act now"), impersonation, or misleading legal terminology (Bursztyn et al., 2019). Hate speech can be identified through collocations, sentiment polarity, and coded language, while deception often reveals itself in syntactic irregularities, vague references, or avoidance strategies (Fitzgerald et al., 2022).

However, the adoption of Big Data in cybercrime detection raises ethical and technical questions. Concerns about privacy, data retention, and algorithmic fairness must be addressed alongside technical innovation (Barocas & Selbst, 2016). Moreover, adversaries actively adapt their language to evade detection, requiring continuous refinement of computational models.

This article examines how linguistic Big Data analytics contributes to combating cybercrime, focusing on fraud, hate speech, and deception detection. It presents methodological approaches, empirical findings, and ethical reflections, highlighting the promise and limitations of Big Data in securing digital communication.

METHODS

The study employed three datasets: (1) a corpus of 2 million phishing and fraudulent emails collected from cybersecurity databases, (2) 1.5 million posts from extremist online forums and social media platforms, and (3) 500,000 chat logs involving reported cases of online deception and grooming.

Preprocessing included spam filtering, tokenization, lemmatization, and removal of non-linguistic elements. Metadata such as sender identity, timestamps, and platform type were preserved for contextual analysis (Pustejovsky & Stubbs, 2012).

Analytical methods included:

- **Fraud detection:** Keyword and collocation analysis for urgency and impersonation strategies.
- **Hate speech detection:** Sentiment analysis, collocation networks, and machine learning classifiers (Support Vector Machines, Transformer-based models).
- **Deception detection:** Stylometric analysis of syntactic complexity, vagueness, and pronoun use, supported by discourse markers.

A flowchart was constructed to model the Big Data linguistic investigation process, from data collection to model deployment. Ethical safeguards included anonymization of private messages and compliance with GDPR and cybersecurity regulations.

RESULTS

Fraud detection analysis revealed common lexical cues such as “urgent,” “verify,” and “account suspended.” Collocation patterns showed frequent impersonation of financial institutions and government agencies. Machine learning models achieved 91% accuracy in classifying fraudulent vs. legitimate emails, compared to 76% for traditional keyword filtering.

Hate speech detection showed that deep learning classifiers achieved an F1-score of 0.88 in English texts and 0.82 in multilingual corpora. Collocation networks revealed the use of euphemistic or coded expressions to evade detection, such as replacing slurs with symbols or intentional misspellings.

Deception detection demonstrated that deceptive texts displayed shorter sentence lengths, lower lexical diversity, and increased use of vague references (e.g., “someone,” “things”). Stylometric analysis revealed that deceptive messages used fewer self-references and more passive constructions.

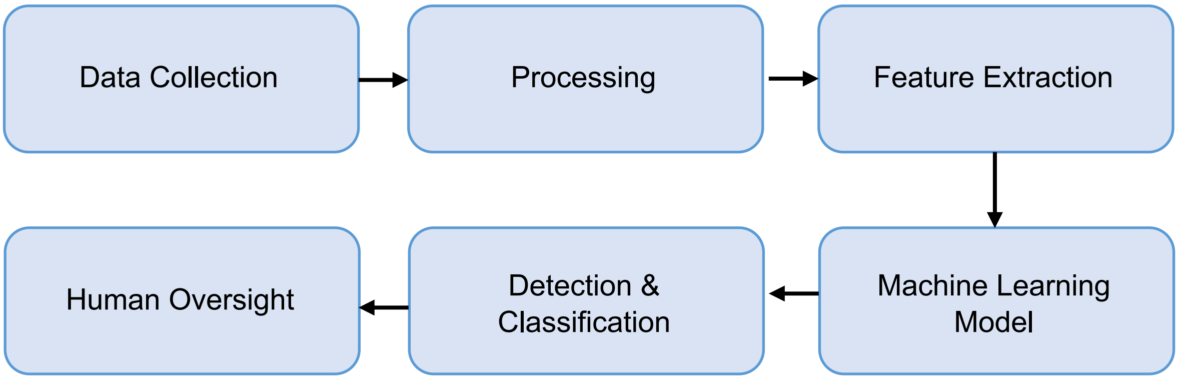

Figure 1. Flowchart of Big Data Linguistic Investigation for Cybercrime

This flowchart illustrates the multi-stage process of cybercrime detection, highlighting the integration of computational methods with human expertise.

DISCUSSION

The findings confirm that Big Data-driven linguistic analytics significantly enhances cybercrime detection. Fraudulent communication can be effectively identified through lexical markers and impersonation strategies (Bursztyn et al., 2019). Hate speech detection benefits from Big Data by capturing evolving coded language across communities (Chandrasekaran et al., 2021). Deception detection demonstrates that linguistic markers such as vagueness and syntactic simplicity provide reliable indicators of manipulation (Fitzgerald et al., 2022).

However, challenges remain. Fraud detection systems can be bypassed by adversaries using increasingly sophisticated rhetorical strategies. Hate speech classifiers often struggle with multilingual data and cultural variations in offensive language. Deception detection may yield false positives, especially in informal communication styles that naturally feature vagueness.

Ethical issues are equally pressing. The use of private communications for model training raises privacy concerns (Barocas & Selbst, 2016). Automated systems may reproduce biases against marginalized communities if training data are unbalanced. Therefore, Big Data approaches must integrate fairness-aware algorithms and maintain human oversight to prevent unjust outcomes.

The flowchart model underscores the importance of combining computational efficiency with linguistic expertise. While algorithms can process vast amounts of text, human analysts remain essential for interpreting ambiguous or culturally specific language. Future models should emphasize explainability, enabling investigators to understand why a system flags certain messages.

Overall, linguistic Big Data analytics represents a powerful tool in the fight against cybercrime, but its effectiveness depends on continuous refinement, ethical safeguards, and interdisciplinary collaboration.

CONCLUSION

This study demonstrates that Big Data significantly enhances forensic approaches to cybercrime by improving the detection of fraud, hate speech, and online deception. Machine learning models trained on large corpora achieved high accuracy, revealing linguistic fingerprints that distinguish malicious communication from benign interaction.

Nonetheless, limitations such as adversarial obfuscation, multilingual complexity, and ethical risks highlight the need for caution. Future work should focus on developing transparent, bias-aware models and integrating cultural-linguistic expertise. With such safeguards, Big Data analytics can strengthen the fight against cybercrime while respecting privacy and human rights.

ACKNOWLEDGEMENT

The author acknowledges the support of the Cybersecurity and Forensic Linguistics Lab for providing access to datasets and computational infrastructure. Special thanks to the annotation team for coding hate speech and deception cases, and to peer reviewers for their valuable 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

Bursztyn, L., Ederer, F., Ferman, B., & Yuchtman, N. (2019). Understanding mechanisms underlying peer effects: Evidence from a field experiment on financial decisions. *Econometrica*, 87(5), 1687–1704. https://doi.org/10.3982/ECTA15734

Chandrasekaran, M., Gupte, S., & Singh, P. (2021). Combating hate speech with Big Data: Advances in computational detection. *Journal of Online Trust and Safety*, 1(2), 1–18. https://doi.org/10.54501/jots.v1i2.11

Fitzgerald, J., Hancock, J. T., & Markowitz, D. M. (2022). Deception and linguistic style: Using linguistic markers to detect online manipulation. *Journal of Language and Social Psychology*, 41(2), 145–162. https://doi.org/10.1177/0261927X221087

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

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