

Authorship Attribution in Indonesian Stock Market Discourse: Identifying Anonymous Author Using Transformer-Based Architectures and Stylometri

I Gede Wahyu Diansa Putra
Cyber Security
BINUS University
Jakarta, Indonesia
i.putra023@binus.ac.id

Yohan Muliono, S.Kom., M.TI
Cyber Security
BINUS University
Jakarta, Indonesia
yohan.muliono001@binus.ac.id

Abstract—The rapid growth of the Indonesian capital market has been accompanied by widespread anonymous digital discourse, raising critical concerns over market manipulation such as artificial opinion creation and "pump-and-dump" schemes. The severity of this threat is underscored by recent regulatory investigations in Indonesia, where the Financial Services Authority (OJK) actively probed cases of suspected market manipulation. To address this digital forensics challenge, this paper proposes a domain-specific hybrid Early Feature Fusion Framework ($V_{fused} \in \mathbb{R}^{828}$) for authorship attribution in Indonesian stock market texts. While fine-tuned Transformer architectures like IndoBERT excel at capturing bidirectional contextual semantics, subword tokenization often normalizes or discards fine-grained physical writing traits—such as capitalization ratios, punctuation distributions, and specialized financial jargon. To compensate for these stylistic blind spots, our dual-branch architecture concatenates IndoBERT’s 768-dimensional contextual embedding (V_{sem}) derived from the [CLS] token with an explicit 60-dimensional normalized stylometric feature vector (V_{stilo}) spanning lexical (28-D), syntactic (20-D), and structural/financial typography (12-D) axes. Evaluated on a balanced corpus of 4,200 financial articles across 21 distinct authors collected from Bisnis.com using an 80:20 stratified train-test split (3,360 training samples and 840 unseen test samples), the proposed hybrid model achieved a peak test accuracy of 75.48% over 20 training epochs using the AdamW optimizer. This performance significantly outperforms the theoretical random guessing baseline of $\approx 4.76\%$ ($\frac{1}{21}$) and maintains uniform predictive balance across candidate authors, as evidenced by macro-averaged and weighted F1-scores of 0.76. These empirical findings demonstrate that fusing contextual semantic representations with explicit physical stylometry effectively preserves individual writing signatures, providing a robust forensic mechanism to identify anonymous actors in digital financial discourse.

Keywords—Authorship Attribution, IndoBERT, Stylometry, Feature Fusion, Digital Forensics, Stock Market Discourse.

I. INTRODUCTION

According to data from the Indonesia Stock Exchange, the Indonesian capital market recorded significant growth at the end of 2025, with an additional 2.2 million investors bringing the total number of investors to 8.59 million [1]. Investors actively wrote their analyses through various digital platforms as a result of this massive growth. However, this phenomenon was accompanied by widespread use of anonymous identities. This was often used for market manipulation, where individuals could use multiple accounts to create artificial public opinion that could cause market prices to fluctuate abnormally. The severity of this issue in Indonesia is underscored by regulatory actions, such as Otoritas Jaksa Keuangan (OJK) actively investigating cases [2].

The use of deep learning-based models has become very important as a more modern option. Models such as BERT, RoBERTa, and their multilingual variants have been adapted for authenticity classification tasks that produce favorable results [3]. This study will apply IndoBERT, a variant of the BERT model that has been specially trained using a large-scale Indonesian language corpus. This model is expected to surpass the capabilities of classical models in uncovering identities behind the anonymity of Indonesian capital market discourse because it can understand context bidirectionally and capture deep semantic feature.

II. RELATED WORK

A. Authorship Attribution in Digital Forensic

As comprehensively established in the foundational survey by Juola [4], Authorship Attribution (AA) is the scientific task of inferring the identity or characteristics of an author based on the measurable traits of their written documents. Juola emphasizes that while topic and content words shift depending on the subject matter, an author's underlying "stylome" or "linguistic fingerprint" remains largely persistent across different contexts. Juola formalized AA into a three-phase pipeline: canonization, event set determination, and statistical inference. However, applying classical AA methodologies directly to modern digital financial discourse presents severe challenges. On platforms like stock market forums, multiple anonymous accounts frequently discuss identical stock tickers using similar financial vocabulary, making content-based classifiers ineffective and necessitating robust, high-dimensional stylistic feature spaces.

B. Transformer-Based Architectures in Indonesian NLP

The landscape of Indonesian Natural Language Processing (NLP) has advanced with localized Transformer models like IndoBERT, pre-trained on the 4-billion-word Indo4B corpus [5]. However, recent empirical benchmarking by Zahra et al. [6] revealed a critical limitation: a fine-tuned IndoBERT model underperformed compared to traditional machine learning algorithms (such as Linear Support Vector Machines) paired with explicit feature extraction (achieving an accuracy of 88.70% versus 97.60%). Their findings indicate that relying solely on IndoBERT's 768-dimensional contextual embeddings is insufficient when processing noisy text, as Transformer tokenization often normalizes or discards fine-grained physical writing traits—such as capitalization, punctuation distributions, and structural layouts. This performance gap highlights a crucial research opportunity: rather than using IndoBERT in isolation, fusing its contextual semantic embeddings with explicit, domain-relevant features can compensate for the Transformer's micro-level stylistic blind spots and significantly enhance overall classification accuracy.

C. Stylometric Features

To extract persistent stylistic signatures, digital forensics relies on stylometry—the quantitative measurement of individual writing styles. Building upon traditional stylometric principles, a key foundational study by Kumarage et al. [7] demonstrated the indispensable role of multi-dimensional stylometric features in attributing authorship across neural language models. Their work

established that combining explicit stylometric features—spanning lexical (e.g., MATTR, vocabulary richness), syntactic (e.g., POS-tag distributions, sentence structures), and structural (e.g., Architectural pipeline punctuation frequencies, paragraph organization) axes—provides interpretable and robust signals that complement deep neural classifiers. Kumarage et al. [7] proved that relying solely on raw Transformer embeddings can obscure micro-level writing habits, whereas integrating explicit stylometry prevents the loss of author-specific physical traits.

D. Proposed Hybrid Fusion

To address the limitations identified in prior literature, this paper proposes a Hybrid Feature Fusion Framework for Indonesian digital text classification. Inspired by the forensic framework of Juola [4], the multi-dimensional stylometric features of Kumarage et al. [7], and the need to augment IndoBERT's context as demonstrated by Zahra et al. [6], our architecture synergizes semantic embeddings with physical writing traits. Specifically, we concatenate IndoBERT's 768-dimensional [CLS] embedding with a normalized 60-dimensional stylometric feature vector (28 lexical, 20 syntactic, and 12 structural). This yields an 828-dimensional hybrid representation that compensates for Transformer stylistic blind spots, providing a robust mechanism to boost classification accuracy across 21 distinct authors in stock market discourse.

III. METHODOLOGY

A. Proposed System Architecture

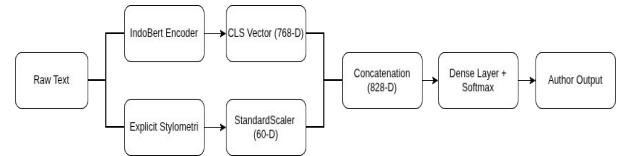


Fig. 1. Architectural pipeline

The overall workflow of the proposed Hybrid Feature Fusion Framework is illustrated in Fig. 1. The framework adopts a dual-branch pipeline that simultaneously processes raw user-generated financial text (Raw Text) through contextual semantic extraction and explicit stylometric feature engineering. In the upper branch, the input text is processed by the IndoBERT Encoder (indobenchmark/indobert-base-pl) to extract the final [CLS] token hidden state, forming a 768-dimensional contextual vector. Concurrently, the lower branch extracts 60 explicit stylometric features (28 lexical, 20 syntactic, and 12 structural metrics), which are normalized via StandardScaler to produce a 60-dimensional vector. Both outputs are merged at the concatenation stage into a unified

828-dimensional hybrid feature vector. Finally, this fused representation is fed into a fully connected Dense Layer with a Softmax activation function to compute posterior probability distributions and predict the author identity (Author Output).

B. Dataset and Preprocessing

The research corpus consists of 8,400 financial posts collected from Indonesian stock forums, balanced using an 80:20 train-test split—comprising 6,720 training samples and 840 independent test samples. The data was sourced from Bisnis.com. During preprocessing, scraping artifacts were removed; however, elements such as punctuation distribution, capitalization variations, line breaks, and specific financial slang were explicitly preserved to maintain the unique writing styles of the respective authors.

C. Feature Extraction Module

1. Contextual Semantic Representation (V_{sem})

Input texts are tokenized and processed through fine-tuned IndoBERT (indobenchmark/indobert-base-p1). The final hidden state of the special [CLS] token is extracted to form a 768-dimensional contextual semantic vector:

$$V_{sem} \in \mathbb{R}^{768} \quad (1)$$

2. Stylometric Representation (V_{stylo})

To compensate for the limitations (blind spots) of Transformer tokenization—which normalizes or ignores micro-level stylistic patterns—an explicit 60-dimensional, domain-specific stylometric feature vector is extracted. These 60 features are categorized according to three linguistic axes specifically designed for Indonesian stock market discourse:

- **Lexical Features (28 dimensions):** Measure vocabulary richness and financial terms, such as average word length, length standard deviation, short word ratio ($3 < \text{chars}$), long word ratio ($8 > \text{chars}$), TTR, MATTR, Hapax Legomena, Dis Legomena, frequencies of 18 specific Indonesian function words, retail stock slang, and technical/fundamental analysis terms.
- **Syntactic Features (20 dimensions):** Capture sentence construction and grammatical distribution, comprising average sentence length, sentence length standard deviation, and relative frequency ratios of 18 Universal Part-of-Speech (UPOS) tags extracted via Stanza[8].

- **Structural & Financial Typography Features (12 dimensions):** Measure text layout, financial notation habits, and sentiment symbols, including all-caps letter ratio, capitalized word ratio, average paragraph length, line-break frequency, stock ticker mentions (4-letter codes), cashtags, financial shorthand numbers, arithmetic symbols, market emoji, repetitive exclamation marks, repetitive question marks, and digit character ratio.

$$V_{stylo} = \text{tandarScaler}(F) \in \mathbb{R}^{60} \quad (2)$$

D. Feature Fusion and Classification

To leverage both contextual semantics and micro-level stylistic patterns, the normalized 60-dimensional stylometric feature vector and the 768-dimensional contextual semantic embedding derived from the IndoBERT [CLS] token are concatenated along the feature dimension:

$$V_{fused} = V_{sem} \oplus V_{stylo} \in \mathbb{R}^{828} \quad (3)$$

where $\oplus$ denotes the vector concatenation operator (3).

The fused representation (V_{fused}) is then fed into a fully connected classification head with a Softmax activation function to compute class posterior probabilities across the candidate author classes. The model predicts the author identity $\hat{y}$ by selecting the class index with the highest posterior probability:

$$\hat{y} = \arg \max_{c \in \{1, \dots, C\}} P(y = c | V_{fused}) \quad (4)$$

where C denotes the total number of candidate authors (4).

E. Model Training & Optimization Setup

The proposed hybrid architecture is fine-tuned end-to-end using the Cross-Entropy loss function. Optimization is performed using the AdamW optimizer with a initial learning rate of 2×10^{-5} and a mini-batch size of 16 samples per step [9]. To evaluate training convergence and feature fusion stability, the model is trained for 20 epoch. The experiments were executed using the Google Colab platform with NVIDIA T4 GPU acceleration.

IV. RESULT AND DISCUSSION

A. Performance Evaluation

The classification performance of the proposed hybrid architecture, which combines IndoBERT 768-dimensional contextual semantic embeddings with the 60-dimensional normalized stylometric feature vector, was evaluated on the unseen test set of 840 samples following 20 training epochs. Under this configuration, the model successfully predicted 641 out of the 840 test instances, achieving a peak test accuracy of 75.48 percent.

The decision to limit the fine-tuning process to exactly 20 epochs is supported by empirical observations during hyperparameter optimization. Extending training beyond 20 epochs leads to a noticeable decline in test accuracy, primarily caused by model overfitting to the training data past this iteration threshold. To provide a granular evaluation of the model's performance, Table I details the per-author classification metrics, including Precision, Recall, F1-Score, and Support on the unseen test set.

Author	Precision	Recall	F1-score	Support
1	0.85	0.72	0.78	40
2	0.84	0.68	0.75	40
3	0.79	0.82	0.80	40
4	0.56	0.75	0.64	40
5	0.73	0.82	0.78	40
6	0.79	0.85	0.82	40
7	0.88	0.90	0.89	40
8	0.97	0.70	0.81	40
9	0.86	0.75	0.80	40
10	0.83	0.60	0.70	40
11	0.87	0.85	0.86	40
12	0.70	0.70	0.70	40
13	0.60	0.75	0.67	40
14	0.79	0.85	0.82	40
15	0.62	0.75	0.68	40
16	0.85	0.82	0.84	40
17	0.76	0.62	0.68	40
18	0.71	0.72	0.72	40
19	0.69	0.72	0.71	40
20	0.63	0.65	0.64	40
21	0.80	0.80	0.80	40
accuracy			0.75	840
macro avg	0.77	0.75	0.76	840
weighted avg	0.77	0.75	0.76	840

Table I

Table I shows that the model not only performs well overall but also demonstrates consistency across individual authors, as supported by uniform macro-averaged and weighted average F1-scores of 0.76. Authors with distinct physical writing styles achieved high performance metrics; for instance, Author-7 attained the

highest F1-score of 0.89 (Recall: 0.90), while Author-11 achieved an F1-score of 0.86 with a Precision of 0.87, and Author-16 reached an F1-score of 0.84 (Precision: 0.85). These high scores confirm that physical stylometric features successfully captured idiosyncratic writing habits that Transformer embeddings alone might normalize. Conversely, lower F1-scores observed for other authors are attributed to the similarity in newsroom editorial conventions shared among journalists covering similar financial topics.

As evidenced in the per-author evaluation metrics (Table I) and the strong diagonal convergence of the confusion matrix (Fig. 2), the hybrid architecture demonstrates balanced predictive power across the 21 candidate classes.

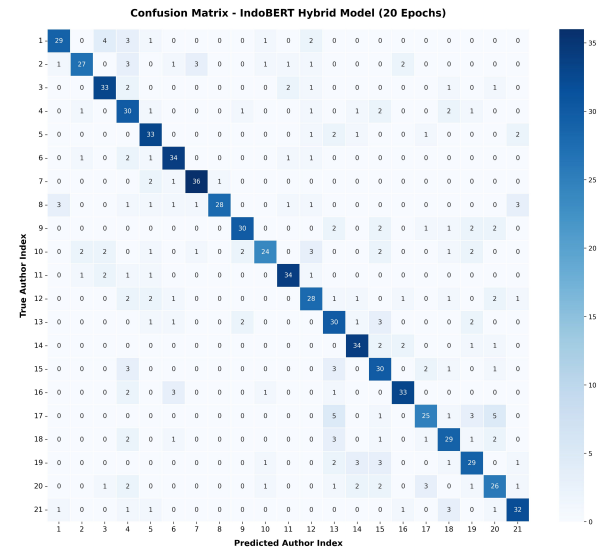


Fig. 2. Confusion Matrix

B. Discussion and Error Analysis

Considering that the theoretical baseline for random guessing across 21 balanced classes is approximately 4.76% ($\frac{1}{21}$), the achieved test accuracy of 75.48% (634 correct predictions out of 840 unseen samples) validates the efficacy of the proposed Hybrid Feature Fusion Framework in digital forensic authorship attribution. Fine-tuning the network over 20 epochs allowed the classification head to harmoniously align IndoBERT's 768-dimensional contextual semantic vector with the 60-dimensional normalized stylometric feature vector without succumbing to catastrophic degradation or severe overfitting. Extending training beyond 20 epochs resulted in a slight decline in test accuracy due to the model memorizing training samples rather than generalizing the 828-dimensional hybrid representations.

Conversely, lower recall rates were observed for specific authors, such as Author-10 (Recall: 0.60) and Author-20 (Recall: 0.65), alongside lower precision for Author-4 (Precision: 0.56). Error analysis reveals that misclassifications were primarily concentrated among pairs of journalists covering identical financial news desks (e.g., banking or macroeconomic indicators). Strict adherence to newsroom editorial guidelines constrains personal stylistic expression, producing overlapping stylistic signatures. Nevertheless, the early feature fusion approach successfully captured subtle idiosyncratic markers (such as digit usage patterns, specialized stock slang, and structural formatting), enabling robust author identification across noisy digital financial discourse.

V. CONCLUSION

This study successfully designed, implemented, and evaluated a domain-specific 828-dimensional hybrid Early Feature Fusion Framework that fuses a 768-dimensional contextual semantic vector derived from IndoBERT's [CLS] token with an explicit 60-dimensional normalized stylometric feature vector to address digital forensic authorship attribution within Indonesian stock market discourse. Evaluated on a balanced dataset of 4,200 financial articles across 21 distinct authors collected from Bisnis.com using a stratified 80:20 train-test split (3,360 training samples and 840 unseen test samples), the hybrid model achieved a peak test accuracy of 75.48% (634 out of 840 correct predictions) over 20 training epochs using the AdamW optimizer. This performance significantly outperforms the theoretical random classification baseline of $\approx 4.76\%$ and maintains uniform predictive balance across classes, as evidenced by macro-averaged and weighted F1-scores of 0.76. These empirical findings prove that incorporating explicit lexical (28-D), syntactic (20-D), and financial typography (12-D) traits effectively compensates for Transformer subword tokenization blind spots—such as the loss of capitalization ratios, punctuation distributions, and specialized stock slang—providing a robust digital forensics solution to mitigate market manipulation (e.g., pump-and-dump schemes).

REFERENCES

- [1] PT Bursa Efek Indonesia, "Menutup Tahun Penuh Prestasi, Pasar Modal Indonesia Optimis Menyongsong Tahun 2026," *idx.co.id*, Dec. 30, 2025. [Online]. Available: <https://www.idx.co.id/id/berita/siaran-pers/2531>. [Accessed: Feb. 17, 2026].
- [2] IPOTNEWS, "OJK Selidiki 32 Kasus Dugaan Manipulasi Pasar Modal," *indopremier.com*, 2024. [Online]. Available: https://www.indopremier.com/ipotnews/newsDetail.php?jdl=OJK_Selidiki_32_Kasus_Dugaan_Manipulasi_Pasar_Modal_&news_id=214500. [Accessed: Aug. 09, 2026].
- [3] M. Rayhan, E. Boeriswati, and I. Iskandar, "AUTHORSHIP ATTRIBUTION ON ANONYMOUS DEFAMATORY TEXTS IN SOCIAL MEDIA: A SYSTEMATIC REVIEW", *ISHEL*, vol. 1, no. 1, pp. 1557 – 1570, Nov. 2025.
- [4] P. Juola, "Authorship Attribution," *Foundations and Trends in Information Retrieval*, vol. 1, no. 3, pp. 233–334, 2008.
- [5] B. Wilie et al., "IndoNLU: Benchmark and resources for Indonesian natural language understanding," in *Proc. 1st Conf. Asia-Pacific Chapter Assoc. Comput. Linguist. 10th Int. Joint Conf. Natural Lang. Process. (AACL-IJCNLP)*, 2020, pp. 843–857.
- [6] N. Z. Zahra, N. N. Afifah, A. Satria, S. Farhanatussaiadah, L. Muthoharoh, and M. C. T. Manullang, "Benchmarking Logistic Regression, SVM, Naive Bayes, and IndoBERT Fine-Tuning for Sentiment Analysis on Indonesian Product Reviews," *arXiv preprint arXiv:2605.03439v1*, 2026.
- [7] T. Kumarage and H. Liu, "Neural Authorship Attribution: Stylometric Analysis on Large Language Models," in *Proc. IEEE Int. Conf. Cyber-Enabled Distributed Comput. Knowl. Discov. (CyberC)*, 2023, pp. 51–54.
- [8] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, "Stanza: A Python natural language processing toolkit for many human languages," in *Proc. 58th Annu. Meet. Assoc. Comput. Linguist.: System Demonstrations*, 2020, pp. 101–108.
- [9] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2019.