EPiC Series in Computing

Volume 111, 2026, Pages 111–122

Proceedings of The 21st Asia Joint Conference on Information Security

Lightweight Authorship Attribution for Japanese Web Reviews: Character n-grams with Collective Attribution toward Actor Analysis

Hiroshi Matsubara[1], Shingo Matsugaya[123], Taichi Aoki[3], and Masaki Hashimoto[1]

[1] Graduate School of Science for Creative Emergence, Kagawa University
hashimoto.masaki@kagawa-u.ac.jp
[2] Trend Micro, Inc.
[3] Japan Cybercrime Control Center

Abstract

We study stylistic-feature-based authorship attribution as a foundational technique for actor analysis in threat intelligence. Using Rakuten Ichiba reviews, we compare four methods—TF-IDF+LR, BERT-Emb+LR, BERT-FT, and Metric+kNN—under unified settings, scaling the number of authors U up to 1,000. In our experiments, the lightweight character n-gram TF-IDF+LR attains accuracy comparable to or higher than the BERT-based methods at $U \geq 20$ while requiring a fraction of their computation time, and retains a Top-10 accuracy of 0.821 at $U = 1{,}000$, useful for candidate retrieval. Applying Collective Attribution (test-time aggregation) further improves accuracy at $U = 1{,}000$ from 0.623 (single review) to **0.985** (ten reviews concatenated). These results suggest that combining lightweight character n-gram features with test-time aggregation is a practical choice for authorship attribution on Japanese short web texts.

Keywords: Authorship attribution; Character n-grams; Collective attribution; Japanese web text; Threat intelligence; Top-k evaluation.

1 Introduction

With the rising sophistication of cyber attacks in recent years, **threat intelligence**—which supports the understanding and tracking of attacker (actor) activities—has become increasingly important. Posts on highly anonymous platforms such as dark web forums can serve as valuable information sources, but tracking the same individual is often difficult due to account anonymization and the use of multiple aliases. To address this challenge, **authorship attribution**, which infers the writer from stylistic features of texts, is expected to serve as a foundational technique that exploits cues independent of content (topic).

In practical post analysis, however, challenging factors such as the brevity of texts, formulaic expressions, and a growing number of authors are involved, so it is necessary to evaluate not only accuracy under specific conditions but also (i) the comparison between lightweight feature-based methods and pre-trained language models, (ii) the scaling behavior with respect to the

T.-T. Tsai, H.K. Kim, Y. Wang and A. Yamada (eds.), AsiaJCIS 2026 (EPiC Series in Computing, vol. 111), pp. 111–122

number of authors, and (iii) computational cost in a cross-cutting manner. Existing studies have proposed approaches based on character n-gram stylistic features (e.g., Stamatatos [1]) and BERT-based approaches (e.g., Fabien et al. [2]), but studies that compare these under unified conditions on Japanese short web texts at large-scale multi-class settings (hundreds to a thousand authors) remain limited. Moreover, the extent to which test-time aggregation of multiple posts can improve identification performance under short-text conditions has not been sufficiently quantified.

In this study, with future application to the dark web in view, we first systematically investigate authorship attribution for Japanese short web texts using a highly reproducible large-scale clear-web review dataset (Rakuten Ichiba reviews). Specifically, we compare four methods—TF-IDF+LR, BERT-Emb+LR, BERT-FT, and Metric+kNN—under unified conditions, and analyze performance and computational cost as the number of authors U is varied incrementally up to 1,000, using Accuracy, Macro-F1, and Top-k Accuracy (which targets candidate retrieval). Furthermore, we introduce **Collective Attribution**, which aggregates multiple reviews from the same author at test time, and quantify the extent to which varying the bundle size b can compensate for the performance degradation under large-scale conditions.

The main contributions of this study are summarized as follows.

- **Systematic comparison of four methods under unified conditions:** We compare TF-IDF+LR, BERT-Emb+LR, BERT-FT, and Metric+kNN on Japanese short web texts under unified preprocessing, data splits, and evaluation metrics, and show that, under our experimental setting, the lightweight TF-IDF+LR achieves accuracy comparable to or higher than the BERT-based methods at medium-to-large scales ($U \geq 20$) with substantially lower computational cost.

- **Characterization of large-scale behavior:** We evaluate performance and computational cost while varying the number of authors U incrementally up to 1,000, characterize the trend of performance degradation with growing U, and confirm the utility of TF-IDF+LR for candidate retrieval (screening) by showing that it retains a Top-10 Accuracy of 0.821 even at $U = 1{,}000$ ($n = 186$).

- **Reinforcement under large-scale conditions via test-time aggregation:** We apply Collective Attribution—a test-time aggregation approach in line with profile-based authorship attribution [3, 4]—to TF-IDF+LR, and show that its concatenation variant improves Accuracy at $U = 1{,}000$ from 0.623 with a single review ($b = 1$) to 0.985 with ten reviews concatenated ($b = 10$).

We note that this study constitutes a foundational investigation conducted on Japanese reviews from the clear web, and does not directly evaluate threat-intelligence-related texts such as those found on the dark web. Nevertheless, since reviews contain difficulty factors common in post analysis—text brevity, formulaic expressions, and topic variation—the findings can serve as a guideline for method selection on texts with similar properties.

2 Related Work

Authorship attribution is a research field that infers the writer of a text from stylistic features, and has evolved from statistical stylometry through machine learning and deep learning to pre-trained language models [5, 1]. In this section, we review authorship attribution based on

Table 1: Comparison between existing studies and this work (○: covered, △: partially covered, ×: not covered)

Study	Japanese	Large-scale	Unified Comparison	Collective
Stylometry studies	×	△	△	×
PLM-based AA	×	△	×	×
Japanese AA	○	×	×	×
Profile/Collective AA	×	×	×	○
This study	○	○	○	○

stylistic features and on pre-trained language models, which are most relevant to this study (Section 2.1), as well as profile-based authorship attribution and Collective Attribution as treated in this study (Section 2.2), and then describe the position of this work. The correspondence between existing studies and this work is summarized in Table 1.

2.1 Authorship Attribution: from Stylometry to Pretrained Models

As a classical reference in authorship attribution, the pioneering work of Mosteller and Wallace [5], which statistically addressed the attribution of the Federalist Papers, is widely recognized, and Stamatatos [1] subsequently organized a comprehensive framework spanning feature design to evaluation design. Among various features, character n-grams have been reported to be relatively less affected by topic dependence and to remain robust under short-text and multilingual conditions [6, 7]. Author identification based on stylistic features has also been studied in Japanese; for example, authorship identification of Japanese texts using random forests has been reported [8]. These findings provide the theoretical background for the character n-gram-based TF-IDF approach adopted in this study.

More recently, authorship attribution using pre-trained language models (PLMs) such as BERT [9] has been widely studied. Representative examples include BertAA [2], which fine-tunes BERT for authorship attribution, and metric learning or contrastive learning approaches that learn an author embedding space [10]. In Japanese as well, studies have been reported that perform authorship attribution using BERT or integrate it with stylistic features [11].

Koppel et al. [12] and Luyckx et al. [13] have pointed out that an increase in the number of candidate authors makes authorship attribution more difficult, and the PAN shared task [14] has continuously addressed extensions to cross-domain and open-set settings. However, most existing studies have remained at the scale of tens to about a hundred authors, and in particular, studies that compare lightweight character n-gram-based methods and PLM-based methods under unified conditions on Japanese short web texts at the scale of hundreds to a thousand authors remain limited.

2.2 Profile-based Authorship Attribution and Test-Time Aggregation

The idea of integrating multiple texts to improve attribution accuracy traces back to early studies based on author profiles [3]. In particular, profile-based verification [4] aims to improve performance by treating multiple documents in an aggregated manner, and methods that combine instance-based and profile-based approaches have also been proposed [15]. The test-time aggregation (Collective Attribution) treated in this study is an approach in line with this lineage, positioned as an operation that, while keeping the trained model fixed, aggregates multiple texts from the same author through concatenation or by averaging predicted probabilities.

However, existing studies have mainly targeted authorship verification or limited-scale settings, and few studies have systematically evaluated the relationship between the number of

aggregated posts and performance under Japanese short web texts and large-scale multi-class conditions. In particular, quantitative guidelines for "how many posts need to be aggregated to reach a practical level" have yet to be organized.

2.3 Positioning of This Work

In summary, while existing studies have individually demonstrated the effectiveness of stylistic-feature-based methods and PLM-based methods, studies that compare multiple methods under unified conditions on Japanese short web texts at large-scale multi-class settings, and further quantify the effect of test-time aggregation, remain limited. Building on these existing studies in a complementary manner, this work conducts a unified comparison of four methods, characterization of scaling behavior, and evaluation of the effect of test-time aggregation, using a highly reproducible large-scale clear-web dataset.

3 Dataset and Methods

3.1 Dataset

The main target of this study is the Rakuten Ichiba dataset [16], which is provided as a large-scale review-posting dataset including author IDs. Rakuten Ichiba is one of the largest e-commerce platforms in Japan on which reviews across a wide range of product categories have been accumulated. The provided data is in TSV format and includes masked author IDs, review bodies, rating points, and registration timestamps. The dataset covers the period from January 2015 to December 2019, and this study primarily analyzes the 2019 reviews. Among the 2019 reviews, the top 100 authors by post count ($U = 100$) yielded a total of 68,222 reviews, while the top 1,000 authors ($U = 1{,}000$) yielded 294,443 reviews. The character length of review bodies has a median of 60 characters and a mean of 118.45 characters, indicating that short texts dominate; this poses a challenging condition under which stylistic features are less evident.

3.2 Preprocessing

To reduce noise in the review bodies and unify model inputs, we applied the following preprocessing steps. (i) Order numbers (bracketed digit strings) and URLs are removed using regular expressions. (ii) Line breaks, tabs, and consecutive whitespace are normalized to a single space, and leading and trailing whitespace is stripped. (iii) Extremely short texts with little information content (10 or fewer characters after preprocessing) are excluded. Only the preprocessed review body (`content`) is used as input; metadata such as `shop_name` and `item_id` are not used because they depend on the product or shop rather than the writing style.

3.3 Compared Methods

We formulate authorship attribution as closed-set multi-class classification, and compare four methods in total: a lightweight stylistic-feature-based method (TF-IDF+LR), and three methods that leverage a pre-trained language model (PLM) at different granularities (BERT-Emb+LR, BERT-FT, and Metric+kNN). The BERT-based methods use the Tohoku University Japanese BERT (`cl-tohoku/bert-base-japanese-v3` [17]) as the backbone, with a maximum input length unified at 128 tokens. Below, we describe the configuration and rationale for each method.

TF-IDF+LR classifies character n-gram TF-IDF features [6] using multi-class logistic regression. In Japanese short texts, subtle orthographic habits—particles, punctuation, sentence-ending expressions, and the choice between kana and kanji—can signal authorship, so we adopt character-level n-grams ($n \in \{2, 3\}$) rather than word-level ones. TF-IDF downweights corpus-frequent expressions, relatively emphasizing author-characteristic orthography. We position this low-cost, reproducible baseline as the reference against which PLM-based methods are examined.

BERT-Emb+LR keeps the pre-trained BERT frozen and extracts a sentence embedding by mean-pooling over the final hidden layer (768 dim, attention-mask weighted), then classifies it with logistic regression. We use mean pooling rather than the [CLS] token because, for representations not adapted to the task, averaging over the sentence tends to be more stable than a single token. This method indicates the performance achievable when a PLM's language understanding is repurposed for authorship attribution without additional training.

BERT-FT fine-tunes all parameters of Japanese BERT and predicts the author through a classification head, following prior work such as BertAA [2]. It re-adapts the general-purpose pre-trained representations toward author-specific stylistic differences. Among the four methods, this most fully leverages the PLM's capability and indicates the performance level originally expected for tasks requiring contextual representations.

Metric+kNN uses Japanese BERT as the backbone and trains a metric space via BatchHardTripletLoss [18] (margin=0.2), pulling same-author embeddings together and pushing different-author ones apart. At inference, the author is decided by k-nearest neighbors ($k = 3$) under cosine distance against the training embeddings [10]. Having no classification layer, it is flexible to changes in author count, and its nearest-neighbor search has natural affinity with candidate retrieval (Top-k). We include it to add a learning paradigm different from the other three.

3.4 Collective Attribution (Test-Time Aggregation)

As a test-time aggregation approach in line with profile-based authorship attribution [3, 4], we use TF-IDF+LR as the base model and keep the trained classifier fixed. Aggregation is applied only at inference time using two schemes: (i) **Concatenation**, where b reviews are concatenated and fed to the classifier as a single input; and (ii) **Probability Averaging**, where b reviews are inferred individually and their predicted probabilities are averaged for the final decision. We evaluate bundle sizes $b \in \{1, 2, 3, 5, 10\}$ under the same setting as Experiment 3, namely $U \in \{100, 200, \ldots, 1000\}$ with $n = 186$ reviews per author. Reviews are shuffled before bundling, and incomplete bundles are discarded.

3.5 Evaluation

We evaluate authorship attribution as closed-set multi-class classification. In addition to Accuracy, we use Macro-F1, which is robust to imbalances in post counts across authors, and Top-k Accuracy ($k \in \{3, 5, 10\}$) for candidate retrieval performance. Results are reported as means over stratified 5-fold cross-validation (StratifiedKFold, shuffle=True, random_state=42); Accuracy and Macro-F1 are reported as mean±standard deviation, and Top-k as the mean. Training and inference times are measured for each fold and used to compare computational cost[1].

[1] All experiments were conducted on a Linux environment (Ubuntu 22.04) equipped with an NVIDIA RTX 5070 Ti (16GB). Main software stack: Python 3.10, PyTorch 2.0.1, Transformers 4.30.0.

Table 2: Experiment 1: Comparison of the four methods ($U = 100$, $n = 400$, 5-fold mean±SD)

Method	Acc	F1	Top-3	Top-5	Top-10	Time(s)
TF-IDF+LR	**0.870**±0.002	**0.870**±0.004	0.943	0.961	0.979	157
BERT-FT	0.861±0.002	0.859±0.003	0.937	0.954	0.970	5,394
BERT-Emb+LR	0.806±0.002	0.806±0.002	0.912	0.943	0.971	628
Metric+kNN	0.800±0.002	0.797±0.003	0.900	0.931	0.964	1,669

4 Experiments and Results

4.1 Experimental Setup

With practical applicability in mind, we conduct three experiments in a stepwise manner: **Experiment 1** (method comparison, Section 4.2) compares the four methods for $U = 100$ with $n \in \{100, 200, 300, 400\}$; **Experiment 2** (imbalanced conditions, Section 4.3) uses all 68,222 reviews of the top 100 authors and a post-count cap $n_{\max} \in \{500, 700, 1000, 1500\}$; and **Experiment 3** (scaling, Section 4.4) increases $U \in \{2, \ldots, 1000\}$ at fixed $n = 186$. We then examine Collective Attribution on the best method from Experiment 3 (Section 4.5).

4.2 Experiment 1: Method Comparison ($U = 100$, n Variable)

Targeting the top 100 authors by post count ($U = 100$), we extracted the same number $n \in \{100, 200, 300, 400\}$ of reviews from each author and compared the four methods under unified conditions. For brevity, Table 2 reports the representative largest-data condition ($n = 400$). At $n = 400$, TF-IDF+LR achieved the highest Accuracy and Macro-F1, both reaching 0.870. BERT-FT followed closely with Accuracy 0.861, but in terms of computation time, TF-IDF+LR required only 157 seconds whereas BERT-FT required 5,394 seconds, a difference of about 34×. BERT-Emb+LR and Metric+kNN achieved lower accuracies of 0.806 and 0.800, respectively. TF-IDF+LR also achieved the best Top-k Accuracy, reaching 0.943 for Top-3, 0.961 for Top-5, and 0.979 for Top-10. The same qualitative ranking was observed across the other n settings, but the full n-wise results are omitted due to the page limit. These results indicate that, under the balanced condition with $U = 100$, TF-IDF+LR shows an advantage in both accuracy and computational efficiency. The interpretation of performance differences across methods and cross-experiment trends is discussed in Section 5.

4.3 Experiment 2: Imbalanced Conditions ($U = 100$)

Since the number of posts per author is uneven in practice, this experiment examined behavior under imbalanced conditions with $U = 100$, using (i) an **imbalanced (full-data) condition** with all 68,222 reviews from the top 100 authors (419–2,489 reviews per author) and (ii) a **post-count cap condition** with $n_{\max} \in \{500, 700, 1000, 1500\}$. As shown in Table 3, TF-IDF+LR achieved the highest accuracy under the full-data condition (Acc 0.900, Macro-F1 0.881), followed closely by BERT-FT (Acc 0.897), with BERT-Emb+LR and Metric+kNN trailing by 0.06–0.07; the method ranking thus matched Experiment 1. TF-IDF+LR also led in Top-k Accuracy (Top-10=0.986). Under the post-count cap condition, TF-IDF+LR remained best at all $n_{\max}$, with accuracy improving gradually but saturating beyond $n_{\max} = 1000$. These results confirm that, even under imbalanced conditions close to practical settings, TF-IDF+LR retains an advantage over BERT-FT in both accuracy and computational cost.

Table 3: Experiment 2: Accuracy comparison under imbalanced conditions ($U = 100$, 5-fold mean)

Condition	TF-IDF+LR	BERT-FT	BERT-Emb+LR	Metric+kNN
Full data (imbalanced, 68,222 reviews)	**0.900**	0.897	0.841	0.831
$n_{\max} = 500$	**0.881**	0.875	0.819	0.809
$n_{\max} = 700$	**0.888**	0.883	0.825	0.815
$n_{\max} = 1000$	**0.892**	0.887	0.830	0.819
$n_{\max} = 1500$	**0.896**	0.894	0.837	0.828

Computation times under the full-data condition: TF-IDF+LR 311s, BERT-FT 9,742s, BERT-Emb+LR 1,354s, Metric+kNN 1,975s (BERT-FT requires about 31× that of TF-IDF+LR).

4.4 Experiment 3: Scale to 1,000 Authors (U Variable, $n = 186$ Fixed)

To examine how an increase in the number of authors U affects the performance of each method, we evaluated 31 conditions: $U \in \{2, 3, \ldots, 10, 15, 20, \ldots, 100, 200, \ldots, 1000\}$. We uniformly extracted $n = 186$ reviews from each author and compared the performance changes across the four methods. Results at representative U values are shown in Table 4, and the Accuracy trend across all conditions is shown in Figure 1a.

Under small-scale conditions ($U \leq 10$), differences across methods were small; for example, at $U = 10$, Metric+kNN achieved the highest value with Accuracy 0.950. On the other hand, at $U \geq 20$, TF-IDF+LR consistently achieved the highest accuracy, and its gap to the other methods widened as the number of authors increased. At $U = 100$, TF-IDF+LR achieved 0.832 and BERT-FT 0.805, with a gap of 0.027; at $U = 1{,}000$, TF-IDF+LR achieved 0.623 and BERT-FT 0.528, with the gap widening to 0.095. A similar trend was confirmed for Macro-F1 (Figure 1b), with TF-IDF+LR showing the most stable trend under medium-to-large-scale conditions.

Regarding Top-k Accuracy, although Top-1 decreased with increasing number of authors, TF-IDF+LR maintained a Top-10 Accuracy of 0.821 even at $U = 1{,}000$ (Figure 2). This indicates a certain utility for narrowing down a candidate author set of around 1,000 to roughly the top ten (screening usage). All three other methods remained at lower levels than TF-IDF+LR, confirming a difference in candidate retrieval performance under large-scale conditions.

Regarding computation time, all methods showed an increase with growing U (Figure 1c). At $U = 1{,}000$, TF-IDF+LR required 8,886 seconds, while BERT-FT required 24,617 seconds (about 2.8×) and BERT-Emb+LR required 19,262 seconds (about 2.2×). Metric+kNN had a relatively stable training time, but its inference time increased with growing U because inference involves distance computations against the training data. The log-log plot (Figure 1c) visually confirms that TF-IDF+LR consistently maintained a substantially lower computational cost than the BERT-based methods across all conditions.

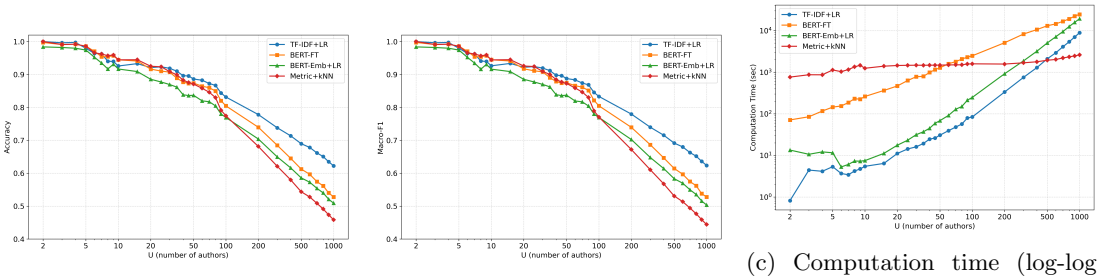
However, exact identification (Top-1) based on a single review at $U = 1{,}000$ remains at 0.623 even for TF-IDF+LR. To address this limitation, the next section (Section 4.5) examines performance improvement via test-time aggregation of multiple reviews. The interpretation of performance differences across methods and cross-experiment trends are discussed in Section 5.

4.5 Collective Attribution Results

To address the performance degradation under large-scale conditions observed in Experiment 3, this section examines the improvement effect of Collective Attribution (test-time aggregation). The target method is TF-IDF+LR, which achieved the highest accuracy in Experiment 3; we

Table 4: Experiment 3: Accuracy under scaling conditions (representative U values, $n = 186$, 5-fold mean)

U	TF-IDF+LR	BERT-FT	BERT-Emb+LR	Metric+kNN
2	**1.000**	**1.000**	0.989	**1.000**
5	**0.994**	0.993	0.978	0.993
10	0.926	0.936	0.914	**0.950**
20	**0.933**	0.919	0.884	0.925
50	**0.886**	0.877	0.839	0.871
100	**0.832**	0.805	0.769	0.800
200	**0.778**	0.740	0.704	0.681
300	**0.738**	0.685	0.649	0.621
400	**0.713**	0.645	0.616	0.580
500	**0.690**	0.613	0.586	0.544
700	**0.662**	0.575	0.554	0.509
1000	**0.623**	0.528	0.509	0.459



(a) Accuracy (log x-axis) (b) Macro-F1 (log x-axis) (c) Computation time (log-log axes)

Figure 1: Experiment 3: Trends with respect to the number of authors U for the four methods.

evaluated bundle sizes $b \in \{1, 2, 3, 5, 10\}$ across the 10 conditions $U \in \{100, 200, \ldots, 1000\}$. We compared two aggregation schemes: Concatenation and Probability Averaging. Training was performed only once per fold at the single-review level, so that differences in b are reflected solely in test-time aggregation.

Comparing the aggregation schemes, Concatenation tended to outperform Probability Averaging across all U conditions (Figure 3, left). In the following, we focus on the results of the Concatenation scheme.

The results of the Concatenation scheme are shown in Figure 3 (right). Accuracy improved substantially at all values of U as b increased. At $U = 1{,}000$, Accuracy improved from 0.623 with single-review attribution ($b = 1$) to 0.840 at $b = 2$, 0.914 at $b = 3$, 0.963 at $b = 5$, and **0.985** at $b = 10$. In particular, at $b = 10$, the difference between $U = 100$ (0.991) and $U = 1{,}000$ (0.985) was only 0.006, substantially mitigating the performance degradation caused by the increase in the number of authors. Even at $b = 2$, Accuracy reached 0.840 at $U = 1{,}000$, indicating that aggregating just two reviews substantially improves identification performance.

These results confirm that, when multiple reviews are available, Collective Attribution can achieve high identification performance even for large-scale candidate sets of around $U = 1{,}000$. The factors behind the superiority of Concatenation over Probability Averaging and the positioning of this method are discussed in Section 5.

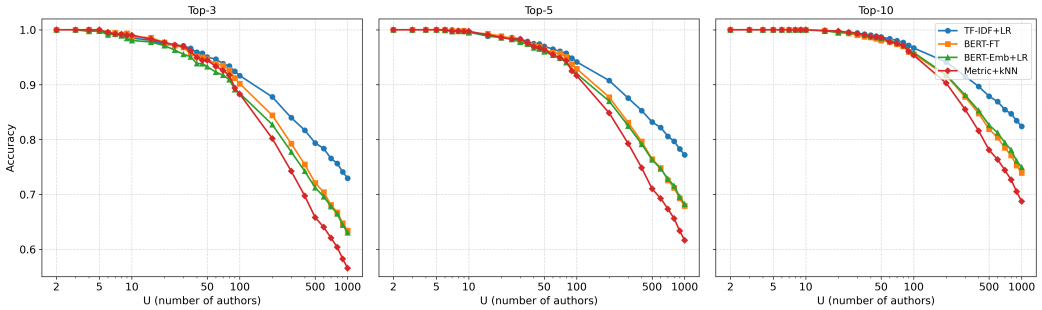


Figure 2: Experiment 3: Trend of Top-k Accuracy with respect to the number of authors U (four-method comparison)

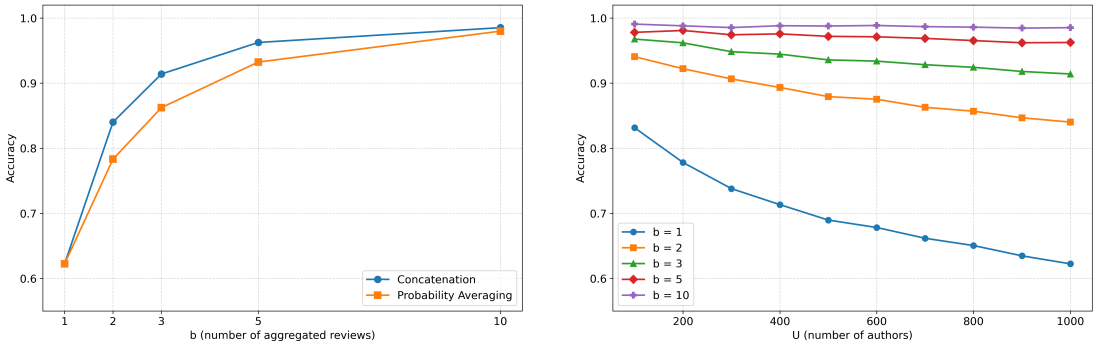


Figure 3: Collective Attribution: (left) comparison of Concatenation and Probability Averaging ($U = 1{,}000$); (right) Accuracy for each bundle size b as a function of U (Concatenation scheme)

5 Discussion

Cross-cutting the results of Experiments 1–3 and Collective Attribution, we describe the main observations under our experimental setting, the implications for practice, and the limitations of this study.

5.1 Cross-Experiment Synthesis

TF-IDF+LR based on character n-grams achieved accuracy comparable to or higher than the BERT-based methods over the wide range of $U = 100$ to $U = 1{,}000$, and was consistently far more advantageous in terms of computational cost. The same trend held under the imbalanced condition (Experiment 2), where TF-IDF+LR maintained its advantage in both accuracy and computational efficiency. This can be interpreted as follows: character n-grams directly capture subtle writing habits in Japanese short texts—such as particles, punctuation, sentence-ending expressions, and the choice between kana and kanji—whereas BERT-FT must adapt pre-trained representations, which are optimized for capturing meaning and topic, toward author-specific stylistic differences. In particular, while the parameter count of the multi-class classification layer grows in proportion to the number of authors U, the per-author training signal is fixed ($n = 186$); the effective training signal per class is therefore diluted as U grows. In contrast, the additional trainable parameters introduced for multi-class classification in TF-IDF+LR are

limited to the weights of the linear classification layer, which is presumed to make it relatively robust to increases in the number of authors.

Metric+kNN achieved the highest accuracy at $U = 10$ (0.950) but the lowest at $U = 1{,}000$ (0.459), a contrast consistent with the characteristics of metric learning. Under small U, a batch covers a large fraction of authors, so BatchHardTripletLoss readily obtains effective hard negatives; at $U = 1{,}000$, a batch of $P \times K = 16$ samples contains at most eight authors, so hard negatives against the remaining 992 cannot be observed and learning efficiency drops markedly. Nearest-neighbor inference also grows more prone to confusing similar authors as U increases. Metric learning is thus strong at small-to-medium scale but challenged under our large-scale multi-class setting.

Even under large-scale conditions where exact identification (Top-1) is difficult, we confirmed that Top-k evaluation yields practically useful performance for candidate retrieval. Even under $U = 1{,}000$ with $n = 186$, TF-IDF+LR maintains a Top-10 Accuracy of 0.821, indicating a certain utility for narrowing a candidate author set of around 1,000 down to roughly the top ten. This shows that Top-1 (exact identification) and Top-k (candidate retrieval) correspond to different operational uses.

Test-time aggregation, which is in line with profile-based authorship attribution [3, 4], has the effect of substantially compensating for the performance degradation observed with single reviews. The accuracy of TF-IDF+LR at $U = 1{,}000$ improved from 0.623 with single-review attribution ($b = 1$) to 0.985 with ten-review concatenation ($b = 10$). Concatenation outperformed Probability Averaging likely because it increases the variety and frequency of character n-grams per input, alleviating feature sparsity—improving large-scale performance through input-side aggregation alone, with the trained model unchanged.

5.2 Practical Implications

The findings of this study have the following implications for the operational design of actor analysis in threat intelligence.

The utility of Top-k Accuracy directly connects to a two-stage operation of candidate generation (retrieval) and confirmation (verification). A practical configuration is to use the lightweight TF-IDF+LR to rapidly narrow the broad candidate set down to several to a dozen candidates (retrieval), and then to confirm the top candidates (verification) by Collective Attribution—which aggregates multiple posts obtained from the target account—or by human inspection of stylistic and metadata cues, or by additional analysis based on domain knowledge. As shown in this study, the fact that TF-IDF+LR's Top-10 Accuracy reaches 0.821 even at $U = 1{,}000$, and that Collective Attribution (Concatenation, $b = 10$) achieves Accuracy 0.985 under the same condition, suggests that the retrieval-then-verification two-stage operation can function practically even on a large-scale candidate pool.

Collective Attribution improves identification without retraining the classifier: when multiple posts can be obtained from the target account (e.g., across multiple threads in the same forum), aggregating them yields more reliable judgments than a single post. Conversely, accounts with only one observed post cannot benefit, so an operational criterion such as "judge only when at least b posts are available" is conceivable.

Although this study uses a closed-set setting, unknown authors are common in practice. Introducing score-threshold judgment (with calibration such as temperature scaling) into Top-k-based candidate presentation offers a developmental path toward an open-set extension that can abstain when no top candidate matches the target.

5.3 Limitations

This study has the following limitations.

- **Closed-set assumption:** This study targets closed-set authorship attribution and does not address open-set settings that include unknown authors.
- **Domain specificity:** This study targets only Rakuten Ichiba reviews. Although reviews include short-text characteristics, formulaic expressions, and topic variation, they have inherent differences from forum posts and threat-intelligence-related texts; generalizability to other domains therefore lies outside the verified scope.
- **Confounding between n and total character count:** While we controlled the per-author review count n, we did not control the character length of each review. The interpretation of "how many reviews are sufficient" may be confounded with "how much total text volume is required."
- **Limited BERT-FT search space:** The search range for the learning rate, warmup, regularization, and similar hyperparameters was limited; the trends could change with broader BERT-FT exploration or continual pre-training.
- **Robustness against LLM-based stylistic transformation:** Given the current ease of stylistic transformation by large language models, the robustness of character n-gram features remains unverified.

6 Conclusion

In this study, we focused on authorship attribution based on stylistic features as a foundational technique to support actor analysis in threat intelligence, and conducted a unified comparison of four methods on Rakuten Ichiba reviews together with an evaluation of the effect of test-time aggregation (Collective Attribution). Under our experimental setting, the lightweight TF-IDF+LR based on character n-grams achieved accuracy comparable to or higher than the BERT-based methods under medium-to-large-scale conditions with substantially lower computational cost, and the utility of candidate retrieval under large-scale conditions was also confirmed via Top-k evaluation. In addition, we showed that the Concatenation scheme of test-time aggregation can improve Accuracy at $U = 1{,}000$ from 0.623 with single-review attribution to 0.985 with ten-review concatenation. These results suggest that, for authorship attribution on Japanese short web texts, the combination of lightweight character n-gram features with test-time aggregation can serve as one practical choice.

To apply the findings of this study to threat-intelligence-relevant texts such as those on the dark web, the following directions are necessary: extension to open-set settings, cross-platform identity linkage (record linkage), expansion to multilingual environments (e.g., English and Russian), re-evaluation under conditions where the character count is controlled, and evaluation of robustness against stylistic transformation by large language models. We leave these as directions for future work.

Acknowledgments

In this paper, we used the "Rakuten Dataset" provided by Rakuten Group, Inc. via the IDR Dataset Service of the National Institute of Informatics. We acknowledge the use of Claude

AI assistant (Anthropic) for the English translation of this paper. This work was supported in part by the JSPS/MEXT KAKENHI under Grant 24K14956.

References

[1] Efstathios Stamatatos. A survey of modern authorship attribution methods. *Journal of the American Society for Information Science and Technology*, 60(3):538–556, 2009.

[2] Maarten Fabien, Esaú Villatoro-Tello, Petr Motlicek, and Shantanu Chakraborty. BertAA: BERT fine-tuning for authorship attribution. In *Proc. 17th Int. Conf. on Natural Language Processing*, pages 127–137, 2020.

[3] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. N-gram-based author profiles for authorship attribution. In *Proc. PACLING*, pages 255–264, 2003.

[4] Nikos Potha and Efstathios Stamatatos. A profile-based method for authorship verification. In *Artificial Intelligence: Methods and Applications*, volume 8445 of *Lecture Notes in Computer Science*, pages 313–326. Springer, 2014.

[5] Frederick Mosteller and David L. Wallace. *Inference and Disputed Authorship: The Federalist*. Addison-Wesley, 1964.

[6] Efstathios Stamatatos. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy*, 21(2):421–439, 2013.

[7] Upendra Sapkota, Steven Bethard, Manuel Montes, and Thamar Solorio. Not all character n-grams are created equal: A study in authorship attribution. In *Proc. NAACL-HLT*, pages 93–102, 2015.

[8] Mingzhe Jin and Masakatsu Murakami. Authorship identification using the random forest method. *Proceedings of the Institute of Statistical Mathematics*, 55(2):255–268, 2007. in Japanese.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. NAACL-HLT*, pages 4171–4186, 2019.

[10] Qingyuan Ai et al. Whodunit? learning to contrast for authorship attribution. In *Proc. AACL-IJCNLP*, 2022.

[11] Risa Konuma, Huaping Zhang, Chunxiao Gao, and Juan Wang. Japanese author attribution using bert finetuning with stylometric features. In *Proc. 1st Int. Conf. on Intelligent Multilingual Information Processing*, volume 2395 of *Communications in Computer and Information Science*, pages 293–307. Springer, 2025.

[12] Moshe Koppel, Jonathan Schler, Shlomo Argamon, and Eran Messeri. Authorship attribution with thousands of candidate authors. In *Proc. SIGIR*, pages 659–660, 2006.

[13] Kim Luyckx and Walter Daelemans. Authorship attribution and verification with many authors and limited data. In *Proc. COLING*, pages 513–520, 2008.

[14] Janek Bevendorff et al. Overview of PAN 2020: Authorship verification. In *CLEF Working Notes*, 2020.

[15] Andrea Bacciu et al. Cross-domain authorship attribution combining instance-based and profile-based features. In *CLEF Working Notes*, 2019.

[16] Rakuten Group, Inc. Rakuten ichiba data. Informatics Research Data Repository, National Institute of Informatics, 2020. dataset, https://doi.org/10.32130/idr.2.1.

[17] Tohoku NLP Group. BERT base Japanese (unidic-lite with whole word masking, CC-100 and jawiki-20230102). Hugging Face model card, 2023. Accessed 2026-04-26.

[18] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. 2017.