# FORMING AN AUTHORSHIP PROFILE THROUGH N-GRAM TRACING

**Devi Ambarwati Puspitasari**
National Research and Innovation Agency of Republik Indonesia (BRIN)
devi018@brin.go.id

## ABSTRACT

In forensic linguistics, authorship profiles have become an important tool for identifying and characterizing individuals based on their writing style and authorship attribution. This study employed N-gram tracing, a corpus linguistics mixed computational method that investigates recurring patterns of sequences of words or characters in text. By analyzing WhatsApp (WA) messages extracted from case evidence, this study examines how N-gram patterns can reveal specific linguistic features associated with author identity. The dataset consists of personal texts and microblogs containing approximately 2.1 tokens. To ensure data integrity, the text was cleaned of non-traditional elements such as hyperlinks and media files during a preprocessing phase. N-grams on both character and word level, including N1-N5, were extracted and examined for diversity, frequency, distribution, and contextual usage patterns. To discern stylistic consistency across texts attributed to a particular individual, a machine learning model was used to calculate the similarity index and evaluate these linguistic fingerprints. Initial results suggest that certain N-gram patterns, such as orthographic selection and lexical choice, are highly indicative of individual influence. Profiling is also enhanced with linguistic markers such as abbreviations, code-switching, and unique styles present in informal communication. This study demonstrates that N-gram tracing is not only effective in identifying authorship but can also provide information on demographic and psychological features such as age, gender, and communication preferences. The fields of forensic linguistics and computational authorship analysis benefit from this study as it provides a robust and scalable technique for profiling authors based on real-world data. Furthermore, it highlights the ramifications in the legal context, emphasizing the potential for N-gram search to aid investigations where digital communication is critical. Extending the analysis to multilingual data and integrating semantic-level profiling to improve accuracy are future steps.

**Keywords:** *Authorship analysis, Forensic Linguistics, linguistic fingerprint, Multilingual Text Analysis, N-Gram Tracing*

## INTRODUCTION

Forensic linguistics has highlighted authorship profiling as a key method for identifying and categorizing people depending on writing style. The increased reliance on digital communication platforms and the expanding volume of textual data in criminal investigations have heightened the demand for precise and efficient tools in authorship attribution (Belvisi et al., 2020; McMenamin, 2002). Authorship profiling, compared to authorship attribution, is developing a profile of an unknown author's demographic or psychological characteristics—such as age, gender, educational background, or regional origin—based purely on their linguistic output. (Coulthard, Johnson, & Wright, 2017). This field integrates forensic stylistics, sociolinguistics, and computational linguistics to investigate how stylistic features may serve as linguistic fingerprints, uniquely identifying or characterizing an author.

Computational methods based on corpus linguistics are central to modern authorship profiling. Regarding these, N-gram tracing has proven to be a useful and robust technique (Puspitasari, Karlina, et al., 2024; Santos et al., 2021). An N-gram refers to a contiguous sequence of n items from a given sample of text or speech. These items can be phonemes, syllables, letters, words, or even parts of speech (Belvisi et al., 2020; Grieve et al., 2019). In this study, we employed word-based and character-based N-gram tracing (Belvisi et al., 2020; Grant, 2022) as a mixed computational method to investigate recurring linguistic patterns in texts attributed to specific authors. The rationale behind using N-grams lies in their ability to capture both syntagmatic structures and stylistic choices in a statistically measurable form (McEnery & Hardie, 2012).

The foundation of N-gram tracing lies in frequency analysis and pattern recognition. For instance, bigrams (2-grams) and trigrams (3-grams) are particularly useful for identifying habitual collocations (Belvisi et al., 2020; Toussaint et al., 2022), function word usage, and phraseology that may escape more superficial analysis. Such patterns often emerge unconsciously and are difficult for authors to disguise,

especially over longer texts. Studies have shown that even minor variations in frequency and distribution of N-grams can serve as reliable indicators of authorship (Juola, 2021; Perkins, 2021). Forensic applications leverage these subtle differences to distinguish between suspect texts, compare them against reference corpora, and support legal claims involving disputed authorship.

In the present study, N-gram tracing was performed on a corpus of digital texts created by individuals under investigation in a forensic context. The objective of this study was to see if distinct writing signatures could be identified and if they were constant enough across diverse texts to permit accurate authorship profiling. The implementation of a corpus-based approach offered various benefits. First, it allowed for the evaluation of enormous amounts of data, increasing the statistical reliability of the analysis. Second, it enabled the detection of both common and idiosyncratic linguistic patterns, ranging from lexical selection to syntactic composition.

Corpus linguistics tools such as AntConc (Anthony, 2011) and Python-based NLP libraries (NLTK) facilitated the extraction (Fedotova et al., 2022; Maskat et al., 2024), comparison, and visualization of N-gram frequencies. Through these tools, it was possible to generate ranked frequency lists of N-grams for each text sample and calculate similarity scores using statistical measures such as cosine similarity and Jaccard index (Permatasari et al., 2020; Puspitasari et al., 2024). In doing so, this study followed a quantitative paradigm, but with qualitative interpretations of the results—particularly when linking linguistic features to sociolinguistic variables like dialectal background or digital register preferences.

One key insight from the study was the consistency of functional N-gram usage across an individual's texts. For example, the repeated use of trigrams involving modal verbs (e.g., "would have been", "could not do") and discourse markers (e.g., "you know that", "I think that") suggested habitual phrase constructions that functioned almost like linguistic fingerprints. These constructions were often semantically neutral but structurally distinct, and they revealed preferences in modality, stance, and cohesion. Such markers proved useful in distinguishing between speakers in a way that went beyond mere vocabulary choice.

Importantly, the use of N-gram tracing in forensic linguistics must also contend with ethical and methodological challenges. Unlike DNA or fingerprint evidence, linguistic profiling remains probabilistic in nature (Biró, 2020; Peng et al., 2016). While certain patterns may suggest common authorship or shared linguistic background, they rarely provide absolute certainty. Consequently, forensic linguists must contextualize their findings within a framework of likelihood rather than determinism (Grant, 2008). Moreover, analysts must remain cautious about overfitting their models or drawing conclusions from insufficient data, particularly when dealing with short texts or multiple authors with similar sociolects.

Another limitation lies in genre sensitivity. N-gram distributions can vary significantly across genres, registers, or communicative settings (Nini, 2018; Puspitasari et al., 2023). For instance, instant messaging tends to involve more contractions, abbreviations, and emojis compared to academic writing. Therefore, in this study, text normalization and genre control were crucial pre-processing steps to ensure the reliability of N-gram comparisons. Without such normalization, the noise introduced by genre-specific features might overshadow meaningful stylistic patterns attributable to individual authorship.

Considering these limitations, the study emphasized the importance of N-gram tracing in forensic authorship profiling. Its strength resides in its objectivity, replicability, and adaptability to various text kinds. Furthermore, when paired with information (e.g., time stamps, device usage, interaction patterns), N-gram-based studies can contribute to a comprehensive forensic linguistic arsenal that assists law enforcement and the judiciary in digital criminal case investigations.

Authorship profiling in forensic linguistics is a multi-layered procedure that uses linguistic theory, corpus-based methodology, and computer techniques to investigate the relationship between language use and author identity. Unlike traditional stylistic analysis, authorship profiling tries not only to connect a work to a single individual, but also to deduce demographic, psychological, or sociolinguistic factors from linguistic data (Coulthard, Johnson, & Wright, 2017). One of the key approaches in this process is N-gram tracing, a computational method used to identify recurring sequences of characters or words within texts. These patterns, whether they are frequent collocations, function word combinations, or character sequences, often reflect unconscious habits of language use and can serve as linguistic fingerprints (McEnery & Hardie, 2012).

The use of N-gram analysis is particularly valuable in digital forensic investigations, where communication occurs through short, informal messages such as those found in WhatsApp (WA) chats. In such contexts, N-grams can reveal stylistic consistencies that remain stable across different messages, including preferences in punctuation, spelling variants, use of abbreviations, or phrase structures. These features can provide insight into an author's identity, sociolect, or communicative style (Grant, 2008).

The present study utilizes N-gram tracing to analyze a corpus of WhatsApp messages derived from forensic case data. The work shows that even seemingly insignificant digital texts can reveal considerable linguistic information by evaluating frequency patterns and stylistic indicators buried in N-grams. The findings highlight the empirical relevance of corpus-driven methods in forensic linguistics and emphasize the need of data-driven tools in modern authorship profiling, particularly as digital text grows increasingly crucial to criminal investigations. This research addresses two research questions:

1. How is the accuracy of N-gram tracing in identifying stylistic patterns that contribute to authorship profiling in digital forensic texts, such as WhatsApp messages?
2. What specific linguistic features, as revealed through N-gram analysis, are most indicative of author identity in informal digital communication?

This study revolves on the hypothesis that N-gram tracing may efficiently identify consistent stylistic trends within an individual's WhatsApp conversations, which can then be used for authorship profiling. It is hypothesized that such patterns, particularly recurrent word and character sequences, are stable and distinct enough to distinguish one author from another, even in informal and brief digital messages. The null hypothesis, on the other hand, proposes that N-gram analysis across communications from the same author will reveal no consistent or unique patterns.

Furthermore, the study proposes that some linguistic elements recorded by N-gram analysis, particularly those involving functional word combinations (e.g., modal verbs, discourse markers) and punctuation-based sequences, are more predictive of author identification than content terms. These components often represent unconscious writing patterns and stylistic preferences, making them useful markers for forensic comparison. The alternative, or null hypothesis, states that these traits do not outperform content words in detecting authorship in digital communication environments.

## METHODOLOGY

The present study utilizes a corpus-based computational approach to authorship profiling referred to N-gram analysis, with an emphasis on personal writings and microblog messages. The dataset contains around 2.1 million tokens collected from individual WhatsApp conversations and publicly available microblog content, resulting in a diverse linguistic sample that includes both private and semi-public digital communication. To ensure data integrity and analytical consistency, a thorough preprocessing phase was conducted. During this stage, non-traditional textual components like hyperlinks, emoticons, media file references, and metadata were removed. The corpus information is presented in Table 1.

**Table 1. Corpus Composition and Preprocessing Overview**

| Component | Source | Quantity / Volume | Description |
|---|---|---|---|
| Total Tokens (Words) | WhatsApp & Microblogs | 2,100,000 tokens | Token count after preprocessing |
| WhatsApp Text Messages | Private Chat Logs | 145,000 messages | Collected from individual users with consent |
| Microblog Posts (e.g., Twitter/Threads) | Public Online Sources | 57,000 posts | Focused on short, informal written texts |
| Average Message/Post Length | Both Sources Combined | ~11.2 tokens per entry | Reflects informal, brief nature of communication |
| Removed Hyperlinks | All Sources | 38,475 URLs removed | Identified via regex filtering |
| Removed Emojis & Emoticons | WhatsApp + Microblogs | ~64,820 emoji instances removed | Normalized or removed for uniform analysis |
| Removed Media References (e.g., images) | WhatsApp metadata | 13,420 media tags removed | Examples: "image omitted", "video file" |
| Removed Metadata | WhatsApp & Blog Platforms | 74,103 metadata tags removed | Includes timestamps, sender IDs, platform-specific tags |
| Final Cleaned Corpus Size | After Preprocessing | ~1,909,182 tokens | ~9% reduction from original raw corpus |
| N-gram Units Extracted (Word Level) | N=1 to N=5 | 1-gram: 1.9M<br>2-gram: 1.6M<br>3-gram: 1.2M<br>4-gram: 978K<br>5-gram: 763K | Frequency tables generated for each N-level |

| | | | |
|---|---|---|---|
| N-gram Units Extracted (Character Level) | N=1 to N=5 | 1-gram: 6.8M<br>2-gram: 5.4M<br>3-gram: 4.2M<br>4-gram: 3.6M<br>5-gram: 3.0M | Captures stylistic subtleties (e.g., spelling patterns, suffixes, etc.) |

Additionally, standardized formatting was applied to correct irregularities in spelling and punctuation while preserving key orthographic choices relevant to authorship. Following this, the data was segmented and annotated according to individual authorship attribution, enabling intra-author and inter-author comparisons. Central to the analysis was the extraction of N-grams on both the word and character levels, ranging from unigrams (N1) to 5-grams (N5). These were examined across several dimensions, including frequency of occurrence, positional distribution, and contextual usage patterns. Both frequent and rare N-gram patterns were considered, as low-frequency but highly idiosyncratic sequences may serve as more reliable markers of individual authorship.

Diversity indices were also calculated to gauge the lexical variability of each author's linguistic output. To determine the degree of stylistic consistency across texts attributed to a single author, machine learning models were employed. Specifically, vectorization techniques such as Term Frequency-Inverse Document Frequency (TF-IDF) were used to represent N-gram features, and cosine similarity measures were applied to calculate a similarity index across textual samples. This enabled the quantification of stylistic coherence within an author's texts and the detection of deviant patterns in suspect messages. Supervised classification models such as support vector machines (SVM) were further trained to evaluate the predictive power of N-gram features in distinguishing between authors.

Preliminary research (Puspitasari et al., 2023; Puspitasari, Fakhrurroja, et al., 2024) indicate that certain patterns—particularly those involving orthographic variation, preferred lexical collocations, and function word sequences—are strongly indicative of individual linguistic fingerprints. For example, authors demonstrated consistent choices in contractions, punctuation spacing, and phrase structures that reappeared across different contexts and platforms. These researchers support the hypothesis that N-gram-based stylistic features can be leveraged not only to differentiate between authors but also to characterize their distinctive linguistic behavior. By integrating computational techniques with forensic linguistic theory, this methodology offers a scalable and replicable framework for digital authorship profiling in forensic contexts. The combination of quantitative frequency analysis and similarity-based machine learning provides a robust basis for examining authorship in an era where short, informal texts dominate interpersonal communication. Moreover, the use of both character- and word-level N-grams allows for a multi-resolution view of stylistic behavior, capturing both micro-level (e.g., spelling choices) and macro-level (e.g., phrasal patterns) linguistic signals.

## RESULTS

### 1. Stylometric Features

Profiling is also enhanced with linguistic markers such as abbreviations, code-switching, and unique styles present in informal communication. This study demonstrates that N-gram tracing is not only effective in identifying authorship but can also provide information on demographic and psychological features such as age, gender, and communication preferences. The fields of forensic linguistics and computational authorship analysis benefit from this study as it provides a robust and scalable technique for profiling authors based on real-world data. Furthermore, it highlights the ramifications in the legal context, emphasizing the potential for N-gram search to aid investigations where digital communication is critical. Extending the analysis to multilingual data and integrating semantic-level profiling to improve accuracy are future steps.

The findings of this study confirm that informal digital communication, particularly from platforms like WhatsApp and microblogs, exhibits a range of stylometric features that are idiosyncratic and largely stable over time. Among the most prominent are the use of abbreviations, emoticons, and spelling variants, all of which are effectively captured using N-gram tracing at both the character and word levels. For instance, users consistently employ abbreviations such as "btw," "idk," or regional variants that reflect linguistic background or social identity. Similarly, orthographic habits such as repeated use of exclamation marks or lowercase personal pronouns (e.g., "i" instead of "I") often persisted across the corpus, thereby functioning as individual markers. The common stylometric features captured by N-gram Tracing shown in Table 2.

Another significant discovery is the role of code-switching—alternating between languages within a single message—which was particularly prevalent among bilingual or multilingual users. N-gram analysis successfully captured these shifts, especially when using character-level N-grams that detect sub-word patterns and transitions between different linguistic systems. This has implications not only for identifying users but also for understanding their linguistic environment, education level, or cultural affiliation. The frequency and contextual positioning of such code-switching patterns served as a reliable feature in determining the communicative tendencies of individuals.

**Table 2. Common Stylometric Features Captured by N-gram Tracing**

| Type of Stylometric Feature | Example from WhatsApp Data | N-gram Level Captured | Linguistic / Social Meaning |
|---|---|---|---|
| Abbreviations | “btw” (by the way), “idk” (I don't know), “gmn” (gimana), “otw” (on the way) | Word-level N1–N2 | Reflects informal style; often associated with younger age and communication efficiency |
| Emoticons & Emojis | “😂”, “🙂”, “:D”, “:(“ | Character-level N2–N4 | Expresses emotional tone and non-verbal communication preferences |
| Spelling Variants | “akuuu” vs “aqu/aku”, “iyaa” vs “iya/iyah, “okeey” vs “oke/okeh”, “ngga” vs “ga” | Character-level N3–N5 | Indicates personality traits, regional background, or individual writing style |
| Consistent Lowercase Use | “ok” used instead of “OK”, “pr” used instead of “PR”, “gw” used instead of “GW” | Character-level N1–N2 | A personal orthographic habit, often seen in quick and informal communication |
| Punctuation Repetition | “!!!”, “??”, “...”, “?!?” | Character-level N2–N4 | Used to emphasize emotion or tone in digital conversation |
| Code-Switching | “*aku lagi ngoding nih, soalnya* ***deadline*** *besok* [I'm coding right now, because the deadline is tomorrow]” | Character-level >N5 | Reflects bilingual background; common in highly multilingual communities |

The present study employs N-gram analysis—a computational technique that examines sequences of 'n' items (such as characters or words) in a given text—to detect and analyze code-switching patterns in a corpus comprising approximately 2.1 million tokens from WhatsApp conversations and microblogs. By focusing on both character-level and word-level N-grams, the analysis captures subtle shifts between languages, providing insights into the linguistic behavior of bilingual users.

One of the significant findings is that character-level N-grams are particularly effective in identifying code-switching instances. This is because character-level analysis can detect sub-word patterns and transitions that may not be evident at the word level, especially in informal digital texts where spelling variations and abbreviations are common. For example, the transition from Indonesian to English within a single sentence can be captured by analyzing sequences of characters that deviate from the typical patterns of one language and align with another. The following examples are extracted from the corpus shown in Table 3.

**Table 3. Code-switching found in the data**

| Example | Language Switch | Contextual Interpretation |
|---|---|---|
| "Aku lagi ngoding nih, soalnya deadline besok." | Indonesian → English | The user switches to English for the term "ngoding" (coding), reflecting a technical context. |
| "Besok meeting di office, jangan telat ya." | Indonesian → English | "Meeting" and "office" are English terms embedded in an Indonesian sentence, indicating a professional setting. |
| "Gue udah submit tugasnya, tinggal upload ke drive aja." | Indonesian → English | "Submit" and "upload" are English verbs used within an Indonesian framework, common in academic or work-related communication. |

These examples demonstrate how code-switching is employed to convey specific meanings, often influenced by the topic of conversation, the speaker's linguistic background, and the social context. The use of English terms within Indonesian sentences, particularly in professional or technical discussions, suggests a level of proficiency and comfort with both languages.

The frequency and contextual positioning of code-switching patterns serve as reliable features in determining communicative tendencies. For instance, users who frequently switch languages in technical discussions may have educational or professional backgrounds that necessitate bilingual proficiency. Similarly, the use of English terms in casual conversations may indicate exposure to English-language media or social circles where such code-switching is normative.

Moreover, the analysis of code-switching patterns can provide insights into the linguistic environment and cultural affiliations of users. In regions where English is commonly used in education and business, code-switching may be more prevalent and socially accepted. Conversely, in areas with less exposure to English, code-switching may be limited to specific contexts or social groups.

The implications of these findings extend beyond linguistic analysis. In forensic linguistics and author profiling, understanding code-switching behavior can aid in identifying individuals based on their unique linguistic patterns. For example, consistent use of certain English terms within Indonesian sentences may point to specific educational backgrounds or professional experiences. Additionally, recognizing code-switching tendencies can assist in tailoring communication strategies in multilingual societies, ensuring messages resonate effectively across different linguistic groups.

## 2. Demographic Indicators

By applying a machine learning model to the extracted N-gram features, the study was able to predict certain demographic attributes with a reasonable degree of accuracy. For example, younger users were more likely to use internet slang and emoticons, whereas older users tended toward full word forms and conventional punctuation. Gender-based variations also emerged: female users were found to use more expressive punctuation (e.g., multiple exclamation marks), emotive lexical items, and interjections, while male users showed a preference for abbreviated directives and clipped syntactic forms as shown in Table 4. These observations reinforce prior findings in sociolinguistics and suggest that N-gram tracing, when calibrated appropriately, can uncover not only who authored a message but also likely demographic features.

**Table 4. Demographic Indicators Detected via N-gram Patterns**

| Demographic Variable | Key Indicators via N-grams | Observed Patterns |
|---|---|---|
| Age Group | Use of slang, abbreviations, emoji | Younger users: more slang/emojis; Older: full word forms |
| Gender | Expressive punctuation, interjections (female users) | Female: “omg!!!”<br>Male: “ok” |
| Language Background | Code-switching, lexical borrowings | Bilinguals: frequent language switching mid-sentence |
| Education Level | Formality of syntax, orthographic correctness | Higher education: fewer spelling variants |

The practical relevance of these findings is most clearly illustrated in forensic scenarios. In digital forensic investigations, the identification of anonymous authors or the authentication of suspicious messages can be critical. In such cases, the ability to attribute a text to an individual based on distinctive N-gram patterns provides investigators with an evidentiary tool that is both data-driven and methodologically transparent. By applying similarity measures such as cosine similarity on N-gram frequency vectors, investigators can compare unknown texts against reference profiles and determine likely matches. The interpretability of the N-gram results—particularly when coupled with visualization tools like frequency heat maps or distribution graphs—makes the technique accessible for both expert analysts and legal professionals.

In addition to its forensic utility, the methodology proves scalable. It can be extended across various platforms and adapted to larger datasets, making it applicable to law enforcement agencies, cybercrime units, and legal firms dealing with digital evidence. As data volume increases, so too does the potential to refine models, reduce false positives, and improve classification accuracy. Further integration with semantic analysis (e.g., topic modeling, sentiment analysis) may enhance the interpretive depth of the results, offering a more nuanced view of authorship and intent.

One of the novel contributions of this study is the validation of linguistic profiling in real-world, multilingual contexts. In analyzing WhatsApp messages from speakers who frequently switched between English and regional languages (e.g., Indonesian and Javanese), the study uncovered unique patterns of hybridized language use. These patterns included intra-sentential code-switching, borrowing of lexical items, and culturally specific emotive expressions. N-gram tracing captured these nuances effectively, underscoring the method's adaptability to diverse linguistic ecologies.

**Table 5. Forensic Utility of N-gram Profiling**

| Use Case | Description | Benefit |
|---|---|---|
| Authorship Verification | Matching unknown texts to known author profiles | High accuracy in identifying likely authors |
| Group Chat Author Isolation | Distinguishing individuals within shared accounts | Enables disambiguation in group communication |
| Psychological Profiling (Preliminary) | Inferring affective state and preferences via language use | Supports behavioral insights for investigative contexts |
| Multilingual Forensic Application | Handling messages with mixed-language structure | Adaptable to regional and multilingual digital texts |

Importantly, the study also addresses limitations, notably the challenge of data sparsity in short texts and the difficulty of isolating authorial signals in group chats or shared accounts. To mitigate these issues, the study incorporated normalization techniques and controlled for conversational context when possible. Future work may explore techniques such as stylometric clustering or unsupervised learning to further refine author identification in such complex environments.

From a legal perspective, the implications of this research are profound. Courts increasingly rely on digital evidence, yet there remains skepticism about the scientific rigor of authorship attribution methods. By grounding N-gram analysis in reproducible statistical models and real-world data, this study helps bridge the gap between linguistic insight and forensic admissibility. These benefits are shown in Table 5. It suggests that linguistic profiling can serve as a complementary form of evidence, particularly when corroborated with metadata, contextual information, or testimony.

The use of N-gram tracing thus emerges not merely as a technical tool but as a forensic methodology with the potential to reshape investigative practices. It empowers analysts to move beyond subjective stylistic judgments toward quantifiable, empirical evaluations of text. Moreover, it opens avenues for interdisciplinary collaboration—between linguists, data scientists, and legal professionals—to co-develop standards and best practices for digital authorship verification.

## Discussion

The present study investigated the effectiveness of N-gram tracing in authorship profiling, particularly in the context of informal digital texts such as WhatsApp messages. The first research question focused on evaluating the accuracy of N-gram tracing in identifying stylistic patterns that contribute to authorship profiling. The findings demonstrate that N-gram tracing, particularly when applied at both the character and word levels, achieves a high degree of accuracy in identifying authorship based on stylistic consistency. This is supported by computational results derived from similarity indices such as cosine similarity, which revealed significant intra-author stability across messages. When trained using supervised machine learning models, especially support vector machines (SVMs) and random forest classifiers, the system achieved accuracy rates ranging from 82% to 93% in correctly attributing authorship within a multilingual WhatsApp corpus. The test results shown in Table 6.

**Table 6. Authorship Attribution Accuracy Using Machine Learning Models**

| Model | N-gram Level | Features Used | Accuracy (%) | Precision | Recall | F1 Score |
|---|---|---|---|---|---|---|
| Support Vector Machine | Word-level N3 (Trigrams) | Frequency of word trigrams | 89.4% | 0.91 | 0.88 | 0.89 |
| Support Vector Machine | Char-level N4 | Character patterns and transitions | 93.1% | 0.94 | 0.92 | 0.93 |
| Random Forest | Word-level N2 (Bigrams) | Word pair frequencies | 85.7% | 0.86 | 0.85 | 0.85 |

| | | | | | | |
|---|---|---|---|---|---|---|
| Random Forest | Char-level N3 | Sub-word spelling & stylometric cues | 82.3% | 0.84 | 0.81 | 0.82 |
| Baseline Model (Naive Bayes) | Word-level N1 | Word frequency | 74.6% | 0.76 | 0.74 | 0.75 |

These findings are consistent with prior research by Grieve (2007), who demonstrated that character-level N-grams are effective for stylometric analysis, and by Kestemont (2014), who emphasized their robustness across genres and platforms. The integration of N-gram features into machine learning classifiers enables precise modeling of authorship attribution thus validating the hypothesis that such patterns can be reliably used for forensic identification.

A crucial factor behind the success of N-gram tracing lies in its capacity to capture micro-level stylistic elements that are often unconscious and difficult to manipulate deliberately. These include habits such as punctuation spacing, frequent collocations, and orthographic idiosyncrasies—elements often overlooked by more traditional stylometric features like sentence length or vocabulary richness. For example, consistent use of lower-case personal pronouns ("i" instead of "I") or frequent insertion of emojis in specific syntactic positions were strong identifiers in several participant profiles. These patterns were found to remain consistent across messages and contexts, reinforcing the notion that such traits constitute part of an individual's linguistic fingerprint (Coulthard, 2013).

The second research question examined which specific linguistic features, as revealed through N-gram analysis, are most indicative of author identity in informal digital communication. The results indicate that function word usage, character sequences reflecting spelling choices, and punctuation patterns were among the most predictive indicators of individual authorship. Function words such as prepositions, conjunctions, and pronouns are typically used unconsciously and, as a result, offer strong stylistic signals (Argamon et al., 2009). The analysis revealed that users tend to adopt stable preferences in conjunction use (e.g., frequent use of "so" or "but" to link clauses), emotive discourse markers (e.g., "ugh," "lol"), and repeated sequences like multiple exclamation marks or trailing ellipses.

In addition to function words and punctuation, code-switching behavior emerged as a highly distinctive feature in multilingual users. Character-level N-grams allowed the model to detect frequent language switches, often occurring within the same sentence or even word. This is particularly relevant in contexts such as Southeast Asia, where multilingualism is common, and users blend English with regional languages such as Bahasa Indonesia or Tagalog. Studies by Estival et al. (2007) and Soler-Company and Wanner (2017) have similarly shown that multilingual patterns and borrowed phrases provide salient features for author profiling, especially in environments with high linguistic diversity.

Furthermore, demographic features such as age and gender were indirectly captured through stylometric behaviors. Younger users often favored internet slang (e.g., "idk," "smh"), emojis, and abbreviations, while older users tended to use complete words and proper punctuation. These distinctions align with findings by Goswami et al. (2009), who showed that age and gender influence digital language use in microblog platforms. Female users in the study were observed to use more expressive punctuation and emotional interjections (e.g., "omg!!!", "awww"), whereas male users favored concise statements and directive forms (e.g., "k", "sure"). While these observations are not definitive predictors on their own, they provide additional layers of profiling that can be triangulated with other evidence.

The hypothesis underpinning this study posited that N-gram tracing would reveal stable and individual-specific patterns in digital texts that could be used for authorship attribution and profiling. The results affirm this hypothesis, indicating that N-gram patterns do indeed reflect underlying linguistic habits that are both measurable and distinctive. Importantly, the study also rejected the null hypothesis that no significant patterns would emerge, as statistical tests demonstrated significant differences in N-gram distributions across authors. These findings validate previous scholarship emphasizing the forensic potential of stylometric profiling (Grant, 2007; Stamatatos, 2009).

Additionally, the hypothesis extended to the notion that character-level N-grams may outperform word-level N-grams in identifying subtle stylistic patterns. The evidence supports this claim, particularly in short or informal texts where word-level features may be sparse or noisy. Character N-grams were especially useful in capturing orthographic variations and idiosyncratic spelling, which are key in informal communication such as WhatsApp chats. These findings corroborate the work of Sapkota et al. (2015), who demonstrated that character N-grams provide greater flexibility and robustness across variable linguistic inputs.

As a result, this study provides a valuable contribution to the field of forensic linguistics. by demonstrating that N-gram analysis, when coupled with computational techniques, offers a scalable,

empirical, and context-sensitive method for authorship profiling. It offers evidence-based support for the growing use of digital linguistic evidence in legal investigations and sets the stage for further developments in multilingual authorship attribution, semantic profiling, and real-time forensic applications.

## CONCLUSION

This research demonstrates how N-gram analysis can reveal sociolinguistic information such as age group, gender preferences, and linguistic background, in addition to determining stylistic consistency. The model's capacity to adapt to the complexities of informal digital communication is bolstered by its successful detection of code-switching patterns, emoticon usage, and spelling variation as significant components. The implications are significant from a forensic standpoint: N-gram tracing, when combined with machine learning, gives investigators a data-driven, empirically supported approach to reducing suspect pools, confirming authorship claims, and profiling unknown communicators in cybercrime or digital threat contexts. This study contributes to the expanding field of forensic linguistics literature by providing a methodological framework that is both robust in theory and practical. It also provides up the possibilities for future studies on cross-platform stylometric comparison, deeper semantic-level profiling, and applications in languages other than Indonesian-English bilingualism. Furthermore, this study confirms that N-gram-based profiling, when founded on solid computational and linguistic principles, can be an effective tool for forensic identification and characterization of digital authors, bridging the gap between language, identity, and technology in today's rapidly evolving communicative landscape.

## REFERENCES

Anthony, L. (2011). AntConc. In *AntConc (Version 3.2.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/*.

Argamon, S., Koppel, M., Pennebaker, J. W., & Schler, J. (2009). Automatically profiling the author of an anonymous text. Communications of the ACM, 52(2), 119-123.

Belvisi, N. M. S., Muhammad, N., & Alonso-Fernandez, F. (2020). *Forensic Authorship Analysis of Microblogging Texts Using N-Grams and Stylometric Features*.

Biró, E. (2020). Linguistic Identities in the Digital Space. *Acta Universitatis Sapientiae, Philologica*, *11*(2). https://doi.org/10.2478/ausp-2019-0011

Coulthard, M. (2013). An introduction to forensic linguistics: Language in evidence. Routledge.

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

Estival, D., Gaustad, T., Pham, S. B., Radford, W., & Hutchinson, B. (2007). Author profiling for English emails. Proceedings of the 10th Conference of the Pacific Association for Computational Linguistics (PACLING), 263–272.

Fedotova, A., Romanov, A., Kurtukova, A., & Shelupanov, A. (2022). Authorship attribution of social media and literary russian-language texts using machine learning methods and feature selection. *Future Internet*, *14*(1). https://doi.org/10.3390/fi14010004

Goswami, S., Sarkar, S., & Rustagi, M. (2009). Stylometric analysis of bloggers' age and gender. Proceedings of the Third International AAAI Conference on Weblogs and Social Media, 214–217.

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. International Journal of Speech, Language & the Law, 14(1), 1–25.

Grant, T. (2008). Approaching questions in forensic authorship analysis. In C. N. Candlin & M. Gotti (Eds.), Intercultural aspects of specialized communication (pp. 329–344). Peter Lang.

Grant, T. (2022). The Idea of Progress in Forensic Authorship Analysis. In *The Idea of Progress in Forensic Authorship Analysis*. https://doi.org/10.1017/9781108974714

Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. Literary and Linguistic Computing, 22(3), 251–270.

Grieve, J., Clarke, I., Chiang, E., Gideon, H., Heini, A., Nini, A., & Waibel, E. (2019). Attributing the Bixby Letter using n-gram tracing. *Digital Scholarship in the Humanities*, *34*(3), 493–512. https://doi.org/10.1093/llc/fqy042

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

Juola, P. (2021). Verifying authorship for forensic purposes: A computational protocol and its validation. *Forensic Science International*, *325*. https://doi.org/10.1016/j.forsciint.2021.110824

Kestemont, M. (2014). Function words in authorship attribution: From black magic to theory? Proceedings

of the 3rd Workshop on Computational Linguistics for Literature, 59–66.

Maskat, R., Azman, N. A., Nulizairos, N. S. S., Zahidin, N. A., Mahadi, A. H., Norshamsul, S. R., Sharif, M. M. M., & Mahdin, H. (2024). A bi-annotated Malay-English code-switching (Manglish) dataset of X posts for biological gender identification and authorship attribution. *Data in Brief*, *52*. https://doi.org/10.1016/j.dib.2024.110034

McEnery, T., & Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

McMenamin, G. R. (2002). *Forensic linguistics: advances in forensic stylistics*. CRC Press LLC.

Nini, A. (2018). An authorship analysis of the Jack the Ripper letters. *Digital Scholarship in the Humanities*, *33*(3), 621–636. https://doi.org/10.1093/LLC/FQX065

Peng, J., Choo, K. K. R., & Ashman, H. (2016). Bit-level n-gram based forensic authorship analysis on social media: Identifying individuals from linguistic profiles. *Journal of Network and Computer Applications*, *70*, 171–182. https://doi.org/10.1016/j.jnca.2016.04.001

Perkins, R. C. (2021). The Application of Forensic Linguistics in Cybercrime Investigations. *Policing (Oxford)*, *15*(1). https://doi.org/10.1093/police/pay097

Permatasari, D. A., Fakhrurroja, H., & Machbub, C. (2020). Human-Robot Interaction Based on Dialog Management Using Sentence Similarity Comparison Method. *International Journal on Advanced Science, Engineering and Information Technology*, *10*(5), 1881–1888. https://doi.org/10.18517/ijaseit.10.5.7606

Puspitasari, D. A., Fakhrurroja, H., & Sutrisno, A. (2023). Identify Fake Author in Indonesia Crime Cases: A Forensic Authorsip Analysis Using N-gram and Stylometric Features. *2023 International Conference on Advancement in Data Science, E-Learning and Information System (ICADEIS)*, 1–6. https://doi.org/10.1109/ICADEIS58666.2023.10271069

Puspitasari, D. A., Fakhrurroja, H., & Sutrisno, A. (2024). AUTHORSHIP ANALYSIS IN ELECTRONIC TEXTS USING SIMILARITY COMPARISON METHOD. *Linguistik Indonesia*, *42*(1). https://doi.org/10.26499/li.v42i1.544

Puspitasari, D. A., Karlina, Y., Hernina, H., Kurniawan, K., Sutejo, S., & Danardana, A. S. (2024). Language Choices and Digital Identity of High School Student Text Messages in the New Capital City of Indonesia: Implication for Language Education. *International Journal of Language Education*, *8*(1). https://doi.org/10.26858/ijole.v8i1.63833

Santos, F. A. O., Macedo, H. T., Bispo, T. D., & Zanchettin, C. (2021). Morphological skip-gram: Replacing fasttext characters n-gram with morphological knowledge. *Inteligencia Artificial*, *24*(67). https://doi.org/10.4114/intartif.vol24iss67pp1-17

Sapkota, U., Bethard, S., Montes, M., & Solorio, T. (2015). Not all character n-grams are created equal: A study in authorship attribution. Proceedings of NAACL-HLT, 93–102.

Soler-Company, J., & Wanner, L. (2017). On the use of character and POS N-grams for the automatic detection of text complexity levels. Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL), 345–351.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3), 538–556.

Toussaint, P. A., Renner, M., Lins, S., Thiebes, S., & Sunyaev, A. (2022). Direct-to-Consumer Genetic Testing on Social Media: Topic Modeling and Sentiment Analysis of YouTube Users' Comments. *JMIR Infodemiology*, *2*(2). https://doi.org/10.2196/38749