ISSN: 2278-4012, Volume:15, Issue:2, June 2026

# Forensic Linguistics Applications for Determining Disputed Authorship of Digital Communications via Quantitative Stylometry

**I Wayan Eka Mahendra**[1]

[1]Lecturer, Department of Public Administration, Ngurag Rai University, Denpasar, Indonesia.
E-mail: eka.mahendra@unr.ac.id, Orcid: https://orcid.org/0000-0001-6085-943X

***Abstract:*** Court cases involving authorship just as in the cyber world have been on an unprecedented peak with the rapid growth of digital communication media, in the form of emails, instant messaging and social media. The anonymity, brevity and informality of online text are quite challenging insofar as attempting to identify the real author of digitally produced texts. In this paper, will endeavor to respond to these issues by applying quantitative stylometric approaches of authorship attribution of online messages. The unique writing patterns found in the methodology are by using a series of linguistic measures which are Type-Token Ratio (TTR), Measure of Textual Lexical Diversity (MTLD), frequency of function words and character-level n-grams. As has been illustrated in this discussion, one can differentiate between the authors of texts using these stylometric characteristics since outline coherent lexical, syntactic and structural characteristics. These findings highlight the fact that, despite the brevity and the informality of texts, stylistic cues that can be quantified may be applied to effectively determine the author. It is a supplement to the growing literature on forensic linguistics because it presents an empirically valid and theoretically sound way of identifying the authorship and finds an application in the field of cybercrime investigations, plagiarism and linguistic evidence in the court of law.

***Key Words*:** forensic linguistics, stylometry, authorship attribution, digital communication, quantitative analysis, linguistic profiling.

**(Received: 10 March 2026; Revised: 22 April 2026; Accepted: 14 May 2026; Published: 30 June 2026)**

## Introduction

Forensic linguistics is an interdisciplinary field and applies linguistic knowledge, methods, and skills to the law, particularly in the analysis of language in the courtroom. It comprises authorship attribution, discourse analysis, and interpretation of legal texts. However, the field has evolved over the years from the traditional approaches of qualitative analysis into more systematic and computational approaches with the introduction of statistical and machine learning techniques to aid in enhancing objectivity and reliability of linguistic studies (Mani et al., 2025). This evolution has highly enabled the application of forensic linguistics in the modern-day judicial process whereby language evidence is increasingly becoming an area of concern.

Forensic linguistics has gained more and more importance in the legal and cyberspace. With the ever-growing cybercrime, online harassment, plagiarism, and online fraud, it is now necessary to have the writer of the controversial texts identified. Linguistic analysis has become a regular occurrence in the legal profession to help in evidence-based decision-making and to decide cases of anonymous or disputed communications (Azmi, 2025). Moreover, forensic linguistics plays a significant role in studying the defamation, plagiarism, and authenticity of legal texts, which indicates its broad usage in different areas of law and electronic communications (Helmi et al., 2025).

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

The rapid increase in digital forms of communication such as emails and instant messaging software as well as social media has presented new challenges to authorship attribution. The digital communications are more often short, informal, and extremely volatile in structure in comparison to written work. Frequently full of abbreviations, emojis, and non-standard grammar, making it more difficult to linguistically analyze them. In addition, code-switching and the use of more than one language also make it difficult to identify the author, particularly in mixed-language environments. The requirements of such challenges determine the necessity to possess more powerful and flexible ways of analyzing devices, which could cope with such variability.

Although richer literature has been established on the subject, there is still a huge gap regarding the use of holistic quantitative stylometric approaches to online texts. The available studies are more of a conventional corpus study or studies that are preoccupied with a few aspects of linguistics, devoid of developing a series of stylometric indicators. The paper will therefore seek to address this gap by quantitatively applying stylometry to the controversial online messages. The aims are to determine the specific linguistic features, use statistical and computational models, and assess the validity of authorship attribution. By this, the study assists in enhancing the reliability and applicability of forensic linguistic study in online spaces nowadays.

This paper is structured as follows: Section 1 presents the definition of forensic linguistics, the problem of authorship attribution in online conversation and the study goals. Section 2 is a review of available literature on forensic linguistics, stylometry and techniques of authorship analysis. Section 3 presents the research methodology, which entails data collection, preprocessing, feature extraction, and data analysis methods. Section 4 shows the findings of the stylometric analysis and corroborated by providing tables. Section 5 addresses the findings in context to the linguistic individuality and previous researches, as well as limitations. Lastly, Section 6 wraps up the research by summarizing the key findings, implications of the research and future research directions.

## Literature Review

One of the main issues of forensic linguistics is authorship attribution, where the aim is to determine the author of the text on the basis of linguistic evidence. Recent studies put an emphasis on the substitution of manual analysis of linguistic data with computational tools that can be used to carry out a more scalable and objective study (Mani et al., 2025). Data has also been analyzed in the context of authorship profiling on social media and has found that not just identity can be identified by the linguistic patterns, but also geographical and sociolinguistic patterns (Roemling, 2025). In addition, stylometric methods have been shown to be useful in real-world forensic cases, such as analyses of suicide notes when used in case studies (Sinaga, 2025). The notion of linguistic fingerprint also helps to justify the fact that people have regular stylistic patterns, which can be identified (Al-Omari et al., 2024).

Stylometry is the quantitative analysis of the writing style, which became one of the efficient techniques in the field of authorship attribution. It is the analysis of the measurable characteristics of language such as word frequency, sentence length, and sentence structure. Studies have demonstrated that stylometric analysis can be undertaken with the assistance of such tools as R Stylo, which enable the extraction and comparisons of textual features of various authors (Sinaga, 2025). The existence of distinctive writing patterns in even individuals who are genetically identical has also been investigated through quantitative methods, and this implies that stylistic variability is influenced by the effects of cognition and experience (Mohamed, 2025). In addition, punctuation-based modeling is also regarded as a

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

new way of separation of writing styles, but it is concerned with the importance of non-lexical variables in stylometry (Dillon et al., 2025).

Digital communication analysis presents special issues that are not similar to textual analysis. Digital texts have been characterized as short, informal, and flexible, and consequently, might blur systematized styles. Studies of digital interactions, such as the reaction of information services, indicate that user-centered communication could influence the manner in which communication is made, and therefore it is hard to establish who is the author of the communication (Park et al., 2024). Secondly, the realm of linguistic analysis becomes even more problematic by the fact that the multilingual expressions and code-switching can also be applied to the digital world. These points imply that need to have flexible and strong stylometric models in order to be able to process different and dynamic textual information.

Various models and tools capable of aiding in the process of authorship attribution have been developed in a variety of different forms, including machine learning and statistical. Computational stylometrics has been applied to other areas, including literary writing and religious texts, to test authorship claims and identify stylistic inconsistencies (Rosa, 2025). The other emerging trends have also embraced the use of cognitive forensic stylistics, which is a combination of both the psychological and linguistic methodologies of improving the detection of features and reliability (Voice et al., 2025). In addition, stylometric analysis of literary works or historical documents demonstrates that these methods can be used with a great number of genres and contexts (Yang & Lyu, 2025). With these tools, authorship analysis has been significantly enhanced, both regarding accuracy and scalability.

Although these improvements have been made, there are still a few shortcomings to the current research. Many of the works are developed in terms of a limited number of stylometric features, without the integration of lexical, syntactic, and statistical information into one system. Also, there is no consideration of real data of digital communication, which presents certain challenges that cannot be addressed by traditional corpora. The gaps also indicate the need to adopt holistic and context-sensitive solutions on the basis of the integration of different stylometric techniques. Against this deficiency, the present study proposes a quantitative stylometric framework that is particular to online communications to increase the validity and rigor of contentious authorship studies.

## Methodology

Figure 1 demonstrates the overall methodological flow that has been adopted in this paper to determine the controversial authorship in digital communications. This will begin by data collection in different forms such as emails, chat logs, and social media messages and then preprocessing, which will include cleaning of the text, tokenization, and normalization of data to ensure data consistency. The second one is a procedure of stylometric features extraction which comprise lexical features (ex: TTR, MTLD), syntactic feature and frequency of use of function words and n-grams at character level that are all features of the linguistic profile of an author. Python-based tools are then used to process these features to come up with structured feature vectors. Lastly, the comparison of the stylistic patterns is done with the assistance of the most used methods the cosine similarity and the Delta approach to determine who is likely to be the author. This up-down pipeline as Figure 1 illustrates might render the process of quantitative attribution of authorship in forensic linguistics structured and reproducible.

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

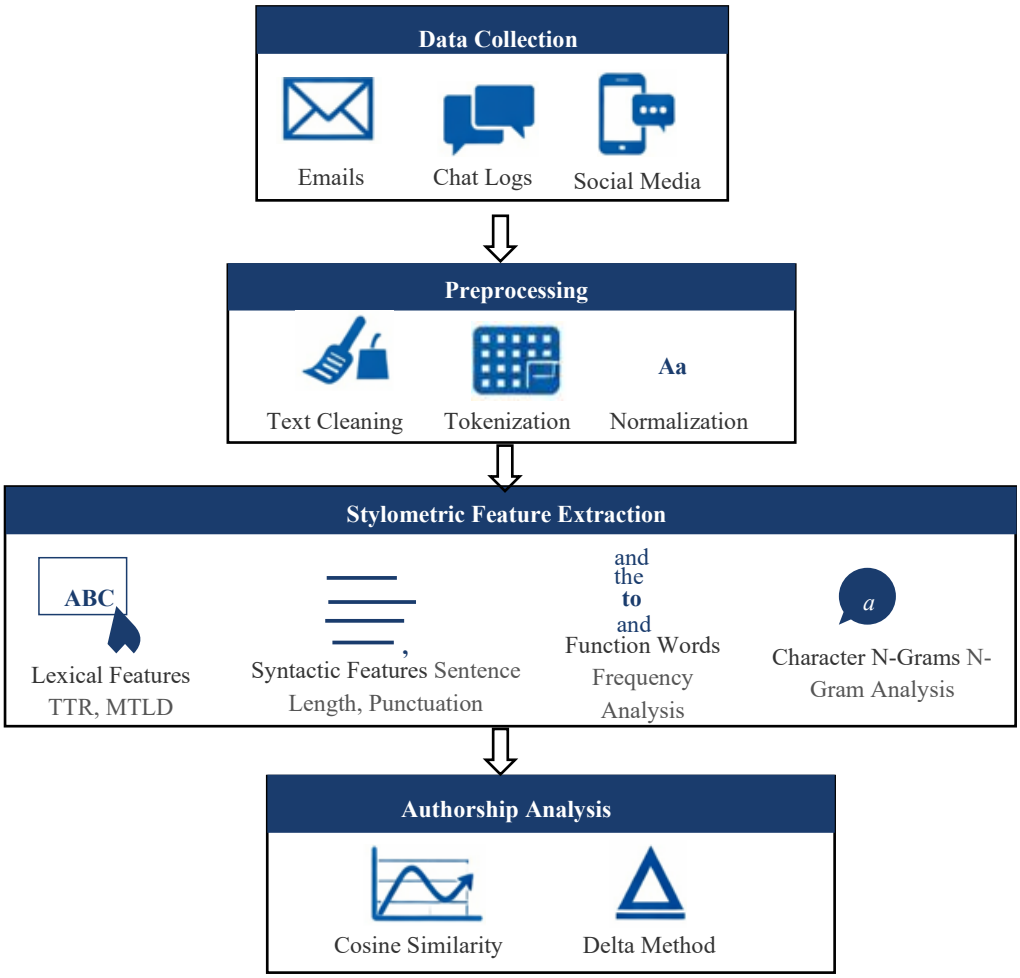


Figure 1: Workflow of Quantitative Stylometric Analysis for Digital Authorship Attribution

**Data Collection**

To obtain the conditions of authorship in the real world the data of this work was collected with the help of different types of the digital communication, e.g., emails, chat messages, messages on the social media, etc. in order to reflect the conditions of the real world. The size of the authors (e.g. 3-5 individuals) was selected; in order to be as eclectic in writing styles as possible and the authors were expected to send the same number of text samples. The size of the aggregate information was to be representative and analytically viable and brief and middle-length texts are the most frequent ones in the Web-based communication. It has tried to incorporate the natural data as compared to some types of data that have been created artificially to ensure authenticity. There were strict ethical rules that were met like the anonymity of personal identifiers, the exclusion of sensitive information and the observance of data privacy rules in such a manner that the dataset could be used in a responsible manner in the forensic linguistic analysis.

**Stylometric Features**

The key about the approach is that it is based on stylometric feature extraction and analysis that signifies personal styles of writing. With the help of lexical features (Type-Token Ratio (TTR) and Measure of Textual Lexical Diversity (MTLD))., the richness of the vocabulary and the variation of vocabulary were estimated. Complex features, such as mean sentence length and punctuation marks allowed learning more

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

about the structure inclinations and writing styles. The frequency distribution was conducted on the functional words (e.g. articles, prepositions and conjunctions) that can be good predictors of the authorial style since these are not consciously used. Additionally, the character-level features were learned using n-grams to capture micro-level style patterns, e.g., spelling and character sequence patterns. The limited number of feature categories enabled the introduction of a linguistic portrait of each author in a multidimensional, holistic way.

### Analytical Techniques

The similarities and differences in authorship were identified using quantitative analysis of the stylometric features obtained. Cosine similarity was used to measure feature similarity between the feature vectors of different authors and to provide a normalized comparison of stylistic feature patterns. A less common but more popular method of computing stylistic distances was also devised, based on the Delta method, a stylometry approach that computes distances between standardised feature values. The approaches allowed finding the most stylistically similar texts between the controversial and known works. The analysis was carried out using different similarity measures, yielding more robust and reliable results for authorship attribution, reducing the likelihood of misclassification, and enhancing the overall quality of the attribution process. The stylometric analysis was implemented in Python and primarily used the Natural Language Toolkit (NLTK) library. NLTK provides simplified preprocessing, tokenization, and feature extraction from text, resulting in an effective framework for working with linguistic data. Numerical computations and manipulations of numerical data were performed using other Python libraries, enabling the calculation of stylometric measures such as TTR and frequency distributions. Python was implemented to ensure reproducibility, scalability, and flexibility in the context of large datasets; hence, it applies to forensic linguistic use of digital text analysis.

## Results

The stylometric analysis shows clear differences in the writing styles of the chosen authors, demonstrating the effectiveness of quantitative methods in authorship attribution. Using lexical, syntactic, and functional feature-extraction techniques, statistical differences were identified in the dataset. Despite these differences, the hypothesis that individual writers share similar linguistic patterns, even in informal, short communications via digital channels, is empirically supported. These features were extracted and systematically compared, as shown in Table 1.

Table 1: Stylometric Feature Comparison

| Feature | Author A | Author B | Author C |
|---|---|---|---|
| Type-Token Ratio (TTR) | 0.62 | 0.48 | 0.55 |
| Avg Sentence Length | 12 | 18 | 14 |
| Function Word Frequency | High | Medium | Low |

As shown in Table 1, Author A has the best type-token ratio (0.62), indicating a richer, more diverse vocabulary than Authors B and C. On the contrary, Author B has the lowest TTR (0.48), indicating a more repetitive lexical usage. Author C is in a moderate level (0.55), which means that the lexical diversity is balanced. Such results demonstrate the significance of lexical properties in author identification, especially in electronic texts, where the vocabulary used by different authors may differ radically.

Syntactically, there were significant differences in mean sentence length. Author B tended to write longer sentences (a mean of 18 words), suggesting a more complex or elaborate writing style. On the other hand, author A employed shorter sentences (an average of 12 words), an indicator of a more to-the-point

ISSN: 2278-4012, Volume:15, Issue:2, June 2026

communication style. Author C again exhibited a middle-ground behavior, with an average sentence length of 14 words. These variation of sentence composition are yet another witness to the stylistic individuality, and this adds to the importance of syntactic analysis to the task of attributing authorship.

The frequency of use of the words was another method by which the frequency of use of function words helped differentiate the writers. Author A has more frequent use of functional words indicating more coherent writing style in the form of grammar. Author B's usage pattern was moderate. In contrast, Author C had a relatively low frequency of function words, suggesting a more content-oriented, less formal manner of communication. The fact that most of the times the use of such words is unconscious means that the distribution of such words is an effective metric of the innate use of the language by the author.

In addition to the personal elements, the syntactic, functional, and lexical features were jointly examined to show the similarity in the style of texts written by the same author. This uniformity was observed even when there were variations in topic, context, and medium of communication. As an example, Author A could afford a high level of lexical diversity and frequent use of the function words in all the samples, but Author B could always resort to the use of long sentence constructions. This consistency in writing style makes stylometric methods more valid for forensic applications.

The observations were further confirmed using similarity measures, such as cosine similarity and the Delta method. Compared feature vectors for each author based on their texts, and found that the texts of a single author showed higher similarity scores than those of different authors. This clustering effect shows that quantitative measures can effectively group texts by stylistic similarity, thereby accurately identifying authors. In disputed cases, the unknown text was always compared to the author with a stylistic profile most similar to it.

Overall, the results suggest that quantitative stylometric analysis is a powerful paradigm for distinguishing authors in digital communication contexts. The differences in vocabulary richness, sentence structure, and the use of function words observed reflect the multidimensionality of the writing style. In addition, the consistency of these features across texts highlights the strength of stylometric techniques in forensic research. These findings indicate that the generalizability of computational linguistics methods for solving authorship problems is greater in contexts where language is informal and dynamic.

## Discussion

The results of this research support the basic assumption of forensic linguistics that each person has a unique linguistic style, commonly known as "linguistic individuality." Stylometric features are efficient at establishing unconscious and habitual aspects of writing, which are difficult to align with time. Lexical indicators (Type-Token Ratio (TTR)) and MTLD are used to determine the richness of the vocabulary, and syntactic indicators reveal what the structure should be (length of the sentence and punctuation). The most powerful indicators amongst them seem to be the function words, which are not used consciously but automatically and thus cannot be consciously changed. The findings in this study are consistent with previous studies, as support the validity of the frequency of use of function words and punctuation in attributing authorship. Additionally, the successful differentiation of authors in small digital texts confirms that, even though the foregoing fears exist, much small data can yield useful stylometric results when quantitative measures are taken. Even though digital communication is quite informal and dynamic, the variability of language use can be controlled when assessed within the context of multidimensional stylometry.

Compared with the previous research, the current analysis is valid and adds to the existing research in stylometry and forensic linguistics. Past literature has shown the successful application of computational tools and statistical models to identify authorship across a wide range of texts, including literary works, social media content, and forensic cases. The contribution of this work is its specificity to the topic of digital communication, given that the problems of brevity, informality, and code-switching are even more pronounced in this context. However, it has certain limitations to be taken into account. The data set is not very large, which could limit the extrapolation of the results; a larger corpus would be more statistically robust. Also, code-mixing and multilingualism, which are typical of digital communication, can introduce variability that is not fully reflected in standard stylometric features. The other restriction is context dependency, in which style of writing may vary according to the issue, the people, or the medium. Despite these restrictions, this paper demonstrates that quantitative stylometric analysis of authorship attribution can be a viable option and can work effectively, particularly when a diverse set of features and analysis tools is applied.

## Conclusion

This paper shows how quantitative stylometric methods can be effective in establishing authorship in controversial digital communication situations. By examining lexical, syntactic, and functional aspects, such as TTR, MTLD, sentence structure, and the frequency of use of functional words, the study demonstrates that there are stable and distinguishable patterns in individual writing style. The results indicate that even short, informal texts typical of Internet sources can provide sufficient linguistic information to support plausible conclusions about authorship, provided the appropriate statistical instruments are used. A mixture of different stylometric features enhances the power and efficiency of the analysis and adds to the concept of linguistic uniqueness in forensic analysis. In practice, the research findings have significant implications across numerous spheres. Stylometric analysis can be used to identify the culprits in cybercrime cases and track cyber threats. It might be implemented in academia to detect plagiarism and verify authorship.

Furthermore, the stylometric evidence could be used in the legal procedures as the objective knowledge of language. Nonetheless, the research notes that it was constrained by the amount of data and the variability of researching with the digital language. The next step in this work can be made with even more advanced applications of artificial intelligence, processing of multilingual data, and deep learning, which will allow for even more precise identification of the author in an even more complex digital world.

## References

Al-Omari, M.D., Elhersh, H., Al Huneety, A., & Mashaqba, B. (2024). Authorship analysis of three Jordanian columnists: is there a linguistic fingerprint?. *Cogent Arts & Humanities*, *11*(1), 2434345. https://doi.org/10.1080/23311983.2024.2434345

Azmi, S.D. (2025). The Intersection of Linguistics and Law: Forensic Language Analysis in Judicial Practice. *Archipel: Journal of Indonesian Interdisciplinary Studies*, *1*(1), 17-23. https://doi.org/10.65739/archipel.v1i1.5

Dillon, R., Gotelli, M., & Bruzzone, A. (2025). Experimental Modeling of Writing Styles for Authorship Verification via Punctuation Analysis. *Procedia Computer Science, 274,* 1238-1243. https://doi.org/10.1016/j.procs.2025.12.122

Helmi, R., Maulina, S., & Mardhatillah, F. (2025). Forensic linguistics in legal contexts: Examining plagiarism, legal documents, and defamation. *Beyond Words*, *13*(1), 27-41. https://doi.org/10.1177/21582440251334276

Mani, R.T., Palimar, V., Pai, M.S., Shwetha, T.S., & Krishnan, M.N. (2025). An evolution of forensic linguistics: From manual analysis to machine learning–A narrative review. *Forensic Science International: Reports*, *11*, 100417. https://doi.org/10.1016/j.fsir.2025.100417

Mohamed, E. (2025). Do Identical Twins Write Identically? Evidence from Authorship Attribution. *Journal of Quantitative Linguistics,* 1-31. https://doi.org/10.1080/09296174.2025.2568210

Park, J.R., Poole, E., & Li, J. (2024). Stylometric features in librarian's responses to user queries: implications for user interaction in digital information services. *Global knowledge, memory and communication*, *73*(3), 375-390. https://doi.org/10.1108/GKMC-03-2022-0055

Roemling, D. (2025). Forensic Authorship Profiling Using Geolocated Social Media Data: A Corpus Linguistic and Cartographic Approach. *Applied Corpus Linguistics*, 100146. https://doi.org/10.1016/j.acorp.2025.100146

Rosa, A. (2025). Computational Stylometrics and the Pauline Corpus: Limits in Authorship Attribution. *Religions*, *16*(10), 1264. https://doi.org/10.3390/rel16101264

Sinaga, T.F. (2025). A forensic linguistic investigation of Mahira's suicide note using stylometric analysis in R Stylo. *Langkawi: Journal of The Association for Arabic and English*, *11*(1), 160-176. https://doi.org/10.31332/lkw.v11i1.11838

Voice, M., Harrison, C., Grant, T., & Giovanelli, M. (2025). Towards a cognitive forensic stylistics: An intercoder reliability test for replicable feature finding in the Operation Heron corpus. *Language and Literature*, *34*(4), 327-348. https://doi.org/10.1177/09639470251337632

Yang, Y., & Lyu, G. (2025). A Stylometric Analysis on Authorship of Quelling the Demons' Revolt. *Sage Open*, *15*(4), 21582440251405558. https://doi.org/10.1177/21582440251405558