Linguistik Indonesia, Februari 2024, 91-112 *Volume ke-42, No.1*
Copyright©2024, Masyarakat Linguistik Indonesia
ISSN cetak 0215-4846; ISSN online 2580-2429

AUTHORSHIP ANALYSIS IN ELECTRONIC TEXTS USING SIMILARITY COMPARISON METHOD

Devi Ambarwati Puspitasari[1], Adi Sutrisno[2], Hanif Fakhrurroja[3]
Badan Riset dan Inovasi Nasional[1,3], Universitas Gadjah Mada [2]
devi018@brin.go.id[1], adisutrisno@ugm.ac.id[2], hani010@brin.go.id[3]

Abstract

The most recent changes to the criteria in legal process for scientific evidence have emphasized scientific methods of authorship analysis. This study examined the authorship of electronic texts using a quantitative method based on forensic stylistics and computer technologies. This study uses 300 digital texts produced by 100 authors, including 100 questioned texts (Q-text) and 200 known texts (K-text). Personal texts of WhatsApp messages are used in this study as electronic texts. Authorship analysis was conducted by tracing the n-gram and testing all the text sets using the Similarity Comparison Method (SCM). Based on the results of the word 1-gram test, the SCM accuracy was found to be quite high, ranging from 85% to 96%. The findings of employing the tiny set are promising, with the various stylistic traits offering dependable accuracy ranging from 92% to 98.5%. The character-level n-gram tracing indicates a key feature of authorship attribution.

Keywords: authorship analysis, electronic texts, forensic stylistics, WhatsApp chat

Abstrak

Kriteria proses hukum dalam pembuktian ilmiah telah mengalami perubahan dengan menekankan penggunaan metode ilmiah dalam analisis kepenulisan. Penelitian ini mengkaji kepenulisan (authorship) pada teks elektronik dengan menggunakan metode kuantitatif berdasarkan stilistika forensik dan teknologi komputasi. Penelitian ini menggunakan 300 teks digital dari 100 penulis, yang meliputi 100 teks yang dipertanyakan kepenulisannya/anonim (Q-text) dan 200 teks yang diketahui profil kepenulisannya (K-text). Teks elektronik yang dianalisis dalam penelitian ini adalah teks-teks pribadi dalam bentuk pesan WhatsApp. Analisis kepenulisan dilakukan dengan menelusuri n-gram dan menguji seluruh kumpulan teks dengan menggunakan Metode Perbandingan Similaritas (Similarity Comparison Method/SCM). Berdasarkan hasil uji 1-gram pada level kata, SCM menunjukkan hasil dengan akurasi yang cukup tinggi, berkisar antara 85% hingga 96%. Penggunaan kumpulan teks pendek pada penelitian ini menunjukkan hasil identifikasi kepenulisan dengan akurasi yang dapat diandalkan, yaitu sekitar 92% hingga 98,5%. Selain itu, penelusuran n-gram pada tingkat karakter terbukti andal dalam mengidentifikasi fitur penting atribusi kepenulisan.

***Kata Kunci**: analisis kepenulisan, teks elektronik, stilistika forensik, obrolan WhatsApp*

INTRODUCTION

After the Covid-19 pandemic, the number of legal cases relating to Indonesia's Law on Electronic Information and Transaction (UU ITE) has not altered much. A thousand cases of infractions of the UU ITE were registered in the Supreme Court of the Republic of Indonesia's Directory. The majority of these infractions were related to hacking and falsifying papers. It is intriguing that multiple cases were uncovered with evidence pointing to text fabrication and authorship. Proofing authorship disagreement cases in Indonesia have not reached an authorship analysis due to the difficulty of establishing personal identification in electronic documents, particularly short texts with restricted characters and words.

Author detection in anonymous texts, particularly brief electronic communications, is accomplished by matching a person's writing style to that of one of the detected writers (Coulthard, 2004). This problem has been addressed using a variety of solutions with varying features. One of the most prevalent techniques is to use stylistics to investigate the author's writing style (Aziz, 2021). This intricate tactic emphasizes identifying contemporary facts in the text and using subjective inspection tools (Tarrayo, 2020). Meanwhile, another commonly used method is the quantitative method, which evaluates the extraction of numerous statistical aspects (McMenamin, 2019). Whenever faced with a short text and the demand for proof that rejects relative aspects, qualitative approaches are seen as inadequately substantial and scientific, despite stylistics being regarded as an excellent method for resolving the question of authorship authentication (Bailey, 2000; Snee, 2016).

In recent years, various authorship analysis investigations have been conducted in computational linguistics and artificial intelligence, particularly machine learning and natural language processing (Alshammari & Alanazi, 2021; Casillas & Ramirez, 2019). As a result of this collaboration, machines can analyze texts in accordance with the linguistic framework, give statistical significance, and provide arguments based on extensive and scientifically precise data (Theophilo et al., 2021).

The scientific method's heuristic requirements have been stressed as a criterion of scientific evidence in the judicial process (Aziz, 2021). Because of several legal case testimony (Grant, 2007), the linguists' arguments were regarded as lacking any scientific foundation. Furthermore, the situation in which a scientific approach is challenged with another scientific method that leads to a result is the most common in establishing legal matters using language (McMenamin, 2019). When the outcomes of the methods are compared, relative results are produced. Language demonstrations often differ from one approach to the next and from one expert to the next. To avoid this relativity, language-related evidence must now provide a measurement with a high degree of certainty.

Though linguistic methods of assessing evidence are viewed as relative (McMenamin, 2019), they nonetheless demonstrate sufficient rigor to create objective facts, dependable outcomes, and accurate conclusions that are indicative of quantitative analysis. The primary drawback of qualitative evaluation is that it cannot make absolute decisions based on the evidence, especially in circumstances involving genuine examination. Quantitative approaches, on the other hand, are possible to demonstrate the author's identity with great certainty due to their detailed measurements and computations (Gorsuch, 2009; Ikeo, 2008). By using quantitative methodologies, authorship analysis can be predicated on assertions of method correctness. Furthermore, the conclusions of the investigation can be communicated using statistical data, exact measurement results, and definitive declarations, such as claims against the author's identity. This quantitative technique in the evaluation of forensic stylistics is referred to by a few experts (Coulthard, 2004; McMenamin, 2022). The forensic stylistic method of determining the comparative text (known text/KT) as a reference to the text of existing evidence is the first step in determining the authorship of a case. This began with identifying the text's subject or owner based on the circumstances of the current case or references to police investigations' outcomes.

In the case of Jenny Nicholl, for instance, if the questioned text (QT) is a text message (SMS) (Grant & Baker, 2007), then, at that point, the nearest reference as a known text (K-text) would be messages and steady private messages from the same author. The term "questioned text (QT)" refers to a text whose authorship is unknown or whose authorship is suspect. Meanwhile, the term "comparative texts or known text (KT)" are texts whose owner or author is clearly identified, and which serve as references for examining their authorship since they are thought to be the closest in authorship to the text under consideration.

The style of Indonesian authors is distinct in their written works (Puspitasari, 2021, 2022; Puspitasari & Sukma, 2022). In addition to characters, other linguistic units such as words and phrases will create a very personal style of writing (McMenamin, 2019). Personal text in Indonesia may be influenced by local languages and online word trends. In Indonesia, digital forensic research was unable to provide substantive confirmation of text authorship claims. It is impossible to establish authorship just based on device ownership and location. Forensic linguistic analysis is required in cases of authorship disputes in Indonesia to make an accurate claim of authorship of a text. However, unlike digital forensics, the subsequent study of forensic linguistics in Indonesia has been unable to generate proof of authorship analysis with limited legal application.

Considering the interest in straightforward logical proof in measured phonetic examination and proof in the legitimate cycle, this study conducted a more quantitative investigation of authorship analysis of Indonesian electronic texts. The work in forensic stylistics serves as the basis for the quantitative method (Eder et al., 2016; Neme et al., 2015), specifically using computational technology to extract linguistic features from text and carry out statistical tests (Anwar et al., 2019; Belvisi et al., 2020a; Frye & Wilson, 2018). Due to the demand for absolute scientific evidence in forensic linguistic analysis and evidence in the legal process, this study conducts an analysis of the authorship of electronic texts using n-gram tracing and stylometric features in recognizing Indonesian authorship attribution. This study conducted several statistical tests using the similarity comparison method by building a statistical test app employing the JC (Jaccard Coefficient) and TF-IDF (term frequency–inverse document frequency) formula. Therefore, the research questions in this study are formulated as follows:

(1) Which lingual forms in Indonesian electronic texts can be identified as authorship attribution as a guideline in the authorship analysis?
(2) What is the accuracy of the two similarity comparison method (i.e., JC and TF-IDF)?

Authorship Identification

A person's writing or speaking style is referred to as language style. The language style is a set of lexical, syntactical, and character qualities that are generally communicated through accents in speech. Each person's language style distinguishes them (Fobbe, 2020). The selection of linguistic traits, according to Grant (2007), is more of a tendency than a law. Linguistic application is not fixed; otherwise, identical features will continue to emerge. However, not every feature may appear in every paragraph. Many sample texts or lengthy texts are required for authorship analysis, both of which are sadly uncommon in forensic investigations. Because no reliable and comprehensive language style markers database has been identified thus far, there is no such thing as a semantic unique finger imprint (Brennan et al., 2012; Ison, 2020).

According to Grant (2007), the selection of linguistic features is more of a tendency than a law. Linguistic application is not fixed; otherwise, similar traits will arise. Not every feature, however, may occur in every paragraph. For authorship analysis, several sample texts or lengthy texts are necessary, both of which are tragically uncommon in forensic investigations. There is no such thing as a semantic unique finger imprint because no credible and comprehensive language style markers database has been established thus far.

N-gram Tracing

In a homicide investigation, a series of instant messages was connected through an exact spelling analysis (Chiang, 2021; Coulthard, 2013). Nini (2018) looked at the word request shared to see how similar the short letters related to the Jack the Ripper case were. N-gram tracing, a novel quantitative method for attribution of authorship to short texts, was associated with the 139-word Bixby Letter (Grieve et al., 2019). As initially proposed by Patodkar & I.R (2016), Award (2007) and Nini (2018), the answer to the issue of examining short texts regarding scientific semantics is to consider no elements as opposed to word recurrence. This method measures the similarities between two texts by dividing the number of features in each text by the total number of features in both texts using Jaccard's coefficients. The Jaccard coefficient is a measure of similarity between two sets. It is also known as the Jaccard similarity coefficient or index. It is defined as the size of the sets' intersection divided by the size of their union (complete formulas and explanations are in the methods section). The Jaccard coefficient is a useful tool in authorship analysis because of its simplicity and efficacy in capturing set-based similarity. It assists researchers in quantifying the degree of linguistic similarity or overlap across authors, hence assisting in the identification of distinct linguistic patterns and contributing to the larger subject of computational stylistics.

The comparative text (known text/KT) as a reference to the text of existing evidence is determined as the first stage in investigating the authorship of a case utilizing the forensic stylistic method, specifically n-gram tracing (McMenamin, 2019). This began with distinguishing the text's subject or proprietor based on the circumstances of the ongoing case or allusions to the results of police examinations. For example, if the questioned text (Q-Text) is an instant message (SMS) from Jenny Nicholl, a homicide case in 2005 in which a few phony instant messages were discovered as evidence, the messages and consistent private messages from a similar proprietor would be the closest reference as known message (K-text). The style variations in the Q-text-labeled corpora were evaluated using the whole range of stylistic variations, including variants and invariant forms of each variable, starting with the lowest n-unit. According to McMenamin (2019), the kinds mentioned in the text can be found, appear, and disappear. The n-gram that follows may use the two levels of n-units for further in-depth analysis: both the word-level and the character-level, which includes non-alphabetic characters.

An n-gram addresses a series of n components in a text that are near to one another. The elements can be any combination of characters, words, symbols, syllables, and so on. For example, the sentence "the rose is red" would generate a vector of n-grams of the form: [(the,rose),(rose,is),(is,red)] when processed at the word level with n=2. Following that, the overall frequency of each n-gram element in the text is determined, and the resulting values are utilized to generate a vector graphic representation of the text. The popularity of this feature can be due to its scalability and language independence. It has been chosen for research in languages other than English. Several research have attempted to identify the value that should be allocated to n in order to successfully represent an author's style, with trials suggesting that as n increases, accuracy improves, but not much after 5 (Belvisi et al., 2020b; Grieve et al., 2019; Nini, 2018). This paper will test both the word-level and the character-level, with n ranging from 1 to 4. Because n-grams are not confined to letters, they were chosen for their ability to manage text length, misspellings, language variances, and the presence of other symbols such as emojis or punctuation.

Stylometric Features

To assess uniqueness, both stylometry and N-gram tracing are used. Stylometry investigates how writers organize words and phrases, and how they use punctuation or paragraph structure. A text's properties must be studied in five major categories: lexical, structural, content-explicit, syntactic, and idiosyncratic. Lexical features are a group of images and words that all fulfill the same purpose. These criteria include capital letter distribution, special characters, average word usage, the number of words in a sentence, and other features. The splendor of an author's words is portrayed here.

The structure of a text, such as the average length or number of paragraphs and sentences, reveals how the author organizes the contents of the text. Indicators in this study include whether the author adds greetings and farewells in an email corpus. Content-specific refers to the frequency of keywords in the text. This classification is particularly relevant for a corpus derived from gatherings or other specified point sources. Although the features are especially useful for monitoring content like terrorism and cyber-pedophilia, they are worthless in a more general setting like Twitter communications because they are topic and environment dependent.

Syntactic features, such as punctuation and function words, focus on the syntax of the text. The words that define the relationships between pieces of a sentence are known as function words. As a result, they are also the most frequently used words in any work. Unfortunately, due to the length of the text, these properties do not considerably help in expressing such writings. Finally, quirky traits highlight attention-grabbing elements that are unique to the author. Emojis, misspelled words, unusual characters, and abbreviations are examples of such features.

Application of Forensic Stylistic Analysis Using the Similarity Comparison Method

Before starting the analysis, both qualitative and quantitative, all data regarding questioned text (Q-text) and all available references (known text/K-text) will be collected. References are selected based on the context of the relationship that is considered closest to the writings in the Q-text or based on police recommendations. Once analysis begins, tools for counting n-grams, statistics on duplication of concordances, and other quantitative work may be needed (Grieve et al., 2019; McMenamin, 2019; Nini, 2018).

The corpora with Q-text labels were examined and measured by the range of stylistic variations, including the variants and invariant forms of each variable, starting with the smallest n-unit. Variations identified in the text can appear and be traced, but can also not be found (McMenamin, 2019). The occurrence of a single variation, as well as variations that occur more frequently or represent repeated habits, was recorded. A series of tests of similarities between Q-text and K-texts should be conducted to examine and record the similarities or differences between texts (Grant & Baker, 2007; Peng et al., 2016). This study calculated the constituent index from n-1 to n-4 on the character-level and word-level based on the JC and TF-IDF formula (Nini, 2018; Permatasari et al., 2020).

Jaccard Coefficient calculates the features between two texts, both double and unique, which are then divided by the total number of features in both texts. Comparison of Q-text and K-text will produce J values with a range of 0–1. If the results of J get closer to 1, it can be concluded that the two texts being compared were written by the same author. Regarding word selection, as is generally the case with stylistic studies on lexical elements, statistical tests can also be added to the frequency distribution test, namely the TF-IDF. The choice of words in the Q-text and K-text will determine the identification of the authors of the two writings. By looking at the frequency distribution, the registers of the two authors will be seen and the similarities and differences will be calculated. TF-IDF (Term Frequency-Inverse Document Frequency) plays an important part in authorship attribution by allowing you to weight terms in a document depending on how important they are in distinguishing that document from others. TF-IDF is able to calculate the frequency distribution of the Q-text with any comparison writing with a weighted value. The highest value of the calculation results is the text with the most frequency distribution of the Q-text registers.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

Information:
A: Jaccard index
A: coefficient A
B: B coefficient
∪: union
∩: intersection
A ∩ B: number of intersections of coefficients A and B
A ∪ B: sum of coefficients A and B

Figure 1. Jaccard Coefficient Formula

$$TF(t,d) = \frac{number\ of\ times\ t\ appears\ in\ d}{total\ number\ of\ terms\ in\ d}$$

$$IDF(t) = log\frac{N}{1+df}$$

$$TF - IDF(t,d) = TF(t,d) * IDF(t)$$

Figure 2. TF-IDF Formula

Table 1. Ilustration How Jaccard Coefficient and TF-IDF Work in the Context of Authorship Analysis

<table>
<tr><th>Method/ formula</th><th>Scenario</th><th>Tokenization</th><th>Result of calculation</th><th>Interpretation</th></tr>
<tr><td>Jaccard Coefficient</td><td>Suppose we have two sentences from two different authors:

Author A: "The cat is on the mat."
Author B: "A cat is sitting on the rug."</td><td>Author A: {"The", "cat", "is", "on", "the", "mat"}

Author B: {"A", "cat", "is", "sitting", "on", "the", "rug"}</td><td>Jaccard Index = 0,4</td><td>The Jaccard Coefficient of 0.4 indicates a moderate level of similarity between the two sentences.

The overlap of words is "cat," "is," "on," and "the," which contributes to the similarity.</td></tr>
<tr><td>TF-IDF</td><td>Consider three documents from three different authors:

Author X: "The sun is shining brightly."

Author Y: "The moon is shining brightly."

Author Z: "The sun and the moon are both shining."</td><td>Calculate the TF-IDF weights for each term in each document.

<table><tr><th>Term</th><th>Author X</th><th>Author Y</th><th>Author Z</th></tr><tr><td>The</td><td>0.405</td><td>0.405</td><td>0.405</td></tr><tr><td>sun</td><td>0.405</td><td>0.405</td><td>0.405</td></tr><tr><td>moon</td><td>0</td><td>0.405</td><td>0</td></tr><tr><td>shining</td><td>0.405</td><td>0.405</td><td>0.405</td></tr><tr><td>brightly</td><td>0.405</td><td>0.405</td><td>0.405</td></tr><tr><td>and</td><td>0</td><td>0</td><td>0.405</td></tr><tr><td>are</td><td>0</td><td>0</td><td>0.405</td></tr><tr><td>both</td><td>0</td><td>0</td><td>0.405</td></tr></table></td><td>Author X: [0.405, 0.405, 0, 0.405, 0.405, 0, 0, 0]

Author Y: [0.405, 0.405, 0.405, 0.405, 0.405, 0, 0, 0]

Author Z: [0.405, 0.405, 0, 0.405, 0.405, 0.405, 0.405, 0.405]</td><td>The vectors highlight the importance of terms within each document.

Author Z has the highest similarity weight</td></tr>
</table>

Table 1 shows concrete examples to illustrate how both Jaccard Coefficient and TF-IDF work in the context of authorship analysis. These examples show how the Jaccard Coefficient catches word overlap and how the TF-IDF allocates weights to terms, emphasizing unique words for each author. In practice, these strategies are employed in tandem to examine and compare the linguistic characteristics of texts in order to identify authorship.

The results of the calculations will become findings and support the basis of claims or conclusions drawn on the identification of authors in the analysis of authorship. Based on the working concepts of JC and TF-IDF, this research builds software to carry out the tokenization process and calculate similarity. This software development employs Python programming, which is intended to process text, as well as authorship analysis features. This program was created with the help of the BRIN research team and funds. We name it Text Curation Engine V.1, and it will be in development for the next two years. Readers with questions about this program can contact the author at the email address written on the first page of this article. Regarding drawing conclusions, McMenamin (2001) presents an example of nine levels of authorship of a text which can be used as criteria from the results of statistical tests.

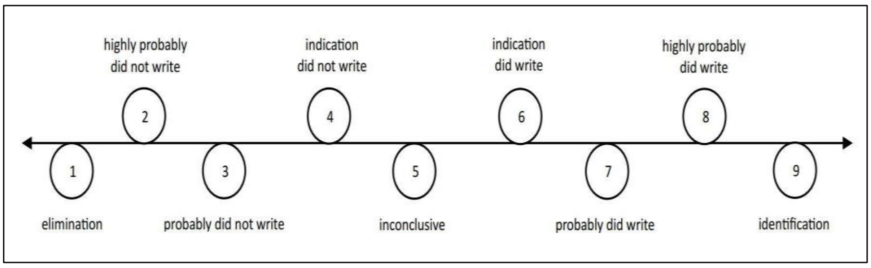

Figure 3. Authorship Level (McMenamin, 2019)

To determine conclusions about authorship claims, McMenamin (2001) made a gradation of the levels of authorship analysis results (see Figure 3). The gradation can be applied as a scale to the range of results of a statistical test. For example, if the result of a statistical test is on a scale of nine, then the text can be concluded that the author has been identified. The value or degree of identity will decrease according to the scale. This will help in determining decisions or claims against a text.

METHOD

The development of digital research methods (DRM) began in 1990 and has become a new approach to social humanities research, adopted and accepted as the academic standard of 100 universities with global repute (Snee, 2016). The concept is a research method that utilizes online media and digital technology, such as big data, online forms, digital text, and voice recognition to support research activities (Snee, 2016). DRM utilizes digital technology to minimize human intervention thereby increasing accuracy, and speed in data collection and processing, and reducing the risk of errors (Rifai, 2020). DRM functions to bridge the interaction between researchers and research objects like conventional methods. In a socio-cultural context, digital data helps identify human behavior and interactions through digital ethnography (Rheingold, 2000). Digital data tends to be well-recorded, dynamic and has a wider scope (Rheingold, 2000; Takwin, 2020; Unik & Larenda, 2019).

This study adapts the DRM by using a mixed method, the statistical, quantitative approach combined with a qualitative descriptive approach to describe the linguistic feature in the authorship analysis. The data used in this study consists of four types, namely (1) characters, consisting of letters, numbers, punctuation marks, and emojis; (2) words; (3) phrases; and (4) sentences to be counted in units of n-grams. N-grams is a contiguous sequence of n items from a given sample of text or speech (Baker, 2010; Baker et al., 2008; Mautner, 2009; Rebuschat et al., 2017). The items can be phonemes, syllables, letters, words, or base pairs according to the application (Baker, 2006). The n-grams typically are collected from a text or speech corpus.

The data in this study are 300 electronic texts of 100 unique authors, consisting of 100 questioned texts (Q-text) and 200 known texts (K-text), comprising 63,414 tokens with a total frequency of 1.9 million words. Electronic texts in this study are personal texts, such as WhatsApp messages, SMS, personal emails, tweets, and Instagram and Facebook posts. This study has received ethical clearance to collect individuals willing to contribute their own texts to the study.

Tabel 2. Author Data Summary

No.	Author	Information	Total (%)
1.	Gender	Male	46%
		Female	54%
2.	Age	14-20 years old	41%
		20-30 years old	23%
		30-40 years old	31%
		40-50 years old	5%
3.	Circle of relationships between respondents (authors)	Individual	45%
		Family	10%
		Friends/community	45%
4.	Source/type of electronic text	Tweet	11%
		Instagram & FB posts	14%
		Whatsapp messages/chats	64%
		E-mail	7%
		SMS	4%
5.	Origin of respondents (authors)	The eastern area of Indonesia (*Indonesia bagian timur*)	30%
		The middle area of Indonesia (*Indonesia bagian tengah*)	30%
		The western area of Indonesia (*Indonesia bagian barat*)	40%

Jaccard Coefficient dan TF-IDF was employed to calculate the features between two texts, both double and unique. The features include the use of characters (capital letter distribution and special characters), average word usage, the number of words in a sentence, and idiosyncratic feature. And regarding the choice of words as a stylistic guideline, statistical tests can also be added to the frequency distribution test, namely TF-IDF. Both calculations will support the basic claim on the identification of the author of the Q-text in the authorship analysis.

A Profile-Based Approach (Belvisi et al., 2020) in the context of authorship analysis as shown in Figure 4, entails creating separate linguistic profiles for each writer based on numerous linguistic traits. Word usage, sentence structure, vocabulary richness, and other stylometric attributes may be included. The fundamental purpose is to generate a unique fingerprint or profile for each author, allowing their writing style to be identified. This research adapted this approach for managing the statistical test in authorship analysis to leverage linguistic profiles to differentiate authors based on their writing styles.

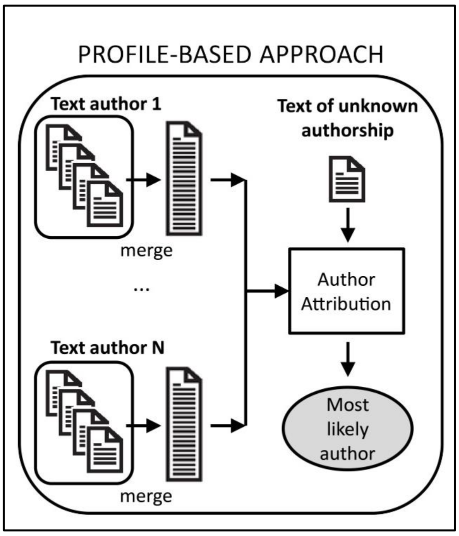

Figure 4. Approaches to Investigate the Set of Documents of the Author (Belvisi et al., 2020)

All features are represented as a vocabulary—a set of words that is unique to each author. The intersection of sets of different authors is computed to compare them. The bigger the intersection, the more similar the two sets are. In addition to stylometric traits, n-grams are distinguished because the writing demonstrates their proficiency regardless of text length or setting. This work employs character and word n-grams with n ranging from 1 to 4.

DISCUSSION

Results of N-gram Analysis in Authorship Identification

The frequency and dominance of individual patterns created by an author can be determined through N-gram calculation. Each author's text set is treated as a distinct corpus of authors when n-grams are calculated. There were 100 authors with their own text set in this study. The author's characteristic or uniqueness was revealed by the n-gram calculation results in each corpus.

To begin, the character-level n-units used in this study ranged from n-1 to n-4. The average number of characters used by the authors is used to select the range. For instance, a few authors will generally utilize one accentuation mark: one space, one period, a comma, a question mark, and one space. Some authors use two periods, two commas, two question marks, and other punctuation characters in the same order. This should be visible in the accompanying model.

Characteristics of characters usage based on the number also affect the n-unit of the alphabet with the smallest range of n-1 (Table 3). An author consistently uses the character '*y*' as a substitute for the word *iya* 'yes', the character '*q*' as a substitute for the word *aku* 'I, me, my', and the character '*g*' as a substitute for the word *enggak* 'no'. In n-2 it is also common to find authors consistently using two characters as a substitute for a word, for example, the character *ya* as a substitute for the word *iya* 'yes', the character *ga* as a substitute for the word *enggak* 'no', the character *aq* as a substitute for the word *aku* 'I, me, my', and others. This is shown in Table 3.

Table 3. Examples of N-Units of Non-Alphabetic Character-Level

Token	n-gram	Examples
..	n-2	**Data 1: Korpus 39.PT** Teks 1: *dia berlari dari lorong sampai di cegat pramugari karna sudah tidak bisa turun***..** [Text 1: he ran from the hallway until he was intercepted by a flight attendant because he couldn't get off anymore**..**] Teks 2: *dan semakin sepi***..** *pesawat sudah sangat siap***..** *tapi ebel belum datang***..** [Text 2: and it is getting quieter**..** the plane is very ready**..** but Ebel hasn't arrived yet**..**]
!!	n-2	**Data 2: Korpus 45.BB** Teks 1: *Iya dong***!!** [Text 1: Yes, please**!!**] Teks 2: *Gw wisuda 2013 anzeeengg***!!** [Text 2: I graduated 2013 anzeeegg**!!**]

Table 4. Examples of N-Units of the Alphabet that Consist of 1-2 Characters

Token	n-gram	Found in…Text Set/s	Frequency (in Corpora/ all authors text)
q	n-1	5	176
aq	n-2	1	7
g	n-1	2	3
ga	n-2	17	136
y	n-1	1	4
ya	n-2	33	172

The data also revealed that authors also tend to have rich writing styles with a large number of n-units, namely n-3 and more. The n-units are punctuation characters, letters of the alphabet, or a combination of both. Based on the data findings, it has been found that authors tend to use the letters 'w' and 'k' to express the emotion of laughter, but the number of characters is consistent. In a different way, some authors tend to use a combination of letters ‘h’ and ‘a’ or ‘e’, sometimes followed by ‘u’ for Sundanese authors. The examples are as follows.

Table 5. Examples of Character N-units

Token	n-gram	Frequency in Text Set	Frequency in Corpora
wkk	n-3	4	12
wkwk	n-4	3	15
wkwkwkk	n-7	3	9
wkwkwkwkk	n-9	2	1
haha	n-4	4	13
haha..	n-6	3	7
heuheu	n-6	4	2
heuheu..	n-8	2	2

In the n-unit of character-level that are compared between authors, each n-unit also has different variations even though they have the same meaning. For example, the word *aku* was found in five different variations across the tokens. The five variations are *aq*, *ak*, *aku*, *qu*, and *q*. However, each author uses each variant consistently. According to McMenamin (2019), findings regarding the author's uniqueness at the n-unit character level can be used as a marker of authorship attribution. This is the author's choice and a part of the authorship style, given the variety and number of characters. According to Eckert (1989), an author's unique set of grammatical patterns, which are typically the result of the author's habitual usage or repetition in some or all his/her writing collections, reveal the author's style. N-units in character-level also include word choice, because the author chose words and wrote it in a very specific and patterned way. For example, the word *aku* 'I/me/my' is written in five variants, *saya* 'I/me/my' is written in three variants, and *gue* 'I/me/my' is written in four variants as (see Table 6).

The word *saya* 'I, me, my′ is a word that is commonly used by the authors in 19 texts set with a frequency of 52 tokens. It means that 19 authors chose *saya* as their style to write a personal pronoun. The same thing can be seen in the word *aku* 'I, me, my' which is found in 14 corpora (set-texts of 14 authors), meaning that 14 authors have the same characteristics. Similarities are also found between authors: the use of n-1 word of personal pronoun is found in more than one text set. This means that several authors have the same choice and writing style of words. It is not surprising that many authors could use this equally because *saya* and *aku* are commonly used personal pronoun in Indonesian. In this case, another attribution is needed to distinguish these authors. This can be done by continuing to calculate the n-grams up to the largest n-unit, both at the character-level, as well as the word-level.

In contrast to the words *aq*, *aqu*, *sya*, *guweh* which are found only in one text set. This can immediately be a sign of the uniqueness of an author, where the choice of words and the way he/she wrote is very distinctive and not the same as other authors. This is in line with what Coulthard (2013) said about using n-gram analysis to find the author's uniqueness. An n-gram is a unit of examination characterized as a grouping of at least one etymological structure at each degree of phonetic investigation, like words and characters. According to Coulthard (2013), comparing two sets of text can reveal similarities as well as distinguishing characteristics when determining authorship evidence.

Table 6. Word Variants of *aku*, *saya*, and *gue*

Word Choice	n-gram	Token Variant	Found in… Text Set/s	Frequency (in Corpora)
aku [I/me/my]	n-1	*q*	5	176
	n-2	*ak*	5	9
		aq	1	7
	n-3	*aku*	14	145
		aqu	1	6
saya [I/me/my]	n-2	*sy*	2	5
	n-3	*sya*	1	8
	n-4	*saya*	19	52
gue [I/me/my]	n-2	*gw*	2	4
	n-3	*gue*	6	22
		gua	3	16
	n-5	*guweh*	1	3

Variant tracking can be achieved by analyzing the frequency of words in a set of texts by an author or between several authors that are being compared. An author can also be very specific using certain words that are influenced by the regional language or social environment. For example, the words *ane*, I, *gue*, and *ulun* are found in the corpora of this research. Based on the author's data, the author from Jakarta chose the word *ane*, as well as the word *gue*. The choice of 'I' word was written by Chinese Indonesian author and the word *ulun* was written by an author from Banjarmasin (a city in Borneo Island). More specifically, each of these words is found in only one corpus, except the word *gue* which is found in six text sets. Considering that many of the authors' data in this study come from or live in Jakarta, it is normal that many authors chose to use the word *gue*. In this case, the six authors with the choice of the word *gue* still need further investigation regarding their uniqueness, while the author with the choice of the words *ane*, I, and *ulun* can immediately have their uniqueness.

Purposive sampling was used to select the authors for this study, which illustrates the different cities or regions of origin for each author. The purpose is to find regional attribution markers. These findings suggest that regional factors will influence vocabulary choice. Text data indicates the author may have spoken with others in the same area, using precise personal pronouns. For various reasons, a writer may also choose personal pronouns that he or she does not normally employ. However, each writer's corpus comprises of various texts drawn from several times and from multiple intended recipients of the messages. By looking at word frequency, we may see the author's usage of personal pronouns as attribution.

An author's usage of language will become standard in a set of documents. One of the standards of language conduct in a public or gathering is a territorial standard connected with topographical area (McMenamin, 2019). The author's upbringing or place of residence have a significant impact on the words used. Three of the six authors who chose the word *gue* in their text set were not of Betawi ethnicity or from Jakarta. One of them is Javanese and the rest are from Bandung, but all three live in Jakarta. These authors are influenced by the environment in which they live and their social interaction. In this case, they violate the language norm, but the choice of words becomes a characteristic of their writing.

Notwithstanding the selection of words, the author's qualities can likewise be seen from how they change words, truncate plural words, and utilization of particles. Some authors typically rewrite words followed by numbers or quotation marks at the word-level of n-1, for example, *itu2* 'that', *baca2* 'read', as shown in data (1) and (2). Some authors consistently abbreviate the word *yang* ‘which' into *yg* and consistently use certain particles such as *dah*, *ae*, *atuh*, and others, as shown in the data (3), (4), and (5).

(1) *timeline stop* ***itu2*** *aja, buka stori pun tak bisa, refresh pun sama.*
[just stop the timeline, can't even open the story, refresh is the same.]

(2) *JPU baca dakwaan, kenapa terbayangnya lagi* ***baca2*** *EPISTEL ya*
[The prosecutor read the indictment, why is in my imagination he was **reading** EPISTEL?]

(3) *Kesian kang ayam* ***yg*** *mompa angin ke ayam nya biar keliatan menarik.*
[It is a pitty to see the chicken seller **who** blows wind on it to make it look attractive.]

(4) *da* ***yg*** *namanya overmacht* ***atuh*** *bapak Wuakakakakaka*
[there is something called overmacht **atuh** Bapak Wuakakakakaka]

(5) *Minimal pake peluru karet* ***atuh***
[At least use rubber bullets **atuh]**

Other language norms are related to the correct linguistic behavior. Based on the n-gram data on word-level, it was found that there was a tendency to deviate from the norms for writing prepositions. In n-1 some authors tend to write the word *di* as a preposition which is written combined with the word that follows as shown in data (6) and (7), for example, *disana* [there], *disini* [here], *disono* [there], and others. Grammatically, it is a wrong way to write prepositions in Indonesian language which should be written separately. In contrast to n-2, some authors write the passive prefix *di-* separately even though it is attached/prefixed to the verb, such as *di bawa* [taken] and *di bungkus* [wrapped], as shown in data (8) and (9).

(6) *Selamat jalan kesayangan gonggong yg kenceng* ***disana***
[Farewell dear, bark loudly **there**]

(7) ***disini*** *aku pengen denger juga suara lato-lato*
[**here** I want to hear the sound of lato-lato too]

(8) *Skrg yg terpenting **di ambil** hikmahnya.... Alhamdulillah banyak yg sayang....*
[Now the most important thing is **to learn** from it.... Alhamdulillah, many love you....]

(9) *udah **di bungkus** pake karet duwa*
[It is already **wrapped** in two rubbers]

Deviation from language norms is not a mistake, but rather a sign of the uniqueness of an author. This is just some examples of author's attributions or uniqueness that this study attempting to uncover. An author's authorship profile may be referenced by ten or more conventional attribution. Then all of these attributions will be recorded, distinguishing one from other authors. For example, two people have the habit of writing word *di* incorrectly, but that is just one attribution signal. There are other attribution markers like the use of other characters and word choices. Even if they are identical, the weight will be determined by the N-gram tracing results. Based on n-gram analysis, namely n-1 and n-2, it was found that an author can be unique in the way he/she chooses words, abbreviates words, uses spaces to separate words, the way they use prefixes, suffixes, infix, etc. The uniqueness of the author can be traced and found in the n-gram analysis, both at the character-level and the word-level. The uniqueness of the author can already be found in the smallest n-units, both characters, and words. Through n-gram analysis, it can also be found how an author follows or even deviates from the norms. Language norms are characteristics and a special marker of an authorship.

Statistical Test Results Using the Similarity Comparison Method (SCM)

As an effort to prove the uniqueness of the authors, this research examines two sets of text from the same and different authors. This test utilized the similarity comparison method by determining n-grams and calculating the coefficient index and frequency weight of two sets of text with the Jaccard Coefficient and TF-IDF formula. As technology develops, researchers use computational technology to obtain fast and accurate statistical test results. These results then become the basis for a qualitative analysis of the authorship style of the text in question. This study developed a calculation program based on a similarity comparison method with the Jaccard Coefficient and TF-IDF formula as a tool for performing statistical tests.

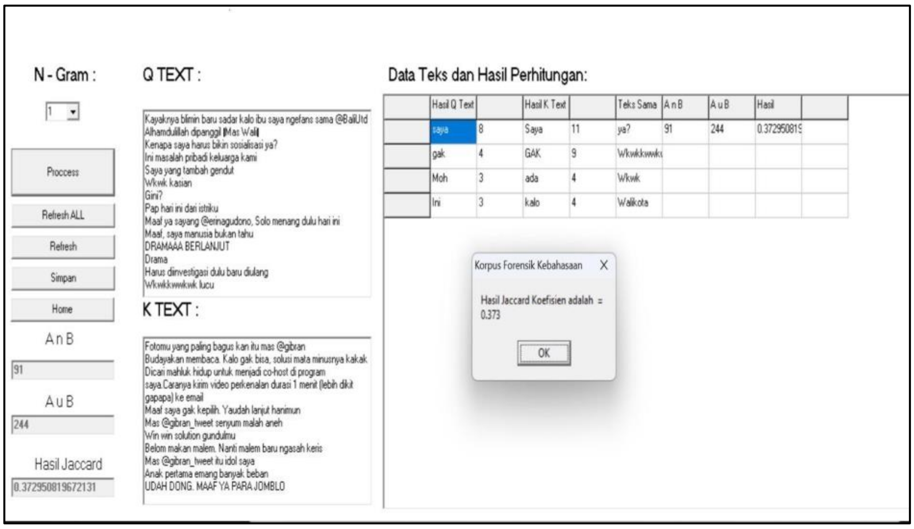

Figure 5. Example of Jaccard Coefficient Calculation Results

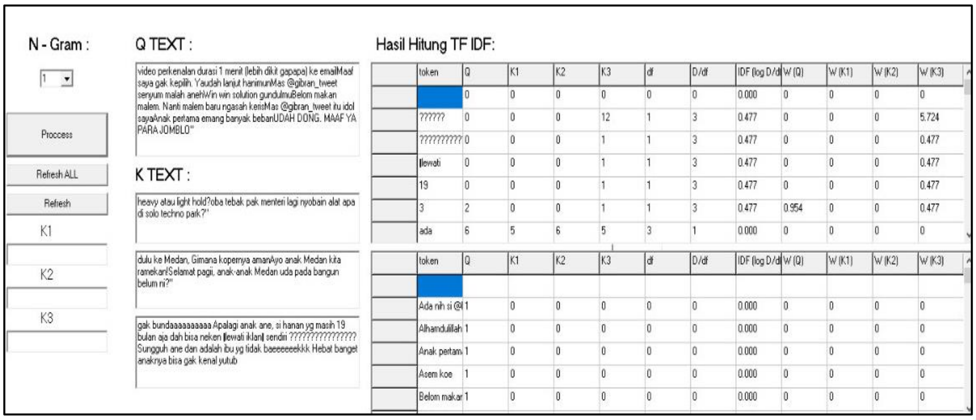

Figure 6. Example of TF-IDF Calculation Results

The frequency with which a term (word or n-gram) appears in a document or set-text is measured by term frequency (TF). The TF denotes the significance of a term within a set-text. Terms that appear frequently are likely to be important in conveying the text's substance and style. The Inverse Document Frequency (IDF) method assesses the rarity or uniqueness of a term over a set of texts. IDF indicates how distinct a term is. Terms that are widespread throughout numerous papers have a lower IDF, whereas terms that are uncommon and unique to a single document have a higher IDF. A term's TF-IDF weight or calculation results are calculated by multiplying its Term Frequency (TF) by its Inverse Document Frequency (IDF). TF-IDF weights terms that are both frequent within a set-text (high TF) and uncommon across the full set of documents (high IDF). This aids in accentuating terms that are unique to a set-text, allowing it to stand out from the crowd. TF-IDF is also used to represent each document as a vector in a high-dimensional space, with each dimension representing a distinct term. This vector's values represent the relevance of each term in the document. This representation enables document comparison based on content and usage of distinguishing terms. TF-IDF is very beneficial in authorship identification since it helps detect one author's distinct lexicon. Terms that are heavily weighted in one author's works but not in others become indicative of that author's style.

Table 7. Statistical Test Results using Jaccard Coefficient

No.	Jaccard Index	Test Result	
		Same Author	Different Author
1.	0.74-1	10%	0%
2.	0.5 – 0.74	10%	0%
3.	0.25 - 0.49	43%	0%
4.	0.1 – 0.24	25%	0%
5.	< 0.1	12%	100%

The Jaccard Coefficient computed the features shared by two texts that are both double and unique, and then divided by the total number of features in both texts. When Q-text and K-text are compared, the J-index ranges from 0 to 1. If the J-index values are near to one, the two texts being compared were created by the same author (Grieve et al., 2019; MacLeod & Grant, 2012; Nini, 2018). This study used n-1 as the unit of analysis in statistical tests and was conducted in two stages: a comparative test of two text sets by the same author and a comparative test of two separate corpora of authors chosen at random by the application. The option of this n-1 is shown in the top-left corner of Figure 5 in the drop-down N-Gram menu. The quantity of documents in this statistical test is 300 text sets from 100 unique author corpora, and the results are shown in Table 5.

Table 7 above shows the results of the statistical test of the similarity comparison method. Where the results of the similarity test for two sets of texts by the same author are shown in the third column (same author), while the results of the similarity test for sets of texts by different authors are in the fourth column (different author). Two sets of documents (Q-text and K-text) from the same author will have a high Jaccard index based on the results of statistical tests. 20% of the compared texts/documents (i.e., Q-text and K-text) has a similarity index greater than 0.5, indicating that the two sets of papers analyzed were written by the same author. Statistical tests show that even though the two sets of texts tested come from the same author, the similarity index or weight is not always high, as can be seen in the results in rows three and four. Approximately 68% of the data shows that the Jaccard index is between 0.1 and 0.49, while the remaining 12% is between 0.01-0.09. Meanwhile, statistical testing on two sets of documents by different authors reveal a low Jaccard index of less than 0.1. According to Nini and Grive (2018), if the Jaccard index results are getting closer to 1, it can be concluded that the two texts being compared were written by the same author, and vice versa, if the Jaccard index results are getting closer to 0, it can be concluded that the two sets of documents were written by different authors.

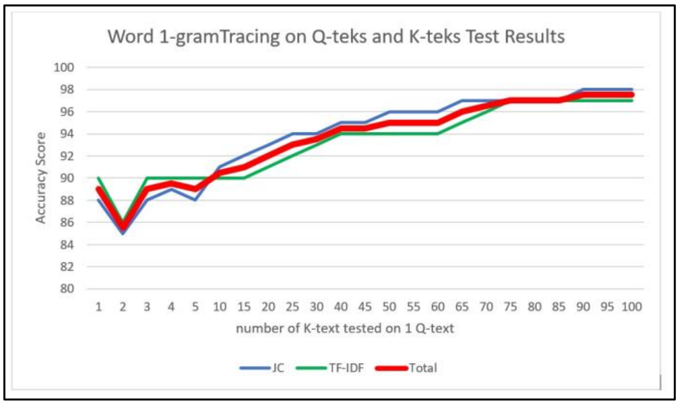

Figure 7. Accuracy Score of N-gram Tracing Using Jaccard Coefficient and TF-IDF

This study created a statistical test application and ran a test with the following scheme: each Q-text is tested with 1 to 100 K-texts. This means that each author's text is compared to the texts of 1 to 100 different authors. This scheme is used to test the accuracy of SCM as well as the ability of the Jaccard index and frequency score to determine the similarity of words used in each batch of texts at the 1-gram level. The SCM accuracy was found to be in the range of 85% to 96%

based on the test results as shown in Figure 7. The test findings were able to determine authorship in the one-on-one scheme between Q-text and K-text.

Indonesian Authorship Attribution

Each author's idiolect is the language in which they talk and write, and it is unique to them. This idiolect will manifest itself in the text through characteristics and unusual decisions, or through what is known as allegorical language (Coulthard, 2004). The linguistic style of a person is how they talk or write. The language style is a set of lexical, syntactical, and personality characteristics that are more likely to be utilized together in a person's communication. Each author's language style is recognized (Juola, 2007).

The selection of linguistic features refers to an author's proclivity to pick and apply features. Even while each component may not present in every piece, there is a pattern that repeats as an indicator of one's writing style (Juola, 2007). The data analysis revealed that an author's style can be found at the lexical level as shown in Table 4 and 6. The terms "diction" and "lexical element" both relate to the author's deliberate selection and usage of specific words to achieve aims (Bacchini, 2016; McIntyre, 2015; Neme et al., 2015).

Based on the n-gram analysis, a pattern of word choice is discovered in the smallest n-unit at the word level, indicating the author's style. Six words are first-person pronouns in the corpus of 100 unique author's text sets, namely *saya*, *aku*, *gue*, *ane*, *ikam*, and I. The first-person pronoun chosen by the author is one of the easiest attributions of authorship to identify. Even though it is relatively common with a high frequency in the corpora, each author has his style of using words. In the n-unit analysis at the character level, the author chooses words that are expressed in the form of a very specific and patterned number of letter characters.

The author's background influences his or her choice of words, where linguistic rules will be followed or even broken. Aside from the choice of words and even the characters used to write a word, adherence to or transgression of conventions will establish an author's style and individuality. In this study, it was discovered that there was norm compliance, specifically regional norms with the phrases *gue*, *ane*, *ikam*, and I, which significantly indicate the area of origin and the environment in which people reside.

Norm violation is also an indication of authorship distribution. According to the findings of this study, 20% of authors violated or departed from correct linguistic behavior. The aberration discovered in this study is the writing of prepositions, based on n-gram data on word units. In n-1, some authors use the word 'in' as a preposition that is coupled with the word follower, such as *disini*, *disana*, *disono*, and others. Unlike n-2, 10 authors wrote the word *di* individually even though it is grammatically positioned as a verb, such as *di ambil* 'taken' and *di bungkus* 'wrapped'. It can be concluded that conformance to and transgression of rules are identifiable attributions of authorship.

The use of the word *di* that violates grammatical rules is an example of a habit that can be a marker of an author’s attribution which will then be combined with other attributions from the analysis and trace of n-grams in a text, and their weight calculated. This writing error, which is actually quite common in Indonesian texts, is an important finding in this research, because this is one of the most dominant markers in the entire data. Even though it cannot be categorized as the most unique, this data is the most dominant one found. Of course, other very personal attribution markers also determine an author's profile.

Stylistic studies often signify an author's authorship based on all stylistic features. However, it is difficult to continue the analysis at the level of coherence, figure of speech, or other aspects that necessitate a bigger range of documents in this study (Brennan et al., 2012; Hoover, 2007; Neal et al., 2017). In personal works, an author prefers to use short, simple sentences. This study discovered short discussions with only one to three words. As a result, this study finds that stylistic elements can only be utilized to attribute authorship to personal texts from WhatsApp, Telegram, and Twitter discussions at the lexical and grammatical levels.

CONCLUSION

Despite the fact that the method for examining evidence with linguistics is considered as relative, linguistics strategies demonstrate sufficient thoroughness to provide objective perceptions of realities, reliable consequences, and legitimate purposes with quantitative examination characteristics. The only disadvantage of qualitative analysis is the inability to make definite conclusions from the facts. With their detailed measurements and computations, the quantitative approaches supported in this study: n-gram tracing, statistical test utilizing SCM, and stylometric characteristics test, cannot precisely verify the author's identity. In-depth manual analysis is still needed regarding other linguistic elements or features, especially regarding meaning, considering that Indonesian is also full of semantic and pragmatic features. This is where the role of linguists is still needed to interpret and validate statistical test results. However, because of its high level of accuracy, the quantitative method is an alternative to identifying an author and might be used to solve criminal cases involving authorship. Using quantitative methodologies, authorship analysis can be predicated on assertions of accuracy. Furthermore, the analytical results can be given with statistical data, very accurate measurement results, and conclusions that are no longer relative, thus claims for authorship.

NOTE

We would like to thank two anonymous reviewers for very helpful comments on the earlier draft.

REFERENCES

Alshammari, N., & Alanazi, S. (2021). The impact of using different annotation schemes on named entity recognition. Egyptian Informatics Journal, 22(3), 295–302. https://doi.org/10.1016/j.eij.2020.10.004

Anwar, W., Bajwa, I. S., Choudhary, M. A., & Ramzan, S. (2019). An empirical study on forensic analysis of Urdu text using LDA-based authorship attribution. IEEE Access, 7, 3224–3234. https://doi.org/10.1109/ACCESS.2018.2885011

Aziz, E. A. (2021). A linguistic contribution for law and justice enforcement 1(1), 1–22. https://ojs.badanbahasa.kemdikbud.go.id/jurnal/index.php/jfk/index

Bacchini, S. (2016). “The routledge handbook of stylistics”. *Reference Reviews*, Vol. 30 No. 4, pp. 20-28. https://doi.org/10.1108/rr-03-2016-0074

Bailey, B. (2000). Qualitative methods in sociolinguistics. Journal of Linguistic Anthropology, 10(2), 285–286. https://doi.org/10.1525/jlin.2000.10.2.285

Baker, P. (2010). *Sociolinguistics and corpus linguistics*. Edinburg University Press.

Baker, P., Gabrielatos, C., Khosravinik, M., Krzyzanowski, M., McEnery, T., & Wodak, R. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. Discourse and Society, 19(3), 273–30. https://doi.org/10.1177/0957926508088962

Belvisi, N. M. S., Muhammad, N., & Alonso-Fernandez, F. (2020). Forensic authorship analysis of microblogging texts using n-grams and stylometric features. *2020 8th International Workshop on Biometrics and Forensics (IWBF), Portugal, 1–6*, https://doi.org/10.1109/IWBF49977.2020.9107953.

Brennan, M., Afroz, S., & Greenstadt, R. (2012). Adversarial stylometry. ACM Transactions on Information and System Security, 15(3), 1–22. https://doi.org/10.1145/2382448.2382450

Casillas, L., & Ramirez, A. (2019). Emotion mining mechanism over texts in social media. Research in Computing Science, 148(7), 227–240. https://doi.org/10.13053/rcs-148-7-17

Chiang, E. (2021). Book Review: Language and online identities: The undercover policing of sexual crime by Tim Grant and Nicci MacLeod, 2020. Pp. x + 195. *International Journal of Speech, Language and the Law*, 28(1), 155–160. https://doi.org/10.1558/ijsll.20645

Coulthard, M. (2004). Author identification, idiolect, and linguistic uniqueness. Applied Linguistics, 25(4), 431–447. https://doi.org/10.1093/applin/25.4.431

Coulthard, M. (2013). On admissible linguistic evidence. Journal of Law & Policy, 21(2), 441-446. https://brooklynworks.brooklaw.edu/jlp

Eder, M., Rybicki, J., & Kestemont, M. (2016). Stylometry with R: A package for computational text analysis. R Journal, 8(1), 107–121. https://doi.org/10.32614/rj-2016-007

Fobbe, E. (2020). Text-linguistic analysis in forensic authorship attribution. JLL, 9, 93–114. https://doi.org/10.14762/jll.2020.093

Frye, R., & Wilson, D. C. (2018). Defining forensic authorship attribution for limited samples from social media. *Proceedings of the 31st International Florida Artificial Intelligence Research Society Conference, FLAIRS 2018*, 248–251.

Gorsuch, G. (2009). Book Review: An introduction to forensic linguistics: language in evidence by Malcolm Coulthard and Alison Johnson. London: Routledge, 2007. Pp. x + 237. *Studies in Second Language Acquisition*, 31(1), 130–131. doi:10.1017/S0272263109090093

Grant, T. (2007). Quantifying evidence in forensic authorship analysis. International Journal of Speech, Language and The Law, 14(1), 1–25. https://doi.org/10.1558/ijsll.v14i1.1

Grant, T., & Baker, K. (2007). Identifying reliable, valid markers of authorship: A response to Chaski. International Journal of Speech Language and the Law, 8(1), 66–79. https://doi.org/10.1558/ijsll.v8i1.66

hooverikeojuolamautner, 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

Hoover, D. L. (2007). Corpus stylistics, stylometry, and the styles of Henry James. Style, 41(2), 174–203. http://www.jstor.org/stable/10.5325/style.41.2.174

Ikeo, R. (2008). Book Review: An Introduction to Forensic Linguistics: Language in Evidence by Malcolm Coulthard and Alison Johnson, 2007. London: Routledge, pp. 237. ISBN 978 0 415 32023 8 (pbk). *Language and Literature*, 17(4), 377–379. https://doi.org/10.1177/09639470080170040505

Ison, D. (2020). Detection of online contract cheating through stylometry: A pilot study. Online Learning, 24(2), 142–165. https://doi.org/10.24059/olj.v24i2.2096

Juola, P. (2007). Authorship attribution. Foundations and Trends® in Information Retrieval, 1(3), 233–334. https://doi.org/10.1561/1500000005

Mautner, G. (2009). Corpora and critical discourse analysis. In P. Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 32–46). Bloomsbury.

McIntyre, D. (2015). Towards an integrated corpus stylistics. Topics in Linguistics, 16(1). https://doi.org/10.2478/topling-2015-0011

McMenamin, G. R. (2019). *Forensic linguistics: Advances in forensic stylistics*. CRC Press LLC.

Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. ACM Computing Surveys, 50(6). https://doi.org/10.1145/3132039

Neme, A., Pulido, J. R. G., Muñoz, A., Hernández, S., & Dey, T. (2015). Stylistics analysis and authorship attribution algorithms based on self-organizing maps. Neurocomputing, 147(1), 147–159. https://doi.org/10.1016/j.neucom.2014.03.064

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

Patodkar, V.N., & I.R, S. (2016). Twitter as a corpus for sentiment analysis and opinion mining. International Journal of Advanced Research in Computer and Communication Engineering, 5, 320–322.. https://doi.org/10.17148/ijarcce.2016.51274

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

Puspitasari, D. A. (2021). Tracing Word Trends on Social Media in 2012 and 2020 Through Corpus Linguistics. In J. Endardi (Ed.), *Demi bahasa bermanfaat dan bermartabat: percikan pemikiran strategi kebahasaan dalam dinamika bahasa, pendidikan, dan budaya era kiwari* (pp. 40–54). Deeppublish Publisher.

Puspitasari, D. A. (2022). Corpus-based speech act analysis on the use of word ‘lu’ in cyber bullying speech. *Proceedings of the 1st Konferensi Internasional Berbahasa Indonesia Universitas Indraprasta PGRI, KIBAR 2020, Indonesia*, 1–10. https://doi.org/10.4108/eai.28-10-2020.2315314

Puspitasari, D. A., & Sukma, B. P. (2022). Potraying The Covid-19 hoaxes at the beginning of the pandemic through a corpus assisted discourse analysis. Ranah: Jurnal Kajian Bahasa, 11(2), 243. https://doi.org/10.26499/rnh.v11i2.5152

Rebuschat, P., Meurers, D., & McEnery, T. (2017). Language learning research at the intersection of experimental, computational, and corpus-based approaches. Language Learning, 67(S1), 6–13. https://doi.org/10.1111/lang.12243

Rheingold, H. (2000). *The virtual community*. The MIT Press. https://doi.org/10.7551/mitpress/7105.001.0001

Rifai, B. (2020). *Pemanfaatan metode riset digital dalam pengembangan ekosistem penelitian dan inovasi*. LIPI.

Snee, H. (2016). *Digital methods for social science: An interdisciplinary guide to research innovation*. Palgrave Macmillan London.

Takwin, B. (2020). Tantangan psikologi siber. Jurnal Psikologi Sosial, 18(1), 3–4. https://doi.org/10.7454/jps.2020.02

Tarrayo, V. N. (2020). Wounds and words: A lexical and syntactic analysis of Casocot's "There are other things beside brightness and light." Indonesian Journal of Applied Linguistics, 10(2), 502–512. https://doi.org/10.17509/ijal.v10i2.28594

Theophilo, A., Giot, R., & Rocha, A. (2021). Authorship Attribution of Social Media Messages. *IEEE Transactions on Computational Social Systems*, 10(1), 10–15. https://doi.org/10.1109/tcss.2021.3123895

Unik, M., & Larenda, V. G. (2019). Analisis investigasi android forensik short message service (SMS) pada smartphone. JOISIE (Journal Of Information Systems And Informatics Engineering), 3(1), 10–15. https://doi.org/10.35145/joisie.v3i1.414