Intisari

Seorang penulis secara tidak sadar meninggalkan bekas atau identitas pada tulisannya melalui gaya bahasa yang konsisten pada tataran grafologi, morfo-sintaksis, dan pilihan leksikal. Analisis terhadap fenomena kebahasaan yang konsisten ini dapat ditemukan melalui analisis kepenulisan (*authorship analysis*), bahkan kuantitasnya pun dapat diukur. Penelitian ini menganalisis kumpulan teks personal dalam Bahasa Indonesia untuk mengidentifikasi kepemilikan suatu teks (*authorship*) dan profil penulis berdasarkan penelusuran N-gram (*N-gram tracing*), serta mengidentifikasi atribusi kepenulisan (*authorship attribution*) pada teks elektonik Bahasa Indonesia. Sumber data dalam penelitian ini adalah set-teks dari penulis unik. Semua teks adalah teks elektronik berupa pesan pribadi dalam bentuk percakapan SMS, WhatsApp, Instagram dan Facebook yang diambil dari Dokumen Putusan MA tahun 2020-2022 dan 100 penulis unik yang merepresentasikan kepenulisan dan profil kebahasaan dari berbagai wilayah di Indonesia. Data berupa karakter, kata, frasa dan kalimat dikelola dalam bentuk korpus data dengan total 2.095 token dan 935 tipe kata untuk data berupa barang bukti kasus dan sejumlah 2,1 juta token untuk data teks non-kasus. Analisis data dilakukan dengan menentukan, menelusuri, dan menghitung register yang dihitung dalam satuan N-unit pada setiap set-teks yang dibandingkan, baik pada level karakter maupun pada level kata. Penelususran N-gram dilakukan untuk mengidentifikasi fitur-fitur linguistik yang menjadi penanda atau atribut seorang penulis (*authorship attribution*). Pada tahap analisis data juga dilakukan uji statistik dengan metode perbandingan similaritas (*similarity comparison method*) dengan mengidentifikasi dan menghitung korelasi N-unit menggunakan Jaccard Coefficient. Penelitian ini juga mengukur akurasi *N-gram tracing* dalam mengidentifikasi penulis. Analisis N-gram pada level karakter menunjukkan, bahwa sebagai N-unit terkecil, karakter menjadi elemen penting penanda atribusi kepenulisan, seperti penggunaan karakter alfabet dan non-alfabet, penggunaan huruf besar atau kecil, dan tanda baca. Hasil analisis data pada level kata menunjukkan, bahwa secara leksikal, pilihan kata merupakan fitur linguistik penanda atribusi kepenulisan yang paling dominan dan berpengaruh dalam mengidentifikasi profil penulis, serta akurat dalam membedakan antara penulis yang satu dengan penulis lainnya.

Kata kunci:
Authorship Analysis; Authorship Attribution; N-gram; Pilihan Leksikal; dan Register.

Abstract

An Author unconsciously leaves a mark or identity in his writing through a consistent language style at the level of graphology, morpho-syntax and lexical choice. Analysis of this consistent linguistic phenomenon can be found through authorship analysis, and its quantity can even be measured. This research analyzes a corpus of personal texts in Indonesian to identify ownership of a text (authorship) and author profile based on N-gram tracing, as well as identifying authorship attribution in Indonesian electronic texts. The data source in this study is a set of texts from unique authors. All texts are electronic texts in the form of private messages in the form of SMS, WhatsApp, Instagram, and Facebook conversations taken from the 2020-2022 Supreme Court Decision Documents (Direktori Mahkamah Agung Republik Indonesia) and 100 unique authors who represent authorship and linguistic profiles from various regions in Indonesia. Data in the form of characters, words, phrases, and sentences is managed in the form of a data corpus with a total of 2,095 tokens and 935-word types for data in the form of case evidence and a total of 2.1 million tokens for non-case text data. Data analysis was carried out by determining, tracing, and calculating registers calculated in N-units for each set of texts being compared, both at the character level and at the word level. N-gram searches are carried out to identify linguistic features that are markers or authorship attribution. At the data analysis stage, statistical tests were also carried out using the similarity comparison method using the Jaccard Coefficient correlation formula. This research also measures the accuracy of N-gram tracing in identifying authors. Analysis of N-grams at the character level shows that as the smallest N-unit, characters are an important element in marking authorship attribution, such as the use of alphabetic and non-alphabetic characters, the use of letters, and punctuation. The results of data analysis at the word level show that lexically, word choice is the most dominant and influential linguistic feature of authorship attribution in identifying the author's profile, as well as accurately distinguishing between one author and another.

Keywords:
Authorship Analysis; Authorship Attribution; Lexical Choice; N-gram and Register.