Surveying Stylometry Techniques and Applications

TEMPESTT NEAL, KALAIVANI SUNDARARAJAN, ANEEZ FATIMA, YIMING YAN, YINGFEI XIANG, and DAMON WOODARD, University of Florida

The analysis of authorial style, termed stylometry, assumes that style is quantifiably measurable for evaluation of distinctive qualities. Stylometry research has yielded several methods and tools over the past 200 years to handle a variety of challenging cases. This survey reviews several articles within five prominent subtasks: authorship attribution, authorship verification, authorship profiling, stylochronometry, and adversarial stylometry. Discussions on datasets, features, experimental techniques, and recent approaches are provided. Further, a current research challenge lies in the inability of authorship analysis techniques to scale to a large number of authors with few text samples. Here, we perform an extensive performance analysis on a corpus of 1,000 authors to investigate authorship attribution, verification, and clustering using 14 algorithms from the literature. Finally, several remaining research challenges are discussed, along with descriptions of various open-source and commercial software that may be useful for stylometry subtasks.

CCS Concepts: • **Applied computing** → **Document analysis**; • **Computing methodologies** → *Natural language processing*; *Machine learning approaches*;

Additional Key Words and Phrases: Adversarial stylometry, authorship analysis, stylometry

ACM Reference format:
Tempestt Neal, Kalaivani Sundararajan, Aneez Fatima, Yiming Yan, Yingfei Xiang, and Damon Woodard. 2017. Surveying Stylometry Techniques and Applications. *ACM Comput. Surv.* 50, 6, Article 86 (November 2017), 36 pages.
https://doi.org/10.1145/3132039

1 INTRODUCTION

The beginning of stylometry is generally traced to the early suggestion of resolving authorship disputes by Augustus de Morgan through the frequency of word lengths in 1851. Subsequently, this prompted the first manual quantitative analysis in the late 1880s by Thomas C. Mendenhall who used word length distributions from the works of Bacon, Marlowe, and Shakespeare for identifying the true author of Shakespeare plays. Many years passed before George Kingsley Zipf discovered a relationship between the rank and frequency of words in 1932, which later came to be known as Zipf's Law. Similar efforts followed, such as George Yule's measurement of word frequency for analysis of vocabulary richness in 1944, which is now known as Yule's Characteristic. However, the research literature largely refers to the work of Mosteller and Wallace on the Federalist Papers in the early 1960s as the foundation of computer-assisted stylometry [117–119], while the Federalist Papers remain as a corpus of interest [60]. The following decades brought about advancements due

Authors' addresses: T. Neal, K. Sundararagan, and Y. Xiang, Department of Computer and Information Science and Engineering; emails: {tempesn, kalaivani.s, yingfeixiang}@ufl.edu; A. Fatima, Y. Yan, and D. Woodard, Department of Electrical and Computer Engineering; emails: {aneez.fatima, yanyiming, dwoodard}@ufl.edu.

https://doi.org/10.1145/3132039

to increases in the availability of data through the emergence of digital technologies and internet media, application of machine learning, information retrieval techniques, and neural networks [114, 115, 174], along with improved natural language processing tools [61]. Refer to Reference [59] for a detailed discussion on the historical development of stylometry.

Seminal research has continued well into the 21st Century, effectively utilizing stylometry for tasks ranging from plagiarism detection to fine arts. For instance, the increase of information shared online has significantly simplified the ability to copy-and-paste. Because this is an egregious offense, particularly in academia, researchers have applied stylometry to detect the copied work and trace it back to the responsible author [7, 54, 165]. On the other hand, several works have sought out to seek the ideal "style" markers related to musical composers and painters [14, 17, 65, 66], while additional applications include countering identity deception in social media applications, preventing and detecting email impersonation, multi-modal authentication on mobile devices, authorship detection in SMS messages, and identifying speech writers [1, 50, 53, 56, 92, 116, 134].

Present-day efforts generally reflect a steady progression of refinements since earlier works laid the framework for feature extraction and authorship analysis techniques. However, as the big data era brings about numerous sources of textual data, such as social media accounts and blogs, new challenges are introduced. For instance, growth of internet communication has caused a shift in approaches to consider more unbalanced datasets during experimentation. However, the majority of research efforts are not robust under these circumstances. Authorial style is more so a behavioral biometric; an author has free will to purposely alter one's style according to genre, topic, manuscript guidelines, and so on. The ability to effectively write in several contexts is likely beneficial for a successful writer, but this same inconsistency further complicates authorship analysis. Moreover, stylometry does not presently offer studies on the general population for making comprehensive conclusions regarding authorial style. As a result, most of the information gained from a single study is isolated to the researcher's dataset. Furthermore, as author sets become larger, stylometry features lose discriminating power; thus, stylometry characteristics do not carry the same identifying capabilities as physiological biometrics, as is generally the case with behavioral modalities. This survey provides a review of some of these challenges by detailing several recent approaches while summarizing the remaining challenges. Particular attention is given to large-scale authorship attribution and verification.

The contributions of this survey are trifold. First, Sections 2 through 4 thoroughly define five stylometry subtasks via the extension of previous surveys beyond authorship attribution, authorship verification, and authorship profiling to include discussions on stylochronometry and adversarial stylometry [43, 71, 86, 110, 144, 155]. A discussion on how these tasks are carried out is provided to include publicly available datasets, feature extraction techniques, and recently explored experimental approaches. Second, Section 5 provides an extensive performance analysis on a corpus of 1,000 authors with small text samples using 14 approaches from the literature to demonstrate the difficulty of large-scale authorship analysis as several research efforts indicate the negative correlation between training samples and number of candidate authors on performance, limiting authorship attribution techniques to small, and in some cases, unrealistic datasets. This section aims to provide insight into what has been termed the *needle-in-a-haystack* problem, where the goal is to accurately identify the author of a document among a large set ($\geq$1,000) of potential authors, each having a small number of training samples [86]. Third, Section 6 summarizes various research challenges and open problems associated with the current state of stylometry as a whole and each individual subtask. Overall, this survey aims to provide the most up-to-date review of stylometry, while providing researchers entering the field with a holistic view of stylometry, recent approaches, and performance expectations when handling larger corpora. We intend to provide a

computational perspective of stylometry; as a result, this survey will primarily focus on the current state of large-scale stylometry studies and the approaches considered in the past few years. While some linguistic details may be briefly mentioned, we provide numerous references that are useful for understanding the historical progression of stylometry and the current state-of-the-art.

2 SUBTASKS

Subtasks in stylometry include authorship attribution, authorship verification, authorship profiling, stylochronometry, and adversarial stylometry, with most emphasis typically placed on the first three. Authorship attribution, or author recognition/detection, is a core stylometry task; however, when precise identification is unfeasible, authorship profiling helps to reduce the search space of the actual author via identification of various demographics such as gender or age. While authorship attribution seeks an exact author, authorship verification aims to determine if documents were written by the same author. Finally, stylochronometry aims to estimate the time period in which a document was written, where several similarities between a test document and a historical document may provide evidence that the test document was also written during the same period. From a high-level perspective, authorship attribution, verification, profiling, and stylochronometry are all similarity detection problems, where the goal is to obtain some measurement of the likeness between two (or more) documents. On the contrary, adversarial stylometry is for subverting attribution, verification, and/or profiling to reduce the likelihood of being recognized. If cleverly done, then authors can also employ stylochronometry techniques to purposely mimic the writing styles of authors in previous time periods for adversarial stylometry.

2.1 Detecting Stylistic Similarities

The general aim of similarity detection problems is to compare the stylistic characteristics of documents [43]. Advancements in various techniques that detect stylistic similarities are beneficial for improving authorship attribution, authorship verification, authorship profiling, and stylochronometry.

2.1.1 Authorship Attribution. Authorship attribution aims to determine the probability that a document was written by a particular author based on stylistic traits rather than the content of the document [69, 71]. Authorship attribution stems directly from stylometry analysis, where it is assumed that features, such as subconscious syntactic idiosyncrasies, are sufficient in defining an author's unique style. Closed- and open-set attribution are typically used as experimental protocols. Closed-set attributes texts to one of the authors in the training samples. Open-set classification allows for an unknown author by generalizing closed-set identification using a threshold for similarity.

Various forms of authorship attribution have been introduced, including k-attribution, cross-domain authorship attribution, and source code authorship attribution. K-attribution, or ranking, is a relaxed authorship attribution problem, where the classifier outputs the k top authors ranked by their probability of being the true author [3]. This is useful in contexts where the exact author cannot be confidently identified. Similar to authorship profiling, it helps to reduce the search space of an authorship attribution problem. Cross-domain, or cross-genre, authorship attribution seeks for links across multiple, nonsimilar domains to answer research questions similar to "which of n bloggers wrote the given Twitter feed?" Hence, domain adaptation aims to identify an author M of a document written in domain A, while only having documents from M in domain B such that features from domain B must be applicable to domain A. While this may be difficult due to domain-specific factor influences, cross-domain authorship attribution has tremendous potential to reveal clues regarding what is necessary for establishing a standard in feature generation. Furthermore,

its application is particularly suitable for scenarios where the domain in which the questioned text is taken is the only available text for a particular author in that domain, though the author has several documents in a different domain. On the other hand, source code authorship attribution is useful for software plagiarism, virus detection, and cyber attacks [108]. Source code authorship attribution differs from code clone detection, in which the intention is to detect copied programs (or program segments) [173]. Source code authorship attribution, however, aims to determine if two programs are written by the same author, regardless of the program's purpose.

2.1.2 Authorship Verification. Authorship attribution is generally trying to find the author responsible for an unlabeled text sample given many labeled text samples, while authorship verification is typically a binary classification problem that decides if two documents were written by the same author [55]. Authorship verification typically produces two answers: *same-author* pairs occur when texts X and Y are assumed to be written by the same person. *Different-author* solutions occur when X and Y appear to be written by two different people; it is this possibility that renders verification as an open-set problem [89].

A few approaches to authorship verification have been proposed. *One-class classification* is presented as an option in Reference [85], where the classifier is trained using samples from the author in question and the test document is compared to the trained model. In this case, there is no notion of negative samples during training, hence single or one-class classification. The *Many-Candidates* method has been proposed to treat the verification problem as an attribution problem via the creation and use of a set of impostors [89]. Even still, the quality of such a set is parameter-defined, and failure to carefully choose such parameters may limit the accuracy of the results.

2.1.3 Authorship Profiling. Authorship profiling is the analysis of a document to determine certain demographics such as gender without directly identifying the author [64, 136, 142, 178, 182]. For instance, it is shown that language use may correlate with age (e.g., pronoun usage decreases with age, but prepositions become more frequent) and teen bloggers may use more non-dictionary words than adults [126]. Unfortunately, authorship profiling is challenging in online environments, as author characteristics may be unavailable, anonymous, or false.

Profiling may require some knowledge of the theme and topics discussed throughout documents to gain an in-depth sense of the emotional state of the author. Hence, in some works, content-specific features are shown to be more discriminative in authorship profiling as opposed to authorship attribution [101]. Understanding prominent themes based on content-specific features may offer a deeper level of information that is more relevant and useful in identifying certain demographics and/or characteristics compared to other surface-level features. However, it is important to "note that the use of content-based features for authorship studies can be problematic. While it is plausible that style-based markers can truly distinguish one class of authors from another, one must be concerned that content markers might just be artifacts of a particular writing situation or experimental setup and might thus produce overly optimistic results that will not be borne out in real-life applications" [10].

An example of authorship profiling was demonstrated during PAN 2015. The PAN[1] 2015 competition placed emphasis on authorship profiling, particularly in identifying the five traits in the Big Five Personality model (i.e., "extraversion, emotional stability, agreeableness, conscientiousness,

[1]PAN (Uncovering Plagiarism, Authorship, and Social Software Misuse) workshop began in 2009 and has grown into a series of evaluation labs consisting of competitions to address some of the limitations in several stylometry subtasks [9]. PAN is established as the main forum for research in stylometric analysis, largely due to several volumes of corpora and challenging tasks [72, 160].

Table 1. Effects of Machine Translation on Simple and Complex Sentences [24]

Original	English → Japanese → English	English → Japanese → German → English
We always eat breakfast at noon.	We are always at noon to eat breakfast.	We have to always eat breakfast at noon.
Even though he was a great soccer player, I doubt that he will be able to recover from his injuries to later become successful in the professional league.	Even though he was a great football player, I doubt that you will be able to recover from his injuries to become a success in the professional league after him.	Although he was a great football player, I doubt that you will be able to recover from his injuries after being him a success in the professional league.

and openness to experience" [15, 180]) as these traits were considered important factors for understanding author personalities based on word usage [160].

2.1.4 Stylochronometry. Stylochronometry is the study and detection of changes in authorial style over time. Language-dependent factors become important as some languages are highly unstable over extended periods of time. Thus, changes found in an author's language may be a reflection of changes in the actual language instead of changes in the author's writeprint. For instance, Reference [31] investigates changes in writing style over time of two Turkish authors using average word length, where results show that average word lengths of newer works were significantly larger than older works for both authors, suggesting a correlation between word length and document age. Furthermore, words of eight characters or less, particularly of four characters, were most likely found in an older work, while words of nine characters or more were most likely found in a newer work. Hence, the authors attribute the observed word length change to either mastery of the Turkish language over time or changes in the language itself.

Overall, there are limited studies in stylochronometry, and those that exist are specific in nature. In other words, these efforts are directed at specific authors, specific time frames, specific languages, and so on. As a result, the reliability of stylochronometry is less established compared to authorship attribution and verification; it is more likely that these studies have yet to reach maturity and widespread acceptance [162].

2.2 Adversarial Stylometry

Adversarial stylometry is defined as an evasion of authorship attribution via alteration of one's style [167]. There are three forms of adversarial stylometry: *imitation*, *translation*, and *obfuscation*. Imitation is writing to closely match the style of another author, while translation involves machine translation services to translate the language of a document to and back from one or many languages. While adversarial stylometry techniques should retain original meaning and only focus on altering the style of the document, machine translation can significantly alter the original meaning of sentences as shown in Table 1. Notice that simpler sentences are easier to translate, but lack obfuscation; complex sentences are harder to translate, but are easier to obfuscate [24]. Shallow obfuscation entails deliberate changes to one's writing style via identifying and adjusting the frequencies of a small set of attribution features for avoiding recognition during authorship attribution. On the other hand, deep obfuscation involves a more thorough analysis of an author's style, and techniques such as unmasking, in which strongly weighted features are iteratively removed in two texts for a measure of similarity depth between the two, are employed for evaluation beyond surface features [75, 85]. Various characteristics of obfuscated texts have been shown to

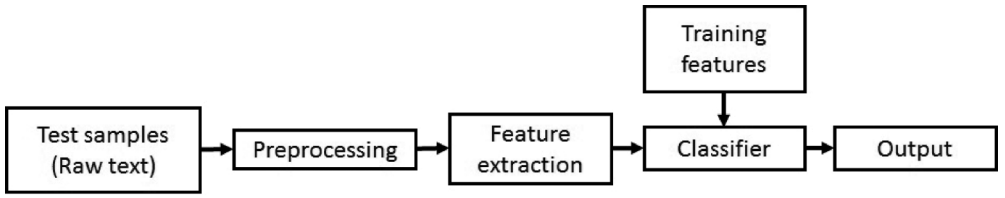


Fig. 1. A typical authorship attribution system.

differ from original texts. For instance, obfuscated texts may be less complex in readability, have a shorter average word length, and contain more adverbial phrases [40].

An adversary may also consider style transformation as a means to text obfuscation. Style transformations can include lexical-level changes, where words or sentences are shuffled for differences in semantics, or content changes, where sentences or phrases are replaced while retaining original meaning. For example, synonym and phrase replacements are considered as a means of style transformation in Reference [80]. Using sample texts correctly classified, automatic phrase and synonym detection and replacement are performed. A different classification following the transformation indicates that the style of the document has been changed successfully. Out of 13 test cases, seven are transformed successfully using the proposed technique. The authors note that style transformations should meet two criteria: (1) confirmation that a style shift has actually taken place, and (2) the transformed document retains its original meaning and remains grammatically correct. Interestingly, character-level features may be harder to transform: "To the extent that authorial 'style' is a function of specific vocabulary items, it is easy for an author to mask his or her style by picking different words, but it is difficult to change large-scale emergent statistics such as character frequency. Consider, for example, how reasonable an editorial request to 'use American spelling' for a journal article appears, especially in comparison with a request like 'use no more than 10% e's' " [74].

Finally, obfuscation techniques are also explored in programming code. According to Reference [122], source code obfuscation is considered effective when its implementation is cheap and prolongs the time required for a hacker to decompile the code. It can be applied at various phases, including at the source-code level and on binary, byte, and machine code until the desired level of obfuscation is reached. Design obfuscation consists of class merging or splitting, data obfuscation consists of restructuring data structures, control obfuscation hides control flow information, and layout obfuscation, which is the closest related to examining the actual stylometry related features of the texts, attempts to significantly alter comments and the naming of variables. Formally, the original program P is mapped to obfuscated program P' such that both programs behave the same, while P' is usually slow, memory inefficient, and much more complex due to the transformation [122].

3 DATA PREPARATION AND FEATURE EXTRACTION

A typical stylometry subtask (particularly for similarity detection) is carried out as shown in Figure 1. Given a corpus of text samples (descriptions of publicly available corpora are given in Table 2), the samples are usually preprocessed for normalization and noise reduction. Examples of preprocessing include tokenization (i.e., splitting a stream of text into words, phrases, etc.), stemming (i.e., only retaining the root or base form of a word), tagging (i.e., replacing words with their grammatical type), removing non-alphabetic characters and spaces, and converting uppercase letters to lowercase. Exact preprocessing tasks, however, are mostly dependent on the document

Table 2. Publicly Available Datasets for Stylometry

	Name	Document Type	Language	# of Authors	# of Samples	Avg. Data per Author	Link
1	Enron [83]	Emails	English	158	200,399	757 samples	http://www.cs.cmu.edu/~enron/
2	Brennan-Greenstadt Corpus [24]	Essays, work reports, and other formal letters	English	45	—	6500 words	https://psal.cs.drexel.edu/tissec/
3	Extended Brennan-Greenstadt Adversarial Corpus [24]	Essays, work reports, and other formal letters	English	12	—	5000 words	https://psal.cs.drexel.edu/tissec/
4	Twisty [179]	Twitter posts	Dutch, German, French, Italian, Portuguese and Spanish	18,168	34,136,161	2000 tweets	http://www.clips.uantwerpen.be/datasets/twisty-corpus
5	CliPs [178]	Reviews and essays	Dutch	—	749	Avg. review length of 128 words/avg. essay length of 1126 words	http://www.clips.uantwerpen.be/bibliography/csi-corpus
6	Personae [106]	Student essays	Dutch	145	145	1400 words	http://www.clips.uantwerpen.be/datasets/personae-corpus
7	ISOT Twitter [25]	Twitter posts	Assumed as English	100	100	3,194 tweets of 301,100 characters	http://www.uvic.ca/engineering/ece/isot/datasets/
8	ISOT Forgery [25]	Twitter posts	Assumed as English	10	300	4,253 characters	http://www.uvic.ca/engineering/ece/isot/datasets/
9	ICWSM 2009/2011 Spinn3r [29, 30]	Blogs, news articles, classifieds, forum posts, and social media content.	Unknown, but likely varied	—	386,576,659	—	http://icwsm.org/data/index.php
10	Slavonic [169]	Web articles	Czech and Slovak	—	30,000	—	http://nlp.fi.muni.cz/projekty/acb/preview
11	PAN [9, 72, 73, 138–141, 158, 161]	Dependent on subtask in various languages.	Varying	Varying	Varying	Varying	http://pan.webis.de/
12	Amazon Commerce Reviews [102]	Online reviews	English	50	2500	50 reviews at 856 characters per review	https://archive.ics.uci.edu/ml/datasets/Amazon+Commerce+reviews+set#
13	Blog Authorship Corpus [147]	Blogs	Mostly English	19,320	681,288	35 posts and 7250 words per person	http://u.cs.biu.ac.il/~koppel/BlogCorpus.htm

type. For instance, preprocessing of HTML documents may include removal of tags, extra spaces, stop words, and normalization of words via capitalization or stemming [175], while URLs, @ mentions, hashtags, and duplicate tweets may be removed from Twitter posts [53]. Features are then extracted from the preprocessed samples. Two main feature extraction approaches exists: *profile-based* approaches join all samples of an author for extraction of a single feature vector, whereas *instance-based* approaches extract a single feature vector per document [155]. Researchers have characterized stylometry features into five main levels: lexical, syntactic, semantic, structural, and domain (or content)-specific [155] (see Reference [52] for an empirical comparison of various features).

The simplest feature representation are lexical features. Lexical features are often word-based, but it is not uncommon for researchers to refer to character and word-based features as both lexical. At the character level, a document is viewed as a sequence of characters, and character-level metrics are extracted to represent the simplest structure of the document. Character-level features are therefore language-independent. These features can also capture subtle differences in style and contextual information. Compared to word-based features, character features are not highly effected by noise (i.e., spelling inconsistencies and grammatical errors), while natural language processors are not required for feature extraction.

A common character-level feature is *n*-grams, or the frequency of *n* consecutive characters in a text. For large values of *n*, complete tokens are captured, retaining more information. Determining the optimal *n* is language-dependent, however, as languages differ in average word lengths. A small *n* for use in Chinese, for instance, may capture thematic information, since the average character length of Chinese words is two. However, in English, it may only capture syllables. Moreover, due to high redundancy, document representations can become extremely large and feature selection is typically required to reduce the dimensions of these features.

Word-level features require word boundaries for use in tokenizers. However, some languages do not contain word boundaries, and language processors are not optimized for all languages. Therefore, extraction of lexical features at the word level is not always a simple procedure. Additionally, word-based features may be noisy if there are many misspellings or use of abbreviations and punctuations. Common features, such as function words, will produce questionable results when attempting to measure the frequencies of several misspelled words that should be mapped to the same term. Hence, for efficient extraction of word-based features, significant preprocessing may be necessary for text normalization and noise removal.

Researchers have also developed various *vocabulary richness* measures to quantify the diversity of sentences. However, these measures are language-dependent, and can only be accurately calculated when natural language processing tools allow accurate sentence extraction on the document's respective language. Nonetheless, vocabulary richness measures attempt to gain an understanding of an author's use of vocabulary and the complexity of an author's language. Several measures have been proposed, including:

(1) Zipf's Law models a linear relationship between the number of vocabulary items appearing r times in a document and r (see Reference [13] for a thorough mathematical explanation).
(2) Yule's K measure assumes that the occurrence of a word is based on chance and can be modeled according to a Poisson distribution [61].
(3) Yule's I measure = $\frac{M_1 M_1}{M_1 M_2}$, where M_1 is the total number of words in a document and M_2 is the sum of weighted word forms with a certain frequency. A larger result indicates a richer vocabulary [79].
(4) The Burrows method considers large sets of high-frequency function words per 1000 words and then applies PCA. This method is considered a standard technique in authorship attribution [61].
(5) Hapax legomena counts the number of words to appear once, while hapax dislegomena counts the number of words to appear twice [125].
(6) Type/token ratio is the ratio of all tokens (or words) to unique tokens (referred to as types).

Several other vocabulary richness measures have been formulated [58], such as Simpson's D [150], due to the researcher's personal notion of authorial style based on word usage.

The *bag-of-words* (BoW) approach generally refers to lexical-level features as it represents a document as a bag (or collection) of words, discarding context, grammar, and word order. Most researchers rely on bag-of-words representations given that lexical-level features typically yield state-of-the-art performance. In fact, a thorough review of the research literature suggests that, though the simplest, character *n*-grams and function words are the most reliable and efficient features for authorship attribution. Character *n*-grams are simple to extract, language independent, and tolerant to noise (e.g., various word forms). On the other hand, function words were one of the first to be considered as stylometry features, and have continued to be utilized in present-day studies [187]. They are typically regarded as context-free (and are therefore less influenced by topic and genre), while revealing social and personal aspects of our lives [35, 171].

Syntactic features capture patterns from sentence structure. Though language-dependent, syntactic information is highly reliable assuming the availability of accurate and robust tokenizers, parsers, and part-of-speech taggers. Noise is introduced, however, when these tools are outdated. Tags, phrasing, and rewrite rules are all syntactic features that represent the unique way in which an author structures sentences [12]. For instance, Reference [19] investigates two novels, *The Adventures of Sherlock Holmes* and *The Memoirs of Sherlock Holmes*, using a very simple syntactical approach that constructs an array of word sequences of specified length coupled with their frequency in the text. In the authorship attribution domain, Reference [135] constructs probabilistic context-free grammars (PCFGs) as syntactic features, where unknown documents are parsed for grammar and assigned to an author using maximum likelihood.

Semantic features capture meaning behind words, phrases, and sentences, such as through analysis of synonyms and semantic dependencies. For instance, Reference [36] proposes a particularly interesting approach that considers synonym-based features as indicators of style for authorship attribution. The authors argue that their approach takes into account meaning of words, where an author's alternatives in word selection are valuable. Substitutions for certain words vary greatly, such as synonyms for "red", while others, such as "computer" are limited in synonyms. Words with a larger set of synonyms are given more weight, given that the author had a larger word space to choose from.

Structural features represent the organization of a document, such as how an author prefers to use indentations and signatures [188]. Structural features are very useful in online contexts, especially when structure is an important component of the document. For instance, emails have a special layout that is highly distinct from essays. Further, use of signatures or greetings is highly specific to emails. Structural features should capture this information and model author-specific structural components.

Domain-specific features are sometimes referred to as content-specific features given that they rely on the content in the document [188]. In this sense, content encompasses the thematic and contextual clues given in the document. Domain-specific features may also include lexical-level features that are only applicable to that specific domain, such as mentions in Twitter posts. Keywords are often used as domain-specific features to determine common themes that author's tend to gravitate to, such as those that indicate intentions to sell a product. Therefore, domain-specific features are confined to the application in which the documents are drawn.

Some feature representations are harder to categorize; they may span several feature levels or stand as a category on their own. For instance, readability features employ various measures that calculate the grade level at which the document can be understood; however, this requires language processors to examine individual words [142]. Idiosyncratic features include misspellings, mistakes in grammar, and purposely chosen words that reflect social and cultural backgrounds [84, 100]. Writeprints is an individual-author-level feature extraction model for authorship attribution [2]. Individual-author features differ from author-group features in that

the latter uses a single feature set across all authors while the former extracts individualized features per author. Writeprints consist of two steps: creation using Karhunen-Loeve transforms on several features extracted from multiple levels and pattern disruption for leveraging absences of features during similarity detection. In Reference [159], typical features are ignored altogether by allowing language processing tools to analyze and extract style markers.

Topic models are an additional feature form that consists of a generative model that considers documents as a collection of topics. An implementation could assume known probability distributions of topics, where a particular topic would have higher probability of having certain terms. It has been shown that author-topic models are useful in understanding what seems to be hidden information about documents. One of the most important and influential topic model implementations is the Latent Dirichlet Allocation (LDA) [22]. LDA is a three-level Bayesian technique for modeling a collection over a set of topics [22, 148]. Thus, LDA is a probabilistic model where each topic is determined by word distributions. The introduction of LDA influenced several researchers to further consider LDA variations via extension of its original implementation. For example, Reference [143] extends LDA for author-topic modeling. Topic distributions are assigned to authors, while word distributions are assigned to topics. Hence, detection of multi-author documents should correlate with a mixture of topic distributions. The authors intend to simultaneously model author and topic information. This technique could be useful in determining when authors prefer to cover a single topic versus multiple topics.

Table 3 summarizes the various levels of features.

4 CLASSIFICATION APPROACHES

Following feature extraction, train and test features are compared to determine the similarity or distance between the two. There are a range of classification approaches to determine this similarity, including machine-learning classifiers and clustering techniques, distance-based models, and probabilistic methods.

4.1 Machine-Learning Classifiers and Clustering

Machine-learning algorithms may be useful for both clustering and classification. Machine-learning classifiers train learning and pattern recognition algorithms to find boundaries between classes that minimize some loss function. Machine-learning algorithms were sought out to handle various multidimensional and sparse feature representations [155]. Machine-learning algorithms are widely used in all stylometry subtasks. For instance, in a PAN Competition, all participants used a machine-learning algorithm for classification, including decision trees, support vector machines, logistic regression, and random forests [138]. On the other hand, clustering is commonly used to find stylistic similarities or to reduce the search space for authorship attribution problems. Clustering has been found useful in source code authorship attribution for detecting web spam [175] and for exposing stylistic similarities in web forum posts [129]. Machine-learning classification differs from clustering in the use of training labels; machine learning is typically a supervised learning problem, where class labels are known and incorporated in the classification process. Clustering is an unsupervised machine-learning procedure, where the algorithm derives a natural separation of the feature space that may or may not correlate with the class labels.

Support vector machines (SVMs) [41] are a common classifier choice largely due to the ability to handle large and sparse datasets. For instance, stylometry techniques are applied to English and Chinese online messages using lexical, syntactic, structural, and domain-specific features from content posted by 40 subjects over a 2-week period [188]. Compared to decision trees and neural networks, SVMs performed best when all features are considered. Another work uses 53 features from 25 web forum posts of six subjects for maximum recognition accuracies of 90% and 86.67%

Table 3. Feature Categories

Category	Description	Examples	Advantages/Disadvantages
Lexical	Divided into character and word-based groups to capture stylistic traits at each level.	(1) No. of characters (all, upper, or lower case) (2) No. of digits/white-spaces/ special characters (3) No. of words (4) Avg. word length (5) Avg. sentence length in words (6) Vocabulary richness	Can be applied to any language/corpus [16, 63]; Character *n*-grams can be used as an alternative to word level features, since the latter is strongly language-dependent with nontrivial feature selection and word segmentation [128]; Character *n*-grams can preserve morphological properties and reduces data sparsity in large vocabularies [128]; Vocabulary richness measures are not always sufficient in capturing even extreme intratextual and intertextual vocabulary variations, and an author's vocabulary richness in one text may differ significantly in another [62]; Lexical richness may suffer to distinguish between authors with similar vocabulary [177].
Syntactic	Capture style in the organization of sentences.	(1) Punctuation frequency (2) Function word frequency (3) No. of sentences beginning with a capital letter (4) Frequency of words with all capital letters	May require robust and accurate NLP tools and a large amount of features for style modeling [99].
Semantic	Attempts to capture the meaning of words and sentences.	(1) Synonyms (2) Semantic dependencies	Semantic analysis can be difficult even with language processors, particularly on unrestricted texts [155].
Structural	Represents how an author organizes a document.	(1) Paragraph length (2) Indentations (3) Use of greeting and farewell statements (4) No. of words/sentences/ characters per paragraph (5) Binary indicator of quotations (6) Binary indicator of URLs in signatures (7) Font size and color	May be more useful in online text, such as emails and blogs, when other feature types are less prominent, but use of structural components, such as greetings and the specific layout of the text body, remain available [188].
Domain-Specific	May include frequency of keywords or other content-specific information that gives rise to information about the theme of the document.	[188] employ *deal*, *obo*, *sale*, *wtb*, *thx*, *paypal*, *check*, *windows*, *software*, *offer*, and *Microsoft* as keywords in online messages that are potentially attempting to sale an item.	May express personal interest in a specific domain [188].
Additional Features	Proposed to fit researchers' needs and can span across multiple levels or stand alone.	(1) Readability metrics (2) Idiosyncratic features (3) Topic models (4) Writeprints models	May require extensive natural language processing tools; may offer information beyond typical surface level features.

for SVMs and decision trees, respectively [113]. Additional examples include extraction of lexical features from emails to train a one-class SVM classifier [1] and SVM classification of lexical and structural features from Chinese blogs, forum posts, and emails [108].

Neural networks are also a popular classifier choice. For instance, *n*-grams are investigated in conjunction with neural networks and Bayesian classifiers on 30 6,000-character samples in Reference [81]. An additional work refers to a collection of neural networks as committee machines that may vary in authorship attribution power. Hence, a voting scheme based on *k*-nearest neighbors is employed to ensure each neural network's information is collectively considered and appropriately weighted [93]. Another work argues that neural network language models are more suitable for capturing complex and longer text patterns, but require a large amount of training data [51]. The authors, therefore, propose feed-forward neural networks due to efficient computational complexity and the ability to handle limited training data. Despite the power of neural networks to

learn when the concept of feature extraction is not concrete, the invention of new features is not explicit and is inaccessible to researchers [137]. Alternative methods may be more useful in certain contexts.

4.2 Nearest-Neighbor/Minimum Distance Techniques

While generally referred to as a machine-learning algorithm, nearest-neighbor approaches are essentially seeking a minimum distance to a labeled training sample(s). Test samples are subsequently assigned the class of this training sample. The notion of distance as a classification method for stylometry subtasks is also regarded as *intertextual distance*. Several distance measures have been proposed to compute the distance between documents in terms of word usage, with a popular choice being Burrows's Delta.

From a high-level perspective, compression-based methods also aim to minimize some distance between train and test samples by concatenating all texts from an author to produce a file A. A compression algorithm produces the compressed file $C(A)$. The text in question, B, is then added to produce $A + B$ to further produce $C(A + B)$. The difference in bits between $C(A)$ and $C(A + B)$ measures the similarity (via minimizing cross-entropy) between the known and unknown texts. A detailed compression algorithm is discussed in Reference [91].

4.2.1 Intertextual Distance. Intertextual distances measure the similarity between the vocabulary used in two texts to determine if they were written by the same person, ideally satisfying the identity, symmetry, and triangle inequality properties. Labbé's work explicitly explores intertextual distance, formally described as "a calculation which considers entire text and which gives a standardized measure of the actual distance between it and another text" by summation of the differences in frequencies of each token [95–97]. Several distance metrics have been created and/or used for this purpose, including Delta, Chi-Square, Kullback-Leibler Divergence, and Stamatatos distances. Refer to References [68, 123] for more detailed discussions on additional intertextual distances.

Delta Burrows provides a simple technique that measures the difference in z-scores, or standard scores, of the relative frequencies of the most frequent words in texts, which he termed *Delta* [28]. Delta has proven to be one of the most robust intertextual distance measures by computing $\frac{1}{n}\sum_{i=1}^{n}|z(f_i(D)) - z(f_i(D'))|$ between documents D and D'.

Chi-Square Distance The chi-square distance, χ^2, is a common metric in intertextual distance models. χ^2 is a non-parametric goodness-of-fit statistical measure for determining if a set of frequencies were drawn from the same population. A lower χ^2 value indicates higher confidence that a sample was drawn from a particular population. Specifically, $\chi^2 = \sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$, where O and E are observed and expected frequencies. In intertextual distance models, the frequencies of lexical features are used, where the population is a collection of candidate author samples. This metric has been used in the authorship attribution problem based on newspaper articles [146].

Kullback-Leibler Divergence Kullback-Leibler Divergence measures the difference between two probability distributions, P and Q, as $KL(P||Q) = \sum_i P(i)log\frac{P(i)}{Q(i)}$, assuming discrete probability distributions. Reference [11] uses stopword graphs and the Kullback-Leibler metric in an authorship attribution problem. In particular, stopwords are utilized as nodes and edges indicate distances between nodes, allowing consideration of the flow and interaction of stopwords. For graph construction, nodes of the graph are initially generated according to a stopword list with no edges. Edge weights are updated by measuring the distance from a stopword w_s to recent occurrences of other stopwords $w_{1,2,\ldots,n}$. Weights

are assigned to each edge based on positions of w_s and $w_{i\in 1,2,\ldots,n}$. After constructing graphs for known and unknown texts, Kullback-Leibler Divergence distances are computed for each test graph against every training graph; the candidate responsible for the training graph with the smallest distance is decided to be the author. Experimentation included testing of 571 stopwords. Train and test documents were 50,000 and 10,000 words, respectively. Accuracies of 94.72% and 82.05% are achieved for binary class and 10-class classification, respectively.

Stamatatos Distances Stamatatos presents new distance measures based on character n-grams, though these metrics have been applied at the word-level. These metrics make use of author profiles, where a profile, P, is a pair (n-gram, normalized frequency) of the L most frequent n-grams in a text sample. The first metric measures the distance between an unknown text profile and training author profile: $d_1(P(x), P(T_a)) = \sum_{g\in P(x)} (\frac{2(f_x(g)-f_{T_a}(g))}{f_x(g)+f_{T_a}(g)})^2$, where $P(x)$ is the profile for the unknown text, $P(T_a)$ is the candidate author's profile, and $f_x(g)$ represents the frequency of n-gram g in $P(x)$. The second metric, d_2, concatenates all training samples as a normalization step: $d_2(P(x), P(T_a), P(N)) = \sum_{g\in P(x)} (\frac{2(f_x(g)-f_{T_a}(g))}{f_x(g)+f_{T_a}(g)})^2 \cdot (\frac{2(f_x(g)-f_N(g))}{f_x(g)+f_N(g)})^2$, where N is the concatenated text [154].

4.2.2 Compression. According to Reference [112], a known drawback to compression-based approaches is slow running time. However, the authors also note that compression models are easy to apply and require little to no preprocessing. They also allow use of off-the-shelf algorithms, such as RAR, gzip, and LZW.

Two compression-based implementations in particular have been shown to perform fairly well in large-scale authorship attribution. In the first work, authors are identified using relative entropy via measurements of distances between unknown texts. The lossless data compression algorithm, LZ77, is used to compress data by detecting duplicates. For an unknown text X, the author, A, of text A_i that minimizes the difference $L_{A_i+X} - L_{A_i}$ is chosen, where L_{A_i} represents the length of text A_i in bits after compression, and $A_i + X$ indicates a new sequence by appending X to A_i [18]. The second approach applies the Partial Matching (PPM) text compression scheme for text categorization. Compared to machine-learning algorithms, this method provides competitive and, in some cases, superior results based on the Reuters-21578 corpus [172]. Both of these approaches are further examined in Section 5.

4.3 Probabilistic Models

Probabilistic models are largely based on the properties of the Bayesian classifier that assumes independence of features [170]. Furthermore, it considers the probabilities that an event will reoccur based on past events. More specifically, probabilistic models want to find the author that maximizes $P(x|a)$ for unknown text, x, and candidate author, a.

Several variations of this model have been proposed, such as the Chain Augmented Naïve Bayes (CAN) model that combines naïve Bayes and statistical n-gram methods. The CAN model uses Markov dependencies to relax the Bayesian assumption of independence amongst words. For each unknown document, d, the candidate, c^*, with the largest probability, $P(c|d)$, amongst all candidates is chosen as the legitimate author. Tree Augmented Naïve Bayes (TAN) is an additional variation.

Machine-learning techniques are often compared to probabilistic models for analyzing performance. In particular, neural networks and Bayesian classifiers are combined for evaluation of 30 samples from authors Alexander Hamilton and James Madison based on character n-grams [81]. The authors determine that, compared to neural networks, Bayesian classifiers require

larger values for n and more features. Specifically, when $n = 5$, 95% accuracy is achieved with 11 features. However, this same accuracy was achieved with neural networks with only three features. It is suspected that assuming that features follow a normal distribution may actually hinder the performance of probabilistic models. Furthermore, Reference [187] compares several machine-learning classifiers, including probabilistic models, naïve Bayes, and Bayesian networks [187]. Using newswire articles and function words as features, both probabilistic models outperform neural networks and decision trees on two-class authorship attribution. Bayesian networks appear to be superior to all other methods, however, and continues to outperform as more documents per author are given. However, this trend is not continued in one-class classification; though comparative performance is observed, neural networks perform the best.

Overall, there is a common thread among these approaches. First, as discussed, machine learning generally takes two forms: supervised and unsupervised learning. Thus, we see that machine learning provides an umbrella for many supervised (e.g., SVM, neural networks, etc.) algorithms and unsupervised (e.g., clustering for authorship profiling) learning. K-nearest neighbors is a common machine-learning algorithm that learns a decision boundary based on the classes of the nearest neighbors to the test sample. This is a nonparametric technique; there is no knowledge of or use of the distribution of the data. Instead, the use of an appropriate distance metric determines the output of the learning algorithm. In References [8, 166], the authors demonstrate that Burrows's Delta is essentially a nearest-neighbor classifier. With a few minor modifications, quadratic Delta can be derived; thinking on this further leads to the probabilistic interpretation using a Gaussian distribution, and the authors show that the Burrows's Delta can be equally assessed using the Laplace distribution. Having already discussed the connection between distance and compression-based approaches, we are finally left with the mathematical relationship between nearest-neighbor algorithms and probability approaches.

Implementing these approaches may initially be a daunting task for those just entering the field. However, fortunately for novice and experienced researchers, several software suites have been developed specifically for stylometry, while others are useful for many machine-learning tasks. Table 4 lists several of these tools.

Last, Table 5 highlights recently published articles within each subtask, labeling each classification method as a machine-learning, distance-based, or probability approach. This table provides a snapshot of articles published since 2010 as previous surveys should provide sufficient coverage of articles prior to this time. This allows an up-to-date review of stylometry research within the current decade. References to works that introduced the data set and/or method used in these works are also indicated. Finally, for datasets not included in Table 2, concise dataset descriptions are provided.

5 PERFORMANCE ANALYSIS

Stylometry research has yet to establish state-of-the-art performance under demanding circumstances. In an ideal scenario, authorship attribution is a considerably easier problem when the set of candidate authors is small ($\leq$20) and each author has training samples of at least 1,000 words [12, 37, 135, 159]. However, online media and communication will produce datasets that do not satisfy these constraints. While a few implementations exist to specifically address this problem [105, 107, 109, 121], all results suggest that as the number of candidate authors increases or the amount of training data decreases, authorship attribution and verification accuracy decreases. This section provides a performance analysis that demonstrates these challenges. First, we describe several metrics commonly used to assess performance, followed by an analysis of the results for the performance assessment of large-scale authorship analysis.

Table 4. Available Software Useful for Stylometry Subtasks

Name	Availability	Description	Link
Pattern	Open-source	Python module for data mining, natural language processing, and machine learning [151]	http://www.clips.ua.ac.be/pattern
Stylometry with R: A Suite of Tools	Open-source	Provides a graphical interface for five components, including classification and keyword generation.	http://dh2013.unl.edu/abstracts/ab-136.html
NLTK: Natural Language Toolkit	Open-source	Regarded as an introductory Python library for natural language processing tasks with an easy-to-use interface [20, 21, 104]	http://www.nltk.org/
Zemberek	Open-source	Turkic language processing library used to generate statistical information for lexical-level features	https://code.google.com/archive/p/zemberek/
JGAAP (Java Graphical Authorship Attribution Program)	Open-source	Developed by Duquesne University. Includes canonicizers for preprocessing and event extraction for feature extraction	http://evllabs.com/jgaap/w/index.php/Main_Page
spaCy	Open-source	Claimed as robust, built for production, and features a whole-document design to read entire documents at once instead of relying on sentence detection. Custom algorithms allow management of both low-level and high-level details. Features a high-performance tokenizer, part-of-speech tagger, named entity recognizer and syntactic dependency parser, with built-in support for word vectors	https://spacy.io/.
JSAN	Open-source	Drexel University's two-part framework that includes JStylo, a Java-based graphical user interface and API, and Anonymouth that uses JStylo results to perform feature clustering, preferential ordering, feature modification, and document reclassification iteratively to achieve writing style anonymization	https://psal.cs.drexel.edu/index.php/Main_Page
OpenNLP	Open-source	Apache's perceptron-based machine-learning toolkit that supports natural language processing tasks such as tokenization, parsing, and coreference resolution	https://opennlp.apache.org/
StyleTool	Open-source	A word frequency-based stylometry tool. It performs principle component analysis and plots clusters using the Squint algorithm.	https://github.com/lnmaurer/StyleTool
Stanford CoreNLP	Open-source	An integrated framework of several linguistic analysis tools. Provides a part-of-speech tagger, sentiment analysis, and bootstrapped pattern learning [111]	http://nlp.stanford.edu/software/
Online Authorship Attribution Tool	Open-source	An online tool to conduct authorship attribution experiments. Several different techniques are implemented to get some degree of interpretation; however conclusive or definite answers are not provided	http://www.aicbt.com/authorship-attribution/online-software/
The Signature Stylometric System	Open-source	A software for literary detection that provides graphical representations of letter/word lengths and frequencies. The developers claim that the software works best with large datasets. Signature supports user or software-defined word lists, phrase, n-gram, and multiple language support.	http://www.philocomp.net/humanities/signature.htm
LexTo	Open-source	Dictionary-based online tokenizer with lexical analysis for the Thai language	http://www.sansarn.com/lexto/
ICTCLAS	Open-source	A Chinese lexical analysis system based on multi-layer Hidden Markov Models. Features word segmentation, part-of-speech tagging and unknown word recognition [186]	http://sewm.pku.edu.cn/QA/reference/ICTCLAS/FreeICTCLAS/English.html
AYLIEN Text Analysis API	Commercial	A package of natural language processing, information retrieval and machine-learning tools to support article, entity, and concept extraction and language detection.	http://aylien.com/
MeaningCloud	Commercial	Includes a set of APIs for language analysis and semantic annotation such as lemmatization, part-of-speech tagging and parsing, topic extraction, sentiment analysis, language identification, spelling, grammar and prooofreading and text classification.	https://www.meaningcloud.com/
Saplo API	Commercial	A JSON-based API that includes personalization and contextual analysis, related texts, entity tagging, sentiment analysis and topic extraction with have libraries to support several languages.	http://developer.saplo.com/
Cloud Natural Language API	Commercial	Google's API for natural language processing that uses both Parsey McParseface (Google's open-source parser), as well as commercial technology. Provides named entity detection, entity type tagging, entity linking, and sentiment analysis with multiple language support	https://cloud.google.com/natural-language/

Table 5. Recently Published Stylometry Papers

Reference	Subtask	Data	Features	Method	Performance
[135] (2010)	Attribution	3–6 authors of various news articles and poems	Syntactic features	Probabilistic (Probabilistic context-free grammars)	77.78–93.34% accuracy
[34] (2010)	Attribution	30 text samples	Lexical and syntactic features	Probabilistic (Conditional Random Field models)	60% accuracy
[32] (2010)	Stylochronometry, Profiling	40 novels from Turkish 40 authors with an average of 63,449 words	Frequent words, sentence, token, and type lengths, syllable counts	Machine learning (Discriminant analysis)	57.27% stylochronometry accuracy, 94.1% gender classification accuracy
[98] (2010)	Attribution	200 Twitter posts from 14,000 users	*n*-grams	Distance (Intersection of *n*-grams [48])	Over 70% accuracy
[87] (2011)	Attribution	10,000 blogs of 2000 and 500 words for known and unknown texts	*n*-grams	Distance (Cosine similarity)	Up to 94% precision
[33] (2011)	Attribution	Exams from 40 students	82 lexical and syntactic features	Distance (*K*-nearest neighbors)	90.7–95.42% accuracy
[24] (2012)	Adversarial Stylometry	Brennan-Greenstadt Adversarial Corpus	Lexical, syntactic, readability, and Writeprint features	Machine learning and distance (Neural networks, SVM, and synonym-based classification)	67% imitation success
[45] (2012)	Adversarial Stylometry	Over 1000 Trip Advisor and Yelp reviews and essays	Lexical and syntactic features	Machine learning (SVM)	60.7–91.2% detection accuracy
[67] (2013)	Profiling	PAN 2013	Common *n*-grams [77]	Distance (Dissimilarity measure)	53.8–58.4% in gender, 42.7–47.3% in age
[49] (2013)	Verification	19 subjects	Writeprints	Machine learning (SVM)	0.00122, 0.00218 FAR,FRR
[26] (2013)	Verification	Enron	*n*-grams	Distance (Percentage of shared *n*-grams)	14.35% EER
[124] (2014)	Attribution	8 subjects provided 22 samples each with an average of 150 words	62 lexical and syntactic features	Machine learning (Back propagation)	88.854–98.312% accuracy
[90] (2014)	Verification, Adversarial Stylometry	Extended Brennan-Greenstadt	Lexical and syntactic features	Distance (*k*-nearest neighbors)	0.06 Imitation success rate
[82] (2015)	Stylochronometry	Works of Mark Twain, Henry James, and a reference corpus from 1860s-1910s	Context sensitive word unigrams	Machine learning (Multiple linear regression models)	± 4–7.2 RMSE
[131] (2015)	Profiling	PAN 2015	Syntactic *n*-grams [132]	Machine learning (SVM)	51.1–58.4% in age, 53.1–65.9% in gender, and 17.1–21.1% in personality traits
[6] (2015)	Profiling	500 news articles from 30 authors [5]	BoW features and keywords which indicate sentiment and emotion	Machine learning and distance (Naive Bayes, decision trees, SVM, *k*-nearest neighbor)	56.1–86.4% with BoW features, 58.7–61.9% with sentiment features, 55.3%–86.4% in hybrid approach
[50] (2016)	Attribution	Text collected from smartphones from 200 subjects for 30 days to 5 months	*n*-grams	Distance (Percentage of common *n*-grams) [26]	0.16/0.11 FAR/FRR
[57] (2016)	Attribution	Extended Brennan-Greenstadt	Frame semantics	Machine learning (SVM)	≈ 0.15–0.8 error as the number of authors increases from 1 to 45
[184] (2016)	Attribution	Turkish news articles from 9 authors	Lexical and syntactic features	Machine learning (Artificial Neural Networks)	80–98% accuracy
[27] (2017)	Verification	Enron and Twitter posts from 100 users with 3,194 posts on average	Lexical, syntactic, and application-specific features	Machine learning (Gaussian-Bernoulli Deep Belief Networks)	8.21–10.08% EER

(Continued)

Table 5. Continued

Reference	Subtask	Data	Features	Method	Performance
[181] (2017)	Profiling	Slovene Twitter posts	Word and character *n*-grams	Machine learning (Linear support vector classification)	87.9–92.6% gender accuracy
[168] (2017)	Profiling	–	Readability metrics, vocabulary richness, and emotional status	Machine learning (Deep learning)	97.7% in gender, 90.1% in age
[183] (2017)	Attribution	PAN 2011, Blog Authorship Corpus	Lexical features	Probability, Machine learning, Distance (Posterior distributions, Nadaraya-Watson kernel regression, Euclidean distance)	30.8–54.2% accuracy
[157] (2017)	Attribution, Verification	CCAT-10 (Subset of Reuters Corpus with 10 authors and 100 texts per author), News articles from 13 authors, PAN 2014, PAN 2015	*n*-grams	Machine learning (SVM) and distance (frequency of *n*-grams)	49.7–79.4% attribution accuracy, 72.4–73.7% verification accuracy

Table 6. Confusion Matrix

	Positively Classified	**Negatively Classified**
Positive Samples	True Positives (TP)	False Negatives (FN)
Negative Samples	False Positives (FP)	True Negatives (TN)

5.1 Assessing Performance

Various measures are used for analyzing subtask performance. Derived from Table 6, the most common performance metrics include accuracy, recall, precision, F_1-score, and area under Receiver Operating Characteristic (ROC) curves (AUC). The research literature suggests that these metrics are the current standard for assessing the performance of the various subtasks, while error rates are rarely provided.

(1) Accuracy measures the percentage classified correctly over all test cases:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN}.$$

(2) Recall is the percentage of positive samples that were classified correctly. Recall measures how often a system correctly classifies positive samples when it encounters them:

$$Re = \frac{TP}{TP + FN}.$$

(3) Precision is the percentage of correctly classified positive samples over all positive classifications. Precision measures how often a system gets positive classifications correct:

$$Pr = \frac{TP}{TP + FP}.$$

(4) F_1-score measures the balance between the precision and recall of a system. A higher f_1-score indicates a more accurate system:

$$F_1 = 2 \times \frac{Pr \times Re}{Pr + Re}.$$

(5) The ROC curve is a plot of the false positive rate ($\frac{FP}{FP+TN}$) against the true positive rate ($\frac{TP}{TP+FN}$) at varying thresholds; the AUC is the area under the curve. A greater AUC indicates higher verification accuracy.

Table 7. Extracted Features

#	Character Features	#	Word Features	#	Syntactic Features
2	No. of characters/digits	2	No. of words/unique words	11	Frequency of punctuations (‘,’,‘.’,‘?’,‘!’,‘:’,‘;’,‘’’,‘”’,‘-’,‘(’,‘)’)
2	No. of alphabets (and uppercase)	1	Average word length	512	Frequency of function words
2	No. of white/tab spaces	2	Average sentence length in characters/words	50	Frequency of POS bigrams
26	Frequency of alphabets	10	Hapax legomenon	50	Frequency of POS trigrams
21	Frequency of special characters	20	Frequency of words of different word lengths	1035	Frequency of syntactic pairs
10	Frequency of digits	6	Fraction of all short, capitalized, uppercase, lowercase, camelcase, othercase words	-	-
150	Frequency of character bi, tri, and 4-grams	4	Yule’s I, Sichel’s S, Brunet’s W, Honore’s R measures	-	-
-	-	100	Frequency of word bi/trigrams	-	-

5.2 Corpus Description and Features

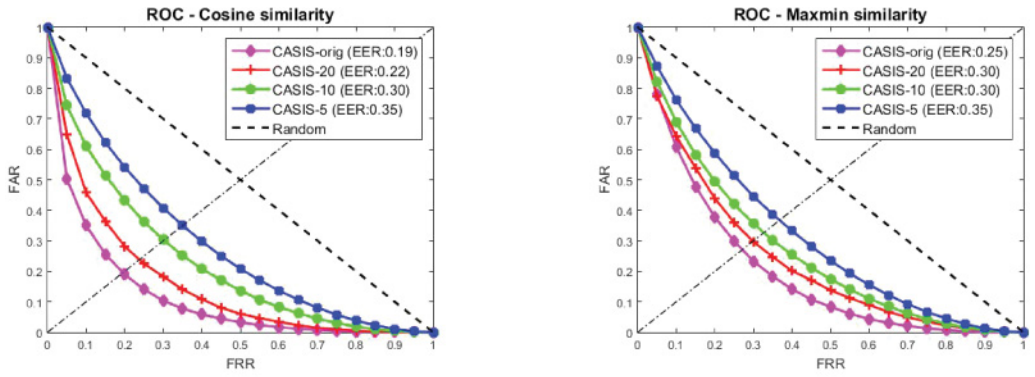
The Center for Advanced Studies in Identity Science (CASIS) corpus is used for experimentation. CASIS consists of 4,000 blog samples from 1,000 non-native English speakers (4 samples per author). Each sample is an average of 1,634 characters, 304 words, and 13 sentences. To study the effect of sentence lengths, author posts were divided such that author samples had an equal number of sentences, yielding subsets with 2, 5, 10, and 20 sentences per post in addition to the original posts (hereafter referred to as CASIS-2, CASIS-5, CASIS-10, CASIS-20, and CASIS-orig, respectively). CASIS-2 contains 984 authors and 27,662 samples, CASIS-5 contains 378 authors and 6,691 samples, CASIS-10 contains 103 authors and 1,597 samples, and CASIS-20 contains 19 authors and 229 samples.

2,016 dimensions of bag-of-words features are extracted for performance evaluation under verification and identification protocols and for feature and clustering analysis. Features include 213 dimensions of character-level features, 145 dimensions of word-level features, and 1,658 dimensions of syntactic features (refer to Table 7).

5.3 Authorship Verification

We investigate the effect of short text lengths on authorship verification performance using five-fold cross-validation on CASIS-orig, CASIS-20, CASIS-10, and CASIS-5. In each fold, similarity scores are computed between every test sample and every training sample using cosine and maxmin similarity metrics. Let $\mathbf{x}$ and $\mathbf{y}$ represent two n-dimensional vectors. Cosine similarity is defined as $\frac{\sum_{i=1}^{N} \mathbf{x}_i \mathbf{y}_i}{\sqrt{\sum_{i=1}^{N} \mathbf{x}_i^2} \sqrt{\sum_{i=1}^{N} \mathbf{y}_i^2}}$ and maxmin similarity is defined as $\sum_{i=1}^{N} \frac{min(\mathbf{x}_i, \mathbf{y}_i)}{max(\mathbf{x}_i, \mathbf{y}_i)}$. Instead of computing the similarity measure using all 2,016 dimensions, the technique of Koppel *et al.* is followed such that 500 dimensions are randomly chosen and similarity is computed using those dimensions [87]. This procedure is repeated for 10 iterations and the average similarity measure is computed.

With the similarity matrices for all five folds, the average false accept and false reject rates (FAR and FRR, respectively) across all folds are computed. ROC curves for all the data subsets are shown in Figure 2. A trend is noticed where the equal error rate (EER) (i.e., where the false accept and false reject rates are approximately equal) increases as the number of sentences per post decreases. This could be attributed to either the increase in the number of samples as the number of sentences per post decreases or to the noisy nature of frequency-based feature representations for shorter



(a) ROC with cosine similarity

(b) ROC with maxmin similarity

Fig. 2. ROC curves for all data subsets.

text lengths. Overall, this verification analysis highlights the difficulty of authorship verification for short text lengths using BoW features.

5.4 Authorship Attribution

Fourteen open-source algorithms were implemented for authorship attribution based on a "large-scale reproducibility study" presented in Reference [133]. The authors reimplement 15 influential author identification algorithms with the intentions of having a "significant impact on researchers entering the field" and to "...lay the groundwork for integrating author identification with information retrieval to eventually scale the former to the web." Thanks to this work, we are able to easily evaluate the robustness of several algorithms in more demanding experimental conditions as the authors intended.

The algorithms that we employ include:

(1) *Stopword Graphs and Authorship Attribution in Text Corpora*: Stopword graphs are used to measure similarity between train and test graphs, where stopwords serve as nodes and edges indicate distances between nodes. Edge weights are updated by measuring the distance from a stopword, w_s, to recent occurrences of other stopwords, $w_{1,2,\ldots,n}$. After constructing graphs for known and unknown texts, Kullback-Leibler Divergence distances are computed for each test graph against every training graph. This algorithm is hereafter referred to as `arun09` [11].

(2) *Language Trees and Zipping*: Benedetto *et al.* identify authors using relative entropy by measuring distances from unknown texts based on the LZ77 algorithm, which compresses data by detecting duplicates. This algorithm is hereafter referred to as `benedetto02` [18].

(3) *'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship*: Burrows introduces *Delta* as an intertextual distance measure. This algorithm is hereafter referred to as `burrows02` [28].

(4) *Mining E-mail Content for Author Identification Forensics*: This work applies SVM to non-sparse author-topic matrices, where features included 170 style markers (e.g., average sentence length) and 21 structural attributes (e.g., a greeting acknowledgment). This algorithm is hereafter referred to as `devel01` [39].

(5) *Local Histograms of Character N-grams for Authorship Attribution*: The locally weighted BoW (LOWBOW) framework is proposed to compute a set of local histograms smoothed by kernels on different parts of the document for preservation of lexical usage and positional information, due to the expectation that authors are expected to use certain words or characters in certain locations. This algorithm is hereafter referred to as `jairescalante11`[44].

(6) *N-gram-based Author Profiles for Authorship Attribution*: This algorithm uses similarity measurements from byte-level *n*-grams. For training, the *L* most common *n*-grams are found for each candidate author; normalized frequencies are then computed for each *n*-gram. Thus, a profile is built as a set of *L* pairs, and the frequencies from an unknown text are compared to the profiles to determine the similarity between the two. This algorithm is hereafter referred to as `keselj03` [77].

(7) *A Repetition-Based Measure for Verification of Text Collections and for Text Categorization*: Khmelev and Teahan introduce *R*-measure for determining the repeatedness of a document for an intuitive assessment of a dataset's characteristics. The *R*-measure is defined as the character-based normalized sum of all substring lengths that are repeated across multiple documents via construction of a suffix tree from the concatenation of all documents. For authorship attribution, the *R*-measure determines the correct class for document T among m candidate authors represented by texts $S_1, \ldots, S_m$ using estimate $\hat{\Theta}(T) = argmax_i R(T|S_i)$. This algorithm is hereafter referred to as `khmelev03` [78].

(8) *Measuring Differentiability: Unmasking Pseudonymous Authors*: Koppel *et al.* present "unmasking" for understanding the depth of difference in authorial style. The intuition behind this technique lies in the importance of only a small number of features that work to distinguish between two authors. During each cross-validation fold, the k strongly weighted positive and negative features are eliminated for observation of performance degradation curves. This algorithm is hereafter referred to as `koppel07` [88].

(9) *Authorship Attribution in the Wild*: Koppel *et al.* extracts 250,000 unique character 4-gram features from 10,000 blogs. For each blog, 2,000 and 500 words are chosen for training and testing, respectively. For each unknown snippet, the authors randomly choose some $k2$ fraction of the feature set, find the best match candidate using cosine similarity, and repeat choosing different subsets $k1$ times. Scores are then calculated for each candidate author to indicate the proportion of times such candidate is considered as the best match. This algorithm is hereafter referred to as `koppel11` [87].

(10) *Augmenting Naïve Bayes Classifiers with Statistical Language Models*: The Chain Augmented Naïve Bayes (CAN) model is applied to achieve 96% accuracy using trigram word-level models with absolute smoothing on a Greek corpus of 20 documents written by 10 modern authors. This algorithm is hereafter referred to as `peng04` [127].

(11) *Syntactic N-grams as Machine-Learning Features for Natural Language Processing*: Sidorov *et al.* introduce syntactic *n*-grams (sn-grams) to preserve syntactical relationships between words. Sn-grams are *n*-grams constructed from following paths in syntactic trees, compared to arbitrarily extraction of consecutive terms as they appear in the text. This algorithm is hereafter referred to as `sidorov14` [149].

(12) *Authorship Attribution Based on Feature Set Subspacing Ensembles*: Stamatatos presents a classifier ensemble based on a feature set subspacing method for authorship attribution using the 1,000 most frequent words of the training corpus. Following feature extraction, subsets from the feature set are selected using two approaches. The first approach involves the *k*-random-classifier (*k*RC), which randomly selects feature subsets of size

m k times. The second approach uses exhaustive disjoint subspacing (EDS) to randomly divide the feature set into equally sized disjoint subsets of size m. Using linear discriminant analysis as the base learning algorithm, k resulting classifiers are generated from k subsets, and a predefined combination method is utilized to ensemble the base classifiers. This algorithm is hereafter referred to as `stamatatos06` [153].

(13) *Author Identification Using Imbalanced and Limited Training Texts*: The Stamatatos distances are applied to multi-topic samples across 50 authors from the Reuters corpus. In most cases, d_2 outperforms d_1, especially when limited texts per author are available for training, and about 70% accuracy is obtained for the best case. It is claimed that the new distance measures provide a more reliable and stable solution than traditional n-grams, while retaining advantages such as language-independence and effectiveness. This algorithm is hereafter referred to as `stamatatos07` [154].

(14) *Using Compression-based Language Models for Text Categorization* Teahan et al. apply the Partial Matching (PPM) text compression scheme for text categorization. For an unknown document, its cross-entropy for the various categories are computed; the lower the cross-entropy, the more likely the document belongs to the category. This algorithm is hereafter referred to as `teahan03` [172].

For each algorithm, fivefold cross-validation was performed using CASIS-2, CASIS-5, CASIS-10 and CASIS-20, while fourfold cross-validation was used on CASIS-orig. In total, 360 experiments were performed. However, 19 of these failed to complete due to limitations in computational resources or the algorithm's inability to handle large-scale authorship attribution. These included CASIS-orig and CASIS-2 on `sidorov14` [149] and CASIS-2 on `jairescalante11` [44] and `koppel07` [88]. As a result, performance is reported for the remaining 341 experiments. Table 9 summarizes the performance of these algorithms on the CASIS partitions, while Table 8 provides attribution accuracy per algorithm.

Overall, it appears that lower-level representations, such as byte and character-level features, outperform higher-level features, such as syntactical and word-level representations. This could be due to the limited amount of text available in the author samples where high-level representations are sparse. It is also observed that when the introduction of the technique itself is emphasized in the initial work instead of its ability to scale, performance is not consistent. Finally, both compression-based methods performed well on the CASIS dataset; this could be an indication that compression-based algorithms are beneficial in large-scale authorship attribution.

Results also show performance decreases across all algorithms as text lengths decrease from 20 sentences per post (CASIS-20) to 2 sentences per post (CASIS-2). This could be attributed to shorter texts per post or due to the increase in the number of authors from CASIS-20 to CASIS-2. Furthermore, CASIS-2 and CASIS-orig have a similar number of authors; however, CASIS-2 has shorter text lengths and more number of samples than CASIS-orig. Hence, the performance of algorithms on CASIS-2 and CASIS-orig sheds light on the effect of text lengths. The general trend is that performance on CASIS-orig is better than CASIS-2 for all approaches except `stamatatos07`, implying the significance of text lengths on authorship attribution.

To further study the effect of varying sentence lengths, fivefold cross-validation is performed on a subset of 19 authors. While the authors remain the same, the number of sentences per post varies. The results are shown in Table 10. It can be seen that performance rates using 10 to 20 sentences per post are closer to that of the original posts. This could be due to the fact that the average sentence length of CASIS dataset is 13 ± 11 sentences per post. However, as the number of sentences per post decreases, performance also decreases.

Table 8. Authorship Attribution Accuracy (%)

	Poor Performance					Moderate Performance			Best Performance					
	arun09	khmelev03	koppel07	peng04	sidorov14	burrows02	devel01	stamatatos06	benedetto02	jairescalante11	keselj03	koppel11	stamatatos07	teahan03
CASIS-orig	12.5	5.6	1.8	13	-	9.83	3.58	32.03	24.55	31.38	60.33	53.68	35.58	59.63
CASIS-20	2.09	12	5.12	32	20.38	48.1	54.01	40.97	92.6	88.94	97.9	87.5	92.43	98.83
CASIS-10	0.25	6.1	1.56	7.5	8.63	21.3	27.48	5.9	76.4	66.55	82.64	76.4	83.2	90.9
CASIS-5	0.07	3.4	0.15	3.6	3.41	8.2	10.81	1.48	52.05	41.26	66.01	53.4	70.02	74.35
CASIS-2	0.07	2	-	12.465	-	3	4.56	0.59	31.06	-	46.99	26.72	46.56	52.47

Table 9. Algorithm Descriptions for Authorship Attribution

Reference	Performance			Comments
	Poor	Moderate	Best	
arun09 [11]	✓			This method may require documents of considerable length as training and testing documents had 50,000 and 10,000 words in the initial study. The authors find a 10% difference between the binary and 10-class problems. Extremely poor performance (near 0%) is expected when 100 classes are considered. The CASIS dataset, however, is attempting a 1000-class problem.
benedetto02 [18]			✓	It is assumed that text compression is a strong technique for large-scale authorship attribution as both compression-based methods performed well.
burrows02 [28]		✓		In the initial study, 47% accuracy is reported when 150 words are used and 30% when using 40 words in the feature set. As CASIS samples become smaller, this algorithm may not be able to extract enough common words.
devel01 [39]		✓		In small samples, there may not be enough semantic or lexical information to infer topics.
jairescalante11 [44]			✓	The use of multiple histograms across samples at different locations expands the training set. Hence, where other algorithms suffer due to small and limited samples, this method creates more training samples per author.
keselj03 [77]			✓	Frequency information at very low-level representations seems to be useful.
khmelev03 [78]	✓			It is suspected that CASIS either contains little repeatability in author samples and/or the algorithm cannot accurately distinguish between a large number of authors due to high R values from groups of different authors.
koppel07 [88]	✓			Repetitively removing strong features implies that there is a large feature set that can withstand the iterative nature of this algorithm. For smaller author samples, the algorithm may not be able to run enough iterations to achieve the same performance found when the authors tested on books.
koppel11 [87]			✓	This algorithm is developed specifically for the needle-in-a-haystack problem, which shows to contribute to the performance on CASIS.
peng04 [127]	✓			Several factors could have contributed to the low accuracy with CASIS: language-dependencies, inability to scale to larger datasets, and/or limited availability of word-level trigrams.
sidorov14 [149]	✓			It is possible that syntactical dependencies are harder to find as samples become smaller, limiting the accuracy of this algorithm.
stamatatos06 [153]		✓		Poor performance on the CASIS dataset could be due to the inability to properly train many classifiers on the information available in smaller samples.
stamatatos07 [154]			✓	Frequency information at very low-level representations seems to be useful.
teahan03 [172]			✓	It is assumed that text compression is a strong technique for large-scale authorship attribution as both compression-based methods performed well.

Table 10. Effect of Varying Sentence Lengths on Authorship Attribution Accuracy (%)

CASIS-orig	CASIS-20	CASIS-10	CASIS-5	CASIS-2
73.68	80.33	61.03	40.48	30.61

Clustering these features in an unsupervised fashion may further provide insight into the effectiveness of such feature representations. If authors can be uniquely identified with such feature representations, then one should find well-separated clusters with each containing samples from an author; that is, the cluster purities will be very high. However, clustering a large number of high-dimensional samples is challenging. Hence, graph-based clustering algorithm Markov clustering is used that has been optimized for scalability and speed [176]. Samples are represented as a graph network with every sample as a node and the similarity between two samples as edge weights. Only the closest K neighbors of every node are retained in the graph

(a) Clustering on CASIS-orig: 1000 authors, 4000 nodes, 196,000 edges resulted in 16 clusters

(b) Clustering on CASIS-2: 984 authors, 27662 nodes, 1,355,438 edges resulted in 96 clusters

Fig. 3. Clustering on CASIS-orig and CASIS-2.

network. To ensure the nearest-neighbor search is scalable for large datasets, Locality-Sensitive Hashing (LSH) is employed to increase the speed of the nearest-neighbor search.[2]

We compute a graph network using LSH for 50 nearest-neighbors and three different similarity measures, which was later provided as input to the Markov clustering algorithm. The clustering of CASIS-orig and CASIS-2 samples are shown in Figure 3.[3] Every circle denotes a graph node or a text sample. Same color nodes indicate a cluster. Results show that there are no separate connected blobs, indicating the difficulty of authorship attribution. However, each cluster may represent a meta-class that characterizes a writing style. Since the number of clusters is far less than the number of authors, it can be inferred that multiple authors are possibly grouped together. Further, it is also possible that an author can be characterized by multiple styles and can belong to more than one cluster.

5.5 Discussion

This section investigates the performance of authorship verification and attribution on a reasonably sized corpus of 1,000 authors. Fourteen open-source algorithms are employed for authorship attribution. Our results are indicative of the difficulty of these subtasks when the number of candidate authors is large or when the text samples are short in length. In verification, error rates remain unacceptably high and increase as text samples become smaller. For authorship attribution, less than half of the 14 algorithms were found suitable for our corpus. We also find that lower-level features are better representations for limited author samples, while compression-based methods generally give good performance. Finally, clustering of the features revealed no separated components, highlighting the difficulty of large-scale authorship attribution. It is observed that the number of clusters are far fewer than the number of authors, indicating that clusters might represent meta-classes with similar style instead of individual authors. Thus, while authorship attribution is robust when using a small set of candidate authors, the presented analysis highlights the difficulty of correctly identifying an author under demanding circumstances. Because of this, large-scale authorship attribution remains a non-trivial research challenge for stylometry researchers.

[2]For clustering, we employ the algorithm available at http://micans.org/mcl/ and the LSH implementation FALCONN available at https://github.com/FALCONN-LIB/FALCONN.

[3]The visualizations were made using BioLayout Express 3D available at http://www.biolayout.org/.

6 RESEARCH CHALLENGES

The current state of stylometry suggests several research challenges. These include lack of sufficient data within a single corpus, exploring and leveraging feature-level correlations, performance decreases in unconstrained conditions, and challenges specific to each subtask.

First, better performing research efforts mainly consists of evaluation on a small set of candidate authors with large amounts of samples per author. Larger corpora offer several challenges that are more difficult to address, such as authors having varying numbers/sizes of samples [144, 163]. Beyond this, there is no gold standard benchmark dataset that is used across all approaches, making performance comparisons cumbersome. A publicly available large-scale, multilingual dataset that contains texts from several document types, topics, genres, and time periods would significantly advance the current state of stylometry [70, 164].

Further, the research literature suggests that factors such as topic, genre, sentiment and style may be highly correlated, but very little efforts have attempted to explore these possible correlations. For instance, there may exists a relationship between semantic features and topic and authorship, while syntactic features may be related to the length of the document [145]. An optimal set of features that efficiently captures these correlations has yet to be discovered.

Generalizing the current levels of features to allow efficient modeling of several authorial styles under multiple circumstances (e.g., languages, genres, document types) is also desired [70, 156]. Currently, there is no standard for feature representation. Frequency vectors are commonly used, but other representations have also been considered, such as optimization of these vectors, matrices, and graphs [130, 152]. However, if classifier decisions could be explained to offer insight into why certain documents are attributed to certain authors, topics, genres, or themes, this could strengthen future efforts [164]. Unfortunately, Juola describes the current feature selection process as an "ad hoc mess," while others regard feature set behavior as classifier dependent, nondeterministic, and ultimately the greatest problem in stylometry [23, 70, 137].

Feature representation has also been shown to have an effect on performance. In one work, tensor space models are evaluated as an alternative to the typical vector space representation of features (R^n feature vectors are transformed into $R^x \times R^y$ feature matrices) [130]. Tensor models require less parameters and are claimed to be more suitable for limited training data. However, it is also unclear how to accurately apply any of these methods to multiple languages. It has been suggested to translate non-Western languages to English and extract features from the translated text, but the process of translation results in loss of clarity and cohesiveness [46].

Moreover, as demonstrated in the performance evaluation, the majority of current methods may fail to generalize to allow consistent performance in unconstrained conditions. For instance, performance has been shown to decrease as the number of candidate authors increase or the amount of training data per author decreases [120]. Ideally, less training data for a particular author should not correlate with the likelihood of that person being the legitimate author of the questioned text [130]. However, it appears that there is no definitive number that indicates how much data is needed for successful authorship analysis [42, 164].

Challenges are also present within each subtask. Authorship attribution is often hindered due to domain-specific factors. For instance, posts extracted from online blogs and social networking sites are often anonymous and use limited and grammatically incorrect vocabulary. Authors may even create their own novel languages to emulate non-verbal expressions. Many times, authors use several aliases and the number of candidate authors easily ranges above the thousands. Authors are not limited to a single language and the number of comments and commenters change dynamically with time in streaming messages. The intentional simplification of natural languages, such as through slang, further complicates application of natural language processing techniques

[3]. Mobile phones, in particular, may be difficult to type on resulting in grammatical and spelling errors. Moreover, emails inherit the same issues, and offer an additional challenge as authors often change their vocabulary based on the content and addressee. Even spam detection is further complicated if spammers change their use of common spam-related terms to purposely circumvent spam-language filters.

Additionally, multilingual authorship attribution tasks are only as good as the available software that can support the considered language. Many non-Western languages do not contain word boundaries, which complicates word segmentation. Hence, some lexical-level features may be unavailable. Additionally, automatic word segmentation tools reportedly incur high computation cost and cannot operate on words not in their dictionaries. As a result, languages with large vocabularies could result in data sparsity. To further demonstrate the difficulties associated with various languages, brief descriptions are offered in Table 11.

Many of the open problems found in authorship attribution studies also exist in authorship verification. There is an inability to properly scale; what is easy in small, balanced datasets is significantly harder otherwise, especially in verification settings where the true author may or may not have representative samples in the database. Moreover, there is a correlation between training set size and verification errors; smaller training sets (in terms of document length) result in higher verification error. This complicates large-scale authorship verification tasks. Error reporting is also limited across both authorship attribution and authorship verification. Common metrics typical of verification settings, such as false accept and false reject rates are often ignored [26].

Authorship profiling is often complicated due to the lack of metadata for the authors. In large datasets, specifically those retrieved from online sources, obtaining such data is difficult and, in some cases, impossible due to privacy regulations [138]. When this information is available, researchers may have to deal with data corruption and additional noise due to fictitious profiles. An open problem that deserves attention is how to properly obtain large, labeled datasets that allow profiling evaluations with automated filtering of noise. Further, smaller feature spaces aid in reducing execution time. For tackling authorship profiling in large domains, this would certainly desired. Finally, it appears that, similar to authorship attribution, performance is correlated with the amount of data available; that is, performance increases as more training data is available. It would interesting to see how authorship profiling methods scale to larger datasets [140].

Stylochronometry often suffers from the inability to assign a time or date to documents with absolute certainty [47]. Researchers must also determine how stylistic clues that suggest a particular time period vary across various genres and languages. It may also be unclear if changes in authorial style are due to changes in time (which correlate with an author's style) or in changes to the language itself [31], and methods that properly handle time-dependent stylistic changes are necessary.

Adversarial stylometry has its own research challenges as well. For instance, style obfuscation is easier when authors are allowed to write freely. It would be interesting to study obfuscation in stricter domains, such as in scientific articles [24]. Style translation may be harder under manuscript guidelines and authorship attribution performance could improve in these settings due to the lack of strong obfuscation transformations. It is also nontrivial to identify which features should be automatically transformed and which require manual inspection for developing the most effective adversarial approach [24]. Further, machine translation currently results in semantic changes; advancement of machine translation services could potentially strengthen machine translation performance as an adversarial stylometry technique. Additionally, shallow obfuscation involves changing feature frequencies, but changing the frequency of terms may affect term dependencies. This consequently affects the meaning of the document. Further research efforts should consider the interdependency between term frequency and semantics

Table 11. Linguistic Characteristics

Language	Description
Thai	According to Reference [113], the Thai language may pose additional difficulties beyond those found in English, Arabic, and Chinese texts. Thai texts have no word boundaries, first-person pronouns can correspond with gender, sentences may end with gender suffixes, and typing words can be difficult given the numerous consonants, vowels, tone marks, and other special symbols.
Chinese	The Chinese language may be challenging in stylometry tasks for several reasons. First, according to Reference [185], ancient and modern Chinese differ in function words due to political changes. Reference [188] states that Chinese has a large vocabulary and no explicit word boundary. POS feature extraction is also primitive for the Chinese language. Reference [103] states that the average word length in Chinese is approximately two characters.
Dutch	According to Reference [126], Flemish Dutch chat messages are influenced by informal writing or writing quickly. The former resulted in regional varieties of the language, while the latter led to different abbreviations. Preliminary experiments found regional dialects to be challenging. Dutch chat messages also include various emoticons.
Arabic	Reference [4] describes various characteristics of Arabic poems. These poems are considered rhymed (i.e., measured) or prose. Rhymed poems are constructed with 15 meters, or seas, measured in syllables. Each sea is restricted in the number of syllables. Additionally, every second part of a verse has to end in the same rhyme. However, it was found that rhyme is not a strong indication of authorship.
Lithuanian	According to Reference [76], the Lithuanian language is rich in vocabulary and morphology, highly inflective (inflection changes words to show expression, such as aspect, voice, or gender), has a complicated word derivation system, free word-order in sentences, and some missing diacritics.
Telugu	Reference [120] describes the Telugu language as one of the 150 languages spoken in India. Indian languages are inflective and derivations lead to scarce feature spaces. The Telugu language in particular is complex and maps English phrases to single words.
Turkish	Reference [31] describes the Turkish language as constituent-free (i.e., grouping of words can change order depending on text flow) having complex word structures with inflectional suffixes. It's alphabet has 29 letters, 8 vowels, and 21 consonants. Some words are translated into full English sentences. In Reference [94], the Turkish language is described as agglutinative, having morphemes stringed together without changing spelling or phonetics.

[75]. Shallow obfuscation also relies on a fair amount of authorship attribution features. However, authorship attribution features are limited in shorter texts. Thus, investigation of adversarial stylometry in shorter documents remains a research challenge [75].

Overall, Daelemans statement from Reference [38] perfectly summarizes these research challenges: "We are looking not just for a system that reaches a certain target accuracy in a task, but for explanations, and for systems that are scalable, and that generalize over different genres and topics in their author identification and profiling results. It seems clear that a systematic study of the components and concepts of style will only be possible by collecting a large balanced dataset for each language of a type that doesn't yet exist in current benchmark efforts."

7 SUMMARY

This survey provides a review of past and present techniques for the various subtasks in the field of stylometry. Specifically, authorship attribution, verification, and profiling, along with stylochronometry and adversarial stylometry, are defined. These subtasks provide the necessary tools to resolve authorship disputes, determine personal characteristics of authors, detect stylistic similarities between several unlabelled texts, estimate the time period in which a document was written, and avoid identification via style translations. All of these tasks rely on a range of representations that attempt to model or explain the authorial style of documents. These representations generally include lexical, syntactic, semantic, structural, and domain-specific features.

Initially, authorial analysis relied on corpora with large training sets and a small set of candidate authors. However, the need to consider more in-the-wild datasets has shifted the focus to large sets of unbalanced and noisy text samples. Since the inclusion of advanced techniques, such as machine-learning algorithms and natural language processing tools, stylometry studies have significantly evolved. For instance, experimental approaches such as application of machine-learning classifiers, intertextual distance models, various compression algorithms, and probabilistic models are useful for a variety of authorship analysis tasks within multiple languages and document types, among others. However, several research challenges remain, such as the lack of sufficient data within a single corpus, exploring and leveraging feature-level correlations, handling performance decreases in unconstrained conditions, and challenges specific to each subtask.

Currently, there is not a general consensus on an optimal feature set and most studies do not consider benchmark datasets. As a result, most methods will fail to scale to larger corpora. There is also little evidence to support consistent performance under various settings, such as variations in topics and genres and small text samples. Evaluation of performance accuracy using 14 algorithms from the research literature further demonstrates these challenges, particularly for authorship attribution and verification. Results indicate that among several approaches in the research literature, many of these may be inefficient at supporting larger sets of candidate authors with smaller text samples. Results also suggest that bag-of-words features may largely overlap among authors, particularly as the number of authors increases.

Finally, while we review several subtasks, multiple experimental approaches (along with available software tools to implement these approaches), and recent approaches, we purposely emphasize the significant challenge of achieving high performance in non-ideal conditions for two reasons. First, it is important that authorial analysis techniques remain relevant. Because we can expect an increase in text samples that originate from online sources, it is irresponsible to assume that text samples will be a certain length, an author will use the same identity across multiple documents, and so on. It is these inconsistencies that tend to complicate stylometry; deriving techniques robust against these factors is necessary. Second, to efficiently apply stylometry to practical applications, researchers must consider a variety of cases. For instance, to continuously verify mobile devices, authorship verification must remain robust with small and possibly grammatically incorrect text samples. However, as the performance analysis demonstrates, this remains as an open research challenge. Thus, this survey aims to provide a holistic review of stylometry, while encouraging exploration of new and/or hybrid approaches that improve performance in a variety of conditions.

REFERENCES

[1] Novino Nirmala, Kyung-Ah Sohn, and T. S. Chung. 2015. A graph model-based author attribution technique for single-class e-mail classification. In *Proceedings of the 2015 IEEE/ACIS 14th International Conference on Computer and Information Science (ICIS'15)*. 191–196.

[2] Ahmed Abbasi and Hsinchun Chen. 2008. Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace. *ACM Trans. Inf. Syst.* 26, 2, Article 7 (April 2008), 29 pages.

[3] S. Afroz, A. C. Islam, A. Stolerman, R. Greenstadt, and D. McCoy. 2014. Doppelgänger finder: Taking stylometry to the underground. In *Proceedings of the 2014 IEEE Symposium on Security and Privacy*. 212–226.

[4] A. F. Ahmed, R. Mohamed, B. Mostafa, and A. S. Mohammed. 2015. Authorship attribution in arabic poetry. In *Proceedings of the 2015 10th International Conference on Intelligent Systems: Theories and Applications (SITA'15)*. 1–6.

[5] K. Alsmearat, M. Al-Ayyoub, and R. Al-Shalabi. 2014. An extensive study of the bag-of-words approach for gender identification of arabic articles. In *Proceedings of the 2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA'14)*. 601–608. DOI: http://dx.doi.org/10.1109/AICCSA.2014.7073254

[6] K. Alsmearat, M. Shehab, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan. 2015. Emotion analysis of arabic articles and its impact on identifying the author's gender. In *Proceedings of the 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA'15)*. 1–6. DOI: http://dx.doi.org/10.1109/AICCSA.2015.7507196

[7] S. M. Alzahrani, N. Salim, and A. Abraham. 2012. Understanding plagiarism linguistic patterns, textual features, and detection methods. *IEEE Trans. Syst. Man Cybernet. Part C (Appl. Rev.)* 42, 2 (March 2012), 133–149. DOI: http://dx.doi.org/10.1109/TSMCC.2011.2134847

[8] Shlomo Argamon. 2008. Interpreting Burrows's Delta: Geometric and probabilistic foundations. *Lit. Linguist. Comput.* 23, 2 (2008), 131. DOI: http://dx.doi.org/10.1093/llc/fqn003 arXiv:/oup/backfile/content_public/journal/dsh/23/2/10.1093/llc/fqn003/3/fqn003.pdf.

[9] Shlomo Argamon and Patrick Juola. 2011. Overview of the international authorship identification competition at pan-2011. In *Proceedings of Clef (notebook Papers/labs/workshop)*.

[10] Shlomo Argamon, Moshe Koppel, James W. Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Commun. ACM* 52, 2 (Feb. 2009), 119–123. DOI: http://dx.doi.org/10.1145/1461928.1461959

[11] R. Arun, V. Suresh, and C. E. V. Madhavan. 2009. Stopword graphs and authorship attribution in text corpora. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC'09)*. 192–196.

[12] H. Baayen, H. van Halteren, and F. Tweedie. 1996. Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. *Lit. Linguist. Comput.* 11, 3 (1996), 121. DOI: http://dx.doi.org/10.1093/llc/11.3.121 arXiv:/oup/backfile/content public/journal/dsh/11/3/10.1093/llc/11.3.121/2/110121.pdf

[13] R. Harald Baayen. 2001. *Word Frequency Distributions*. Vol. 18. Springer Science & Business Media.

[14] Eric Backer and Peter van Kranenburg. 2005. On musical stylometry—a pattern recognition approach. *Pattern Recogn. Lett.* 26, 3 (2005), 299–309. DOI: http://dx.doi.org/10.1016/j.patrec.2004.10.016 In Memoriam: Azriel Rosenfeld.

[15] Murray R. Barrick and Michael K. Mount. 1991. The big five personality dimensions and job performance: A meta-analysis. *Person. Psychol.* 44, 1 (1991), 1–26.

[16] Yasemin Bay and Erbuğ Çelebi. 2016. Feature selection for enhanced author identification of Turkish text. (2016), 371–379.

[17] Abdellghani Bellaachia and Edward Jimenez. 2009. Exploring performance-based music attributes for the stylometric analysis. *World Acad. Sci. Eng. Technol.* 3, 55 (2009), 468–70.

[18] Dario Benedetto, Emanuele Caglioti, and Vittorio Loreto. 2002. Language trees and zipping. *Phys. Rev. Lett.* 88, 4 (2002), 048702.

[19] Steven Benzel. 2015. A simple stylometric comparator: Nifty assignment. *J. Comput. Sci. Coll.* 31, 2 (Dec. 2015), 283–284.

[20] Steven Bird. 2006. Nltk: The natural language toolkit. In *Proceedings of the Conference on Computational Linguistics on Interactive Presentation Sessions (COLING/ACL'06)*. Association for Computational Linguistics, Stroudsburg, PA, 69–72.

[21] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*. O'Reilly Media, Inc.

[22] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.* 3, Jan (2003), 993–1022.

[23] I. N. Bozkurt, O. Baglioglu, and E. Uyar. 2007. Authorship attribution. In *Proceedings of the 22nd International Symposium on Computer and Information Sciences (ISCIS'07)*. 1–5.

[24] Michael Brennan, Sadia Afroz, and Rachel Greenstadt. 2012. Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. *ACM Trans. Inf. Syst. Secur.* 15, 3, Article 12 (Nov. 2012), 22 pages.

[25] M. L. Brocardo and I. Traore. 2014. Continuous authentication using micro-messages. In *Proceedings of the 2014 12th Annual International Conference on Privacy, Security and Trust*. 179–188. DOI: http://dx.doi.org/10.1109/PST.2014.6890938

[26] Marcelo Luiz Brocardo, Issa Traore, Sherif Saad, and Isaac Woungang. 2013. Authorship verification for short messages using stylometry. In *Proceedings of the 2013 International Conference on Computer, Information and Telecommunication Systems (CITS'13)*. IEEE, 1–6.

[27] Marcelo Luiz Brocardo, Issa Traore, Isaac Woungang, and Mohammad S. Obaidat. 2017. Authorship verification using deep belief network systems. *Int. J. Commun. Syst.* (2017). DOI: http://dx.doi.org/10.1002/dac.3259 e3259 dac.3259.

[28] John Burrows. 2002. "Delta": A measure of stylistic difference and a guide to likely authorship. *Lit. Linguist. Comput.* 17, 3 (2002), 267–287. DOI: http://dx.doi.org/10.1093/llc/17.3.267 arXiv:http://llc.oxfordjournals.org/content/17/3/267.full.pdf+html

[29] Kevin Burton, Akshay Java, and Ian Soboroff. 2009. The ICWSM 2009 spinn3r dataset. In *Proceedings of the 3rd Annual Conference on Weblogs and Social Media (ICWSM'09)*. AAAI.

[30] Kevin Burton, Niels Kasch, and Ian Soboroff. 2011. The ICWSM 2011 spinn3r dataset. In *Proceedings of the Annual Conference on Weblogs and Social Media (ICWSM'11)*.

[31] Fazli Can and Jon M. Patton. 2004. Change of writing style with time. *Comput. Human.* 38, 1 (2004), 61–82.

[32] Fazli Can and Jon M. Patton. 2010. Change of word characteristics in 20th-century turkish literature: A statistical analysis. *J. Quant. Linguist.* 17, 3 (2010), 167–190. DOI: http://dx.doi.org/10.1080/09296174.2010.485444

[33] Omar Canales, Vinnie Monaco, Thomas Murphy, Edyta Zych, John Stewart, Charles Tappert Alex Castro, Ola Sotoye, Linda Torres, and Greg Truley. 2011. A stylometry system for authenticating students taking online tests. In *Proceedings of Student-Faculty Research Day, CSIS. Pace University* (2011).

[34] Tanmoy Chakraborty and Sivaji Bandyopadhyay. 2010. Authorship identification using stylometry analysis: A CRF-based approach. In *Proceedings of IEEE Cascom Postgraduate Student Paper Conference, Jadavpur University, Kolkata.* 66–69.

[35] Cindy Chung and James W. Pennebaker. 2007. The psychological functions of function words. *Social Communication* (2007), 343–359.

[36] Jonathan H. Clark and Charles J. Hannon. 2007. A classifier system for author recognition using synonym-based features. In *Proceedings of the 6th Mexican International Conference on Artificial Intelligence (MICAI'07)*. Alexander Gelbukhand Ángel Fernando Kuri Morales (Eds.). Springer, Berlin, 839–849.

[37] Rosa María Coyotl-Morales, Luis Villaseñor-Pineda, Manuel Montes-y Gómez, and Paolo Rosso. 2006. *Authorship Attribution Using Word Sequences*. Springer, Berlin, 844–853. DOI: http://dx.doi.org/10.1007/11892755_87

[38] Walter Daelemans. 2013. *Explanation in Computational Stylometry*. Springer, Berlin, 451–462. DOI: http://dx.doi.org/10.1007/978-3-642-37256-8_37

[39] O. de Vel, A. Anderson, M. Corney, and G. Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Rec.* 30, 4 (Dec. 2001), 55–64. DOI: http://dx.doi.org/10.1145/604264.604272

[40] Rémi De Zoeten. 2015. Computational stylometry in adversarial settings. Master of Science in Artificial Intelligence Thesis, University of Amsterdam. https://esc.fnwi.uva.nl/thesis/centraal/files/f1650865434.pdf.

[41] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Appl. Intell.* 19, 1 (2003), 109–123.

[42] Maciej Eder. 2010. Does size matter? authorship attribution, small samples, big problem. *Proceedings of Digital Humanities* (2010), 132–135.

[43] Sara El Manar El and Ismail Kassou. 2014. Authorship analysis studies: A survey. *Int. J. Comput. Appl.* 86, 12 (2014).

[44] Hugo Jair Escalante, Thamar Solorio, and Manuel Montes-y Gómez. 2011. Local histograms of character N-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1 (HLT'11)*. Association for Computational Linguistics, Stroudsburg, PA, 288–298. Retrieved from http://dl.acm.org/citation.cfm?id=2002472.2002510.

[45] Song Feng, Ritwik Banerjee, and Yejin Choi. 2012. Syntactic stylometry for deception detection. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers—Volume 2 (ACL'12)*. Association for Computational Linguistics, Stroudsburg, PA, 171–175. Retrieved from http://dl.acm.org/citation.cfm?id=2390665.2390708.

[46] Vanessa Wei Feng and Graeme Hirst. 2013. Authorship verification with entity coherence and other rich linguistic features notebook for PAN. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'13)*.

[47] Richard S. Forsyth. 1999. Stylochronometry with substrings, or: A poet young and old. *Lit. Linguist. Comput.* 14, 4 (1999), 467–478.

[48] Georgia Frantzeskou, Efstathios Stamatatos, Stefanos Gritzalis, Carole E. Chaski, and Blake Stephen Howald. 2007. Identifying authorship by byte-level n-grams: The source code author profile (scap) method. *Int. J. Dig. Evidence* 6, 1 (2007), 1–18.

[49] A. Fridman, A. Stolerman, S. Acharya, P. Brennan, P. Juola, R. Greenstadt, and M. Kam. 2013. Decision fusion for multimodal active authentication. *IT Professional* 15, 4 (July 2013), 29–33. DOI: http://dx.doi.org/10.1109/MITP.2013.53

[50] L. Fridman, S. Weber, R. Greenstadt, and M. Kam. 2016. Active authentication on mobile devices via stylometry, application usage, web browsing, and GPS location. *IEEE Syst. J.* PP, 99 (2016), 1–9.

[51] Zhenhao Ge and Yufang Sun. 2016. Domain specific author attribution based on feedforward neural network language models. *Arxiv:1602.07393* (2016).

[52] Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *Lit. Linguist. Comput.* 22, 3 (2007), 251. DOI: http://dx.doi.org/10.1093/llc/fqm020

[53] Andreas Grivas, Anastasia Krithara, and George Giannakopoulos. 2015. Author profiling using stylometric and structural feature groupings. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'15)*.

[54] Stefan Gruner and Stuart Naven. 2005. Tool support for plagiarism detection in text documents. In *Proceedings of the 2005 ACM Symposium on Applied Computing (SAC'05)*. ACM, New York, NY, 776–781. DOI: http://dx.doi.org/10.1145/1066677.1066854

[55] Oren Halvani, Christian Winter, and Anika Pflug. 2016. Authorship verification for different languages, genres and topics. *Digital Investigation* 16, Supplement (2016), S33–S43.

[56] Jonathan Herz and Abdelghani Bellaachia. 2014. The authorship of audacity: Data mining and stylometric analysis of barack obama speeches. In *Proceedings of the International Conference on Data Mining (DMIN'14)*. 1.

[57] R. Hinh, S. Shin, and J. Taylor. 2016. Using frame semantics in authorship attribution. In *Proceedings of the 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC'16)*. 004093–004098. DOI: http://dx.doi.org/10.1109/SMC.2016.7844873

[58] David I. Holmes. 1994. Authorship attribution. *Comput. Human.* 28, 2 (1994), 87–106. DOI: http://dx.doi.org/10.1007/BF01830689

[59] David I. Holmes. 1998. The evolution of stylometry in humanities scholarship. *Lit. Linguist. Comput.* 13, 3 (1998), 111. DOI: http://dx.doi.org/10.1093/llc/13.3.111

[60] D. I. Holmes and R. S. Forsyth. 1995. The federalist revisited: New directions in authorship attribution. *Lit. Linguist. Comput.* 10, 2 (1995), 111. DOI: http://dx.doi.org/10.1093/llc/10.2.111

[61] David I. Holmes and Judit Kardos. 2003. Who was the author? An introduction to stylometry. *Chance* 16, 2 (2003), 5–8.

[62] David L. Hoover. 2003. Another perspective on vocabulary richness. *Comput. Human.* 37, 2 (2003), 151–178.

[63] John Houvardas and Efstathios Stamatatos. 2006. N-gram feature selection for authorship identification. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA'06)*, Jérôme Euzenatand John Domingue (Eds.). Springer, Berlin, 77–86.

[64] Faliang Huang, Chaoxiong Li, and Li Lin. 2014. Identifying gender of microblog users based on message mining. In *Web-Age Information Management (WAIM'14)*, F. Li, G. Li, S. Hwang, B. Yao, and Z. Zhang (Eds). Lecture Notes in Computer Science, vol. 8485. Springer, Cham.

[65] C. R. Jacobsen and M. Nielsen. 2013. Stylometry of paintings using hidden Markov modelling of contourlet transforms. *Signal Process.* 93, 3 (2013), 579–591. DOI: http://dx.doi.org/10.1016/j.sigpro.2012.09.019 Image Processing for Digital Art Work.

[66] S. Jafarpour, G. Polatkan, E. Brevdo, S. Hughes, A. Brasoveanu, and I. Daubechies. 2009. Stylistic analysis of paintings using wavelets and machine learning. In *Proceedinsg of the 2009 17th European Signal Processing Conference*. 1220–1224.

[67] Magdalena Jankowska, Vlado Kešelj, and Evangelos Milios. 2013. CNG text classification for authorship profiling task. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'13)*.

[68] Fotis Jannidis, Steffen Pielström, Christof Schöch, and Thorsten Vitt. 2015. Improving Burrows' Delta—an empirical evaluation of text distance measures. In *Proceedings of the Digital Humanities Conference*.

[69] Matthew L. Jockers and Daniela M. Witten. 2010. A comparative study of machine-learning methods for authorship attribution. *Lit. Linguist. Comput.* (2010).

[70] Patrick Juola. 2007. Future trends in authorship attribution. In *Proceedings of the IFIP International Conference on Digital Forensics, Advances in Digital Forensics III*, Philip Craigerand Sujeet Shenoi (Eds.). Springer, New York, 119–132.

[71] Patrick Juola. 2008. Authorship attribution. *Found. Trends Info. Retr.* 1, 3 (2008), 233–334. DOI: http://dx.doi.org/10.1561/1500000005

[72] Patrick Juola. 2012. An overview of the traditional authorship attribution subtask. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'12) (Online Working Notes/Labs/Workshop)*.

[73] Patrick Juola and Efstathios Stamatatos. 2013. Overview of the author identification task at PAN 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'13) (Working Notes)*.

[74] Patrick Juola and Darren Vescovi. 2011. Analyzing stylometric approaches to author obfuscation. In *Proceedings of the 7th IFIP WG 11.9 International Conference on Digital Forensics, Advances in Digital Forensics VII*, Gilbert Peterson and Sujeet Shenoi (Eds.). Springer, Berlin, 115–125.

[75] Gary Kacmarcik and Michael Gamon. 2006. Obfuscating document stylometry to preserve author anonymity. In *Proceedings of the Conference on Computational Linguistics on Main Conference Poster Sessions (COLING/ACL'06)*. Association for Computational Linguistics, 444–451.

[76] Jurgita Kapočiūtė-Dzikienė, Andrius Utka, and Ligita Šarkutė. 2015. Authorship attribution of internet comments with thousand candidate authors. In *Information and Software Technologies. Communications in Computer and Information Science*, G. Dregvaite and R. Damasevicius (Eds), vol. 538. Springer, Cham.

[77] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics (PACLING'03)*, Vol. 3. 255–264.

[78] Dmitry V. Khmelev and William J. Teahan. 2003. A repetition-based measure for verification of text collections and for text categorization. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval (SIGIR'03)*. ACM, New York, NY, 104–110. DOI: http://dx.doi.org/10.1145/860435.860456

[79] M. Khonji, Y. Iraqi, and A. Jones. 2015. An evaluation of authorship attribution using random forests. In *Proceedings of the 2015 International Conference on Information and Communication Technology Research (ICTRC'15)*. 68–71.

[80] F. Khosmood and R. Levinson. 2010. Automatic synonym and phrase replacement show promise for style transformation. In *Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications (ICMLA'10)*. 958–961.

[81] B. Kjell. 1994. Authorship attribution of text samples using neural networks and Bayesian classifiers. In *Proceedings of the 1994 IEEE International Conference on Systems, Man, and Cybernetics, 1994. Humans, Information and Technology*, Vol. 2. 1660–1664.

[82] Carmen Klaussner and Carl Vogel. 2015. Stylochronometry: Timeline prediction in stylometric analysis. In *Research and Development in Intelligent Systems XXXII (SGAI'15)*, M. Bramer and M. Petridis (Eds). Springer, Cham.

[83] Bryan Klimt and Yiming Yang. 2004. *The Enron Corpus: A New Dataset for Email Classification Research*. Springer, Berlin, 217–226. DOI: http://dx.doi.org/10.1007/978-3-540-30115-8_22

[84] Moshe Koppel and Jonathan Schler. 2003. Exploiting stylistic idiosyncrasies for authorship attribution. In *Proceedings of the Workshop on Computational Approaches to Style Analysis and Synthesis (IJCAI'03)*. 69–72.

[85] Moshe Koppel and Jonathan Schler. 2004. Authorship verification as a one-class classification problem. In *Proceedings of the 21st International Conference on Machine Learning*. ACM, 62.

[86] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2009. Computational methods in authorship attribution. *J. Amer. Soc. Info. Sci. Technol.* 60, 1 (2009), 9–26. DOI: http://dx.doi.org/10.1002/asi.20961

[87] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Lang. Res. Eval.* 45, 1 (2011), 83–94. DOI: http://dx.doi.org/10.1007/s10579-009-9111-2

[88] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *J. Mach. Learn. Res.* 8, Jun (2007), 1261–1276.

[89] Moshe Koppel and Yaron Winter. 2014. Determining if two documents are written by the same author. *J. Assoc. Info. Sci. Technol.* 65, 1 (2014), 178–187.

[90] Markus Krause. 2014. A behavioral biometrics-based authentication method for MOOC's that is robust against imitation attempts. In *Proceedings of the First ACM Conference on Learning @ Scale Conference (L@S'14)*. ACM, New York, NY, 201–202. DOI: http://dx.doi.org/10.1145/2556325.2567881

[91] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev. 2001. Using literal and grammatical statistics for authorship attribution. *Probl. Info. Trans.* 37, 2 (2001), 172–184.

[92] A. M. Kuruvilla and S. Varghese. 2015. A detection system to counter identity deception in social media applications. In *Proceedings of the 2015 International Conference on Circuit, Power and Computing Technologies (ICCPCT'15)*. 1–5.

[93] A. O. Kusakci. 2012. Authorship attribution using committee machines with k-nearest neighbors rated voting. In *Proceedings of the 2012 11th Symposium on Neural Network Applications in Electrical Engineering (NEUREL'12)*. 161–166.

[94] R. S. Kuzu, K. Balci, and A. A. Salah. 2016. Authorship recognition in a multiparty chat scenario. In *Proceedings of the 2016 4th International Conference on Biometrics and Forensics (IWBF'16)*. 1–6.

[95] Cyril Labbé and Dominique Labbé. 2001. Inter-textual distance and authorship attribution Corneille and Molire. *J. Quant. Linguist.* 8, 3 (2001), 213–231. DOI: http://dx.doi.org/10.1076/jqul.8.3.213.4100

[96] Cyril Labbé and Dominique Labbé. 2006. A tool for literary studies: Intertextual distance and tree classification. *Lit. Linguist. Comput.* 21, 3 (2006), 311. DOI: http://dx.doi.org/10.1093/llc/fqi063

[97] Dominique Labbé. 2007. Experiments on authorship attribution by intertextual distance in english*. *J. Quant. Linguist.* 14, 1 (2007), 33–80. DOI: http://dx.doi.org/10.1080/09296170600850601

[98] R. Layton, P. Watters, and R. Dazeley. 2010. Authorship attribution for twitter in 140 characters or less. In *Proceedings of the 2010 2nd Cybercrime and Trustworthy Computing Workshop*. 1–8. DOI: http://dx.doi.org/10.1109/CTC.2010.17

[99] Fabio Leuzzi, Stefano Ferilli, and Fulvio Rotella. 2014. A relational unsupervised approach to author identification. In *New Frontiers in Mining Complex Patterns (NFMCP'13)*, A. Appice, M. Ceci, C. Loglisci, G. Manco, E. Masciari, and Z. Ras (Eds), Lecture Notes in Computer Science, vol. 8399. Springer, Cham.

[100] Jenny S. Li. 2015. An investigation of authorship authentication in short messages from a social networking site. ETD Collection for Pace University. Paper AAI3711057. http://digitalcommons.pace.edu/dissertations/AAI3711057.

[101] Wee-Yong Lim, Jonathan Goh, and Vrizlynn L. L. Thing. 2013. Content-centric age and gender profiling. *Proceedings of the Notebook for PAN at the Conference and Labs of the Evaluation Forum (CLEF'13)*.

[102] Sanya Liu, Zhi Liu, Jianwen Sun, and Lin Liu. 2011. Application of synergetic neural network in online writeprint identification. *Int. J. Dig. Cont. Technol. Appl.* 5, 3 (2011), 126–135.

[103] W. Liu, B. Allison, D. Guthrie, and L. Guthrie. 2007. Chinese text classification without automatic word segmentation. In *Proceedings of the 6th International Conference on Advanced Language Processing and Web Information Technology (ALPIT'07)*. 45–50.

[104] Edward Loper and Steven Bird. 2002. NLTK: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics—Volume 1 (ETMTNLP'02)*. 63–70.

[105] Kim Luyckx and Walter Daelemans. 2008. Authorship attribution and verification with many authors and limited data. In *Proceedings of the 22nd International Conference on Computational Linguistics—Volume 1 (COLING'08)*. 513–520. http://dl.acm.org/citation.cfm?id=1599081.1599146.

[106] Kim Luyckx and Walter Daelemans. 2008. Personae: A corpus for author and personality prediction from text. *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC'08)* (2008).

[107] Kim Luyckx and Walter Daelemans. 2011. The effect of author set size and data size in authorship attribution. *Lit. Linguist. Comput.* 26, 1 (2011), 35–55.

[108] Jianbin Ma, Guifa Teng, Yuxin Zhang, Yueli Li, and Ying Li. 2009. A cybercrime forensic method for chinese web information authorship analysis. In *Proceedings of the Pacific Asia Workshop on Intelligence and Security Informatics (PAISI'09)*, Hsinchun Chen, Christopher C. Yang, Michael Chau, and Shu-Hsing Li (Eds.). Springer, Berlin, 14–24.

[109] David Madigan, Alexander Genkin, David D. Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005. Author identification on the large scale. In *Proceedings of the Meeting of the Classification Society of North America*. 13.

[110] M. B. Malyutov. 2006. *Authorship Attribution of Texts: A Review*. Springer, Berlin, 362–380. DOI: http://dx.doi.org/10.1007/11889342_20

[111] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The stanford corenlp natural language processing toolkit. In *Proceedings of the ACL Conference on System Demonstrations*. 55–60.

[112] Yuval Marton, Ning Wu, and Lisa Hellerstein. 2005. On compression-based text classification. In *Proceedings of the European Conference on Information Retrieval*. Springer, 300–314.

[113] R. Marukatat, R. Somkiadcharoen, R. Nalintasnai, and T. Aramboonpong. 2014. Authorship attribution analysis of thai online messages. In *Proceedings of the 2014 International Conference on Information Science Applications (ICISA'14)*. 1–4.

[114] Robert A. J. Matthews and Thomas V. N. Merriam. 1993. Neural computation in stylometry I: An application to the works of shakespeare and fletcher. *Lit. Linguist. Comput.* 8, 4 (1993), 203. DOI: http://dx.doi.org/10.1093/llc/8.4.203

[115] Thomas V. N. Merriam and Robert A. J. Matthews. 1994. Neural computation in stylometry II: An application to the works of Shakespeare and Marlowe. *Lit. Linguist. Comput.* 9, 1 (1994), 1. DOI: http://dx.doi.org/10.1093/llc/9.1.1

[116] George K. Mikros. 2012. Authorship attribution and gender identification in greek blogs. *Methods Appl. Quant. Linguist.* 21 (2012).

[117] Frederick Mosteller and David Wallace. 1964. Inference and disputed authorship: The Federalist. (1964).

[118] Frederick Mosteller and David L. Wallace. 1962. Notes on an authorship problem. In *Proceedings of a Harvard Symposium on Digital Computers and Their Applications*. 163–197.

[119] Frederick Mosteller and David L. Wallace. 1963. Inference in an authorship problem: A comparative study of discrimination methods applied to the authorship of the disputed federalist papers. *J. Amer. Statist. Assoc.* 58, 302 (1963), 275–309.

[120] S. Nagaprasad, T. Raghunadha Reddy, P. Vijayapal Reddy, A. Vinaya Babu, and B. VishnuVardhan. 2015. Empirical evaluations using character and word n-grams on authorship attribution for Telugu text. In *Intelligent Computing and Applications. Advances in Intelligent Systems and Computing*, D. Mandal, R. Kar, S. Das, and B. Panigrahi (Eds), vol. 343. Springer, New Delhi.

[121] A. Narayanan, H. Paskov, N. Z. Gong, J. Bethencourt, E. Stefanov, E. C. R. Shin, and D. Song. 2012. On the feasibility of internet-scale author identification. In *Proceedings of the 2012 IEEE Symposium on Security and Privacy*. 300–314. DOI: http://dx.doi.org/10.1109/SP.2012.46

[122] Dinh Phuc Nguyen. 2014. Obfuscation techniques for java source code. In *Proceedings of the URECA@NTU 2013-14*. Student research paper, Nanyang Technological University.

[123] Michael P. Oakes. 2014. *Literary Detective Work on the Computer*. Vol. 12. John Benjamins Publishing Company.

[124] P. K. Pateriya, Lakshmi, and G. Raj. 2014. A pragmatic validation of stylometric techniques using BPA. In *Proceedings of the 2014 5th International Conference—Confluence: The Next Generation Information Technology Summit (CONFLUENCE'14)*. 124–131. DOI: http://dx.doi.org/10.1109/CONFLUENCE.2014.6949275

[125] Srikanta Patnaik, S. Naga Prasad, V. B. Narsimha, P. Vijayapal Reddy, and A. Vinaya Babu. 2015. International conference on computer, communication and convergence (ICCC'15) influence of lexical, syntactic and structural features and their combination on authorship attribution for Telugu text. *Proced. Comput. Sci.* 48 (2015), 58–64. DOI: http://dx.doi.org/10.1016/j.procs.2015.04.110

[126] Claudia Peersman, Walter Daelemans, and Leona Van Vaerenbergh. 2011. Predicting age and gender in online social networks. In *Proceedings of the 3rd International Workshop on Search and Mining User-generated Contents (SMUC'11)*. ACM, New York, NY, 37–44.

[127] Fuchun Peng, Dale Schuurmans, and Shaojun Wang. 2004. Augmenting naive bayes classifiers with statistical language models. *Info. Retr.* 7, 3 (2004), 317–345. DOI: http://dx.doi.org/10.1023/B:INRT.0000011209.19643.e2

[128] Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. 2003. Language independent authorship attribution using character-level language models. In *Proceedings of the 10th Conference on European Chapter of the Association for Computational Linguistics—Volume 1 (EACL'03)*. 267–274.

[129] S. R. Pillay and T. Solorio. 2010. Authorship attribution of web forum posts. In *Proceedings of the eCrime Researchers Summit (eCrime), 2010*. 1–7.

[130] Spyridon Plakias and Efstathios Stamatatos. 2008. Tensor space models for authorship identification. In *Proceedings of the 5th Hellenic Conference on Artificial Intelligence: Theories, Models and Applications (AI, SETN'08)*, John Darzentas, George A. Vouros, Spyros Vosinakis, and Argyris Arnellos (Eds.). Springer, Berlin, 239–249.

[131] J. Posadas-Durán, Ilia Markov, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, Alexander Gelbukh, and Obdulia Pichardo-Lagunas. 2015. Syntactic n-grams as features for the author profiling task. *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF'15)*.

[132] Juan-Pablo Posadas-Duran, Grigori Sidorov, and Ildar Batyrshin. 2014. *Complete Syntactic N-grams as Style Markers for Authorship Attribution*. Springer International Publishing, Cham, 9–17. DOI: http://dx.doi.org/10.1007/978-3-319-13647-9_2

[133] Martin Potthast, Sarah Braun, Tolga Buz, Fabian Duffhauss, Florian Friedrich, Jörg Marvin Gülzow, Jakob Köhler, Winfried Lötzsch, Fabian Müller, Maike Elisa Müller, Robert Paßmann, Bernhard Reinke, Lucas Rettenmeier, Thomas Rometsch, Timo Sommer, Michael Träger, Sebastian Wilhelm, Benno Stein, Efstathios Stamatatos, and Matthias Hagen. 2016. Who wrote the web? Revisiting influential author identification research applicable to information retrieval. In *Advances in Information Retrieval (ECIR'16)*, N. Ferro et al. (Eds), Lecture Notes in Computer Science, vol. 9626. Springer, Cham.

[134] R. Ragel, P. Herath, and U. Senanayake. 2013. Authorship detection of SMS messages using unigrams. In *Proceedings of the 2013 IEEE 8th International Conference on Industrial and Information Systems*. 387–392.

[135] Sindhu Raghavan, Adriana Kovashka, and Raymond Mooney. 2010. Authorship attribution using probabilistic context-free grammars. In *Proceedings of the ACL 2010 Conference Short Papers (ACLShort'10)*. 38–42.

[136] Hoshiladevi Ramnial, Shireen Panchoo, and Sameerchand Pudaruth. 2016. Gender Profiling from PhD theses using k-nearest neighbour and sequential minimal optimisation. *Intelligent Systems Technologies and Applications*. 369–377.

[137] Congzhou He Ramyaa and Khaled Rasheed. 2004. Using machine-learning techniques for stylometry. In *Proceedings of International Conference on Machine Learning*.

[138] Francisco Rangel, Paolo Rosso, Moshe Moshe Koppel, Efstathios Stamatatos, and Giacomo Inches. 2013. Overview of the author profiling task at PAN 2013. In *Proceedings of the Conference and Labs of the Evaluation Forum on Multilingual and Multimodal Information Access Evaluation (CLEF'13)*. 352–365.

[139] Francisco Rangel, Paolo Rosso, Martin Potthast, Benno Stein, and Walter Daelemans. 2015. Overview of the 3rd author profiling task at PAN 2015. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'15)*.

[140] Francisco Rangel, Paolo Rosso, Martin Potthast, Martin Trenkmann, Benno Stein, Ben Verhoeven, Walter Daeleman, et al. 2014. Overview of the 2nd author profiling task at pan 2014. In *Proceedings of the CEUR Workshop*, Vol. 1180, 898–927.

[141] Francisco Rangel, Paolo Rosso, Ben Verhoeven, Walter Daelemans, Martin Potthast, and Benno Stein. 2016. Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations. *Working Notes Papers of the Conference and Labs of the Evaluation Forum (CLEF'16)*.

[142] T. Raghunadha Reddy, B. Vishnu Vardhan, and P. Vijayapal Reddy. 2016. A survey on authorship profiling techniques. *Int. J. Appl. Eng. Res.* 11, 5 (2016), 3092–3102.

[143] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 487–494.

[144] Joseph Rudman. 1997. The state of authorship attribution studies: Some problems and solutions. *Comput. Human.* 31, 4 (1997), 351–365.

[145] Upendra Sapkota, Thamar Solorio, Manuel Montes-y Gómez, and Paolo Rosso. 2013. The use of orthogonal similarity relations in the prediction of authorship. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'13)*, Alexander Gelbukh (Ed.). Springer, Berlin, 463–475.

[146] Jacques Savoy. 2015. Comparative evaluation of term selection functions for authorship attribution. *Dig. Scholar. Human.* 30, 2 (2015), 246–261.
[147] Jonathan Schler, Moshe Koppel, Shlomo Argamon, and James W. Pennebaker. 2006. Effects of age and gender on blogging. In *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, Vol. 6. 199–205.
[148] Yanir Seroussi, Fabian Bohnert, and Ingrid Zukerman. 2012. Authorship attribution with author-aware topic models. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers—Volume 2*. Association for Computational Linguistics, 264–269.
[149] Grigori Sidorov, Francisco Velasquez, Efstathios Stamatatos, Alexander Gelbukh, and Liliana Chanona-Hernndez. 2014. Syntactic N-grams as machine-learning features for natural language processing. *Expert Syst. Appl.* 41, 3 (2014), 853–860. DOI: http://dx.doi.org/10.1016/j.eswa.2013.08.015 Methods and Applications of Artificial and Computational Intelligence.
[150] E. H. Simpson. 1949. Measurement of diversity. Nature 163 (1949).
[151] Tom De Smedt and Walter Daelemans. 2012. Pattern for python. *J. Mach. Learn. Res.* 13, Jun (2012), 2063–2067.
[152] P. W. H. Smith. 2010. Using genetic algorithms in word-vector optimisation. In *Proceedings of the 2010 UK Workshop on Computational Intelligence (UKCI'10)*. 1–5.
[153] Efstathios Stamatatos. 2006. Authorship attribution based on feature set subspacing ensembles. *Int. J. Artific. Intell. Tools* 15, 05 (2006), 823–838. DOI: http://dx.doi.org/10.1142/S0218213006002965
[154] E. Stamatatos. 2007. Author identification using imbalanced and limited training texts. In *Proceeedings of the 18th International Workshop on Database and Expert Systems Applications (DEXA'07)*. 237–241. DOI: http://dx.doi.org/10.1109/DEXA.2007.5
[155] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *J. Amer. Soc. Info. Sci. Technol.* 60, 3 (2009), 538–556.
[156] Efstathios Stamatatos. 2016. Universality of stylistic traits in texts. In *Creativity and Universality in Language*, M. Degli Esposti, E. Altmann, and F. Pachet (Eds), Lecture Notes in Morphogenesis. Springer, Cham.
[157] Efstathios Stamatatos. 2017. Authorship attribution using text distortion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. 1138–1149. Retrieved from http://www.aclweb.org/anthology/E17-1107.
[158] Efstathios Stamatatos, Walter Daelemans, Ben Verhoeven, Patrick Juola, Aurelio López-López, Martin Potthast, and Benno Stein. 2014. Overview of the author identification task at PAN 2014. In *Proceedings of the Conference and Labs of the Evaluation Forum (CLEF'14) (Working Notes)*. 877–897.
[159] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. 2001. Computer-based authorship attribution without lexical measures. *Comput. Human.* 35, 2 (2001), 193–214.
[160] Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015. Overview of the PAN/CLEF 2015 evaluation lab. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, J. Mothe et al. (Eds), Lecture Notes in Computer Science, vol. 9283. Springer, Cham.
[161] Efstathios Stamatatos, Martin Potthast, Francisco Rangel, Paolo Rosso, and Benno Stein. 2015. Overview of the pan/clef 2015 evaluation lab. In *Proceedings of the International Conference of the Cross-Language Evaluation Forum for European Languages*. Springer, 518–538.
[162] Constantina Stamou. 2008. Stylochronometry: Stylistic development, sequence of composition, and relative dating. *Lit. Linguist. Comput.* 23, 2 (2008), 181. DOI: http://dx.doi.org/10.1093/llc/fqm029
[163] Urszula Stańczyk. 2016. The class imbalance problem in construction of training datasets for authorship attribution. In *Man–Machine Interactions 4. Advances in Intelligent Systems and Computing*, A. Gruca, A. Brachman, S. Kozielski, and T. Czachórski (Eds), vol. 391. Springer, Cham.
[164] Benno Stein, Moshe Koppel, and Efstathios Stamatatos. 2007. Plagiarism analysis, authorship identification, and near-duplicate detection PAN'07. *SIGIR Forum* 41, 2 (Dec. 2007), 68–71.
[165] Benno Stein, Nedim Lipka, and Peter Prettenhofer. 2011. Intrinsic plagiarism analysis. *Lang. Res. Eval.* 45, 1 (01 Mar 2011), 63–82. DOI: http://dx.doi.org/10.1007/s10579-010-9115-y
[166] Sterling Stein and Shlomo Argamon. 2006. A mathematical explanation of Burrows's Delta. In *Proceedings of the Digital Humanities Conference*. Citeseer, 207–209.
[167] L. M. Stuart, S. Tazhibayeva, A. R. Wagoner, and J. M. Taylor. 2013. On identifying authors with style. In *Proceedings of the 2013 IEEE International Conference on Systems, Man, and Cybernetics*. 3048–3053.
[168] K. Surendran, O. P. Harilal, P. Hrudya, Prabaharan Poornachandran, and N. K. Suchetha. 2017. *Stylometry Detection Using Deep Learning*. Springer, Singapore, 749–757. DOI: http://dx.doi.org/10.1007/978-981-10-3874-7_71
[169] Ján Švec and Jan Rygl. 2015. Slavonic corpus for stylometry research. *Proceedings of the Conference on Recent Advances in Slavonic Natural Language Processing (RASLAN'15)*, 11.

[170] R. H. R. Tan and F. S. Tsai. 2010. Authorship identification for online text. In *Proceedings of the 2010 International Conference on Cyberworlds (CW'10)*. 155–162.
[171] Yla R. Tausczik and James W. Pennebaker. 2010. The psychological meaning of words: LIWC and computerized text analysis methods. *J. Lang. Soc. Psychol.* 29, 1 (2010), 24–54. DOI: http://dx.doi.org/10.1177/0261927X09351676
[172] William J. Teahan and David J. Harper. 2003. Using compression-based language models for text categorization. In *Language Modeling for Information Retrieval*. Springer, 141–165.
[173] M. F. Tennyson. 2013. A replicated comparative study of source code authorship attribution. In *Proceedings of the 2013 3rd International Workshop on Replication in Empirical Software Engineering Research (RESER'13)*. 76–83.
[174] F. J. Tweedie, S. Singh, and D. I. Holmes. 1996. Neural network applications in stylometry: The federalist papers. *Comput. Human.* 30, 1 (1996), 1–10. DOI: http://dx.doi.org/10.1007/BF00054024
[175] Tanguy Urvoy, Emmanuel Chauveau, Pascal Filoche, and Thomas Lavergne. 2008. Tracking web spam with HTML style similarities. *ACM Trans. Web* 2, 1, Article 3 (March 2008), 28 pages.
[176] Stijn Marinus Van Dongen. 2001. Graph clustering by flow simulation. Doctoral Dissertation, Utrecht University. https://dspace.library.uu.nl/handle/1874/848.
[177] P. Varela, E. Justino, and L. S. Oliveira. 2011. Selecting syntactic attributes for authorship attribution. In *Proceedings of the 2011 International Joint Conference on Neural Networks (IJCNN'11)*. 167–172.
[178] Ben Verhoeven and Walter Daelemans. 2014. Clips stylometry investigation (CSI) corpus: A dutch corpus for the detection of age, gender, personality, sentiment and deception in text. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC'14)*. 3081–3085.
[179] Ben Verhoeven, Walter Daelemans, and Barbara Plank. 2016. Twisty: A multilingual Twitter stylometry corpus for gender and personality profiling. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC'16)*.
[180] Ben Verhoeven, Juan Soler Company, and Walter Daelemans. 2014. Evaluating content-independent features for personality recognition. In *Proceedings of the 2014 ACM Multi Media Workshop on Computational Personality Recognition (WCPR'14)*. ACM, New York, NY, 7–10.
[181] Ben Verhoeven, Iza Škrjanec, and Senja Pollak. 2017. Gender profiling for slovene twitter communication: The influence of gender marking, content and style. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*. 119–125. Retrieved from http://www.aclweb.org/anthology/W17-1418.
[182] Cynthia Whissell. 1996. Traditional and emotional stylometric analysis of the songs of Beatles Paul McCartney and John Lennon. *Comput. Human.* 30, 3 (1996), 257–265.
[183] Min Yang, Dingju Zhu, Yong Tang, and Jingxuan Wang. 2017. Authorship Attribution with Topic Drift Model (2017). Retrieved from https://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14152.
[184] O. Yavanoglu. 2016. Intelligent authorship identification with using Turkish newspapers metadata. In *Proceedings of the 2016 IEEE International Conference on Big Data (BIGDATA'16)*. 1895–1900. DOI: http://dx.doi.org/10.1109/BigData.2016.7840809
[185] Bei Yu. 2012. Function words for chinese authorship attribution. In *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*. Association for Computational Linguistics, 45–53.
[186] Hua-Ping Zhang, Hong-Kui Yu, De-Yi Xiong, and Qun Liu. 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing—Volume 17 (SIGHAN'03)*. Association for Computational Linguistics, Stroudsburg, PA, 184–187. http://dx.doi.org/10.3115/1119250.1119280.
[187] Ying Zhao and Justin Zobel. 2005. Effective and scalable authorship attribution using function words. In *Asia Information Retrieval Symposium*. Springer, 174–189.
[188] Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A framework for authorship identification of online messages: Writing-style features and classification techniques. *J. Amer. Soc. Info. Sci. Technol.* 57, 3 (2006), 378–393. Retrieved from http://dx.doi.org/10.1002/asi.20316.

Received August 2016; revised July 2017; accepted August 2017