IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

SPECIAL SECTION ON ADVANCED DATA MINING METHODS FOR SOCIAL COMPUTING

Received February 17, 2020, accepted March 29, 2020, date of publication April 6, 2020, date of current version April 21, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2985916

Online Social Networks and Writing Styles– A Review of the Multidisciplinary Literature

KAH YEE TAI[1], JASBIR DHALIWAL[1], AND SHAFIZA MOHD SHARIFF[2]
[1]School of Information Technology, Monash University Malaysia, Selangor 47500, Malaysia
[2]Malaysian Institute of Information Technology, Universiti Kuala Lumpur, Kuala Lumpur 50250, Malaysia

Corresponding author: Jasbir Dhaliwal (jasbirkaur.dhaliwal@monash.edu)

ABSTRACT In recent years, author identification has become an active research area, where the major differences are caused by paper or online medium, mode of entry and target audience. Much research has been devoted to analyzing writing styles in handwritten, word-processed and online social networks (OSN) texts. Word processing editors that typically include spell and grammar checkers may influence the writing style as it allows an individual to edit a piece of text to perfection. Thus, similarities may exist between OSN and word-processed texts. Moreover, none of the studies to date have made a detailed comparison of the writing styles across multidisciplinary factors. This paper attempts to close the gap between the writing styles in pre- and post-Internet periods as well as provide an in-depth comparison of the writing styles in OSN texts across three major factors: demographics, personality & behavior, and cybersecurity. The aim is to learn from past literature as we advance these techniques to OSN texts. Thus, in this paper, we also propose a novel machine learning prediction model based on tense morphology, to classify age and gender from English blogs, and the PAN 2013 dataset. This model achieves an accuracy of 94%–98% and 95%–97% for age and gender, respectively.

INDEX TERMS Online social networks, survey, writing styles.

I. INTRODUCTION

Author identification has been an active research field for centuries [1]. The major differences are caused by paper or online medium; mode of entry such as pen, computer keyboard, mobile phone keypad and dictation recorder; and target audience that ranges from teenagers to professionals to senior citizens. The rationale behind this research is that some style characteristics can be extracted from a piece of writing, and those characteristics are unique to an individual. Author identification research looks at authorship attribution and authorship verification. Authorship attribution is defined as the task of seeking an exact author, while that of authorship verification is to determine whether the same author [2] wrote the documents. Users' heuristic behavior will show in their online writings if they spend a great deal of time online. Features such as structural traits, language usage, content markers, and stylistic features can be used to form a feature collection for author identification. Regardless whether individuals write under a pseudonym or authornym or real name, the deviation from their writing style makes them identifiable. Individuals can be profiled based on factors such as demographics [3]; personality [4]; handwritten allographs [5]; left- or right-handiness [6]; and even accent [7].

The importance of identifying and profiling authors is undeniably vital in applications, for example, in forensic profiling [8] and marketing profiling [9]. From a forensics perspective, forensics investigators are interested in knowing the profile of the person writing harassing text messages. Similarly, from a marketing perspective, companies are interested in knowing the demographics and even personalities of people who like or dislike their products by analyzing online social networks (OSN) texts. This is because with the evolution of the Internet, people have the freedom to share, create and find information across different platforms, time and geographical locations. The information can range from dictated text to word-processed text to OSN text. This fact further introduces the issue of anonymity on the Internet: anonymity allows people to hide their real identity information such as name, age, gender, and address and may further give rise to plagiarism, information misuse and illegal information sharing.

The associate editor coordinating the review of this manuscript and approving it for publication was Shirui Pan.

Thus, it is no surprise that the study of writing styles (also used interchangeably with writing style techniques in literature) emerged as a separate discipline in the 20th century [10]. Research in this field continued from stylistics [11] and social science [12]–[14] to cybersecurity [15]–[17]. As an multidisciplinary field, several definitions of writing styles have emerged and have varied over time, country, culture, and disciplines, with the most relevant to this paper being as follows:

- Style as an expression of individuality, subjectivity, or emotions of an author [1], [4], [18], [19].
- Style as an author-function [20], [21]. An author chooses to write under a pseudonym, authornym or real name.
- Style as a deviation. These deviations range from normative-prescriptive ideas to language norms of writing that are based on demographic information [19].
- Style as a language choice. Authors choose to adapt their writing styles according to their conversation partners [22].

In today's age, all the above can be formally defined and measured with technology advancements [23]. Much research has been devoted to analyzing the writing styles in handwritten [6], [23]–[26], word-processed [27]–[29] and OSN texts [30]–[34]. Word processing editors that typically include spell and grammar checkers may influence the writing style as it allows perfection since the word processing editors can help authors to limit the use of idiosyncratic features in their text, which include misspellings and mistakes in grammar. The editors also allow the author to revisit the text for changes in spelling, insertion, and deletion of words and sentences, as well as massive scale movement of chunks of text [35]. This continuous activity erases the unique individual writing styles as it mimics a particular style or format of a document to fit the target audience such as the one provided by one of the popular grammar checker tool, Grammarly[1]. Whereas, such a feat is difficult to be achieved in the handwritten text without several manual rewrites. Compared to handwritten and word-processed text readers who read left-to-right, OSN readers have a shorter attention span, whereby skimming is typically done. Thus, for better readability, OSN content is typically divided into headings, lists, hyperlinks, images, and shorter paragraphs.

However, with the increasing number of spell and grammar checkers that are available online, social media text can be perfected, which may somewhat mirror word-processed text. We are not aware of any research that has reviewed common writing styles in handwritten, word-processed, and OSN texts in a multidisciplinary field. Reference [36] reviewed demographic factors of gender, age and location by combining advanced computational linguistics and machine learning with insights from sociolinguistics. Whereas, [37] reviewed 14 demographic factors of computer science and psychology-related research in order to infer research trends for each factor. On the other hand, [10] provided the development and definition of writing style in the German, Dutch and French studies in digital humanities. Reference [2] defined the writing style feature categories in cybersecurity whereas [38] review the use of writing styles in author identification specifically in cybersecurity literature for multi-language documents. Another survey paper regarding authorship attribution by [39] focuses on the computational requirements and settings used in previous research for identifying authors rather than on linguistic or literary issues. In another paper, [40] combines both issues; computational requirement and linguistic element in their review paper, but focusing on the effect of the number of candidate authors and the size of training data settings in order to determine the identity of authors. The comparison summary of these review papers regarding writing styles and authorship identification is shown in Table 1. As we can see from the comparison of other survey papers, none of the studies to date have made a detailed comparison of the writing styles in a multidisciplinary field that has been studied previously individually; demographics, cybersecurity, and psychological factor such as personality, as what is presented in this paper.

Other than that, this paper attempts to close the gap between the writing styles in pre- and post-Internet periods as well as provide an in-depth comparison of the writing styles techniques in OSN texts across three multidisciplinary factors: demographics, personality & behavior, and cybersecurity. Also, this paper extends the work of [10], by providing writing style definitions, as well as adapting the writing style feature categories of [2], [41]–[44] to OSN text across the three stated multidisciplinary factors. The aim is to learn from past literature as we advance the writing style techniques to OSN text. The contributions of this paper are summarized as follows:

- Provides definitions of writing styles used across the stated factors.
- Presents common writing styles during pre- and post-Internet periods.
- Presents an overview of writing style feature categories in OSN text across the stated factors.
- Provides an in-depth comparison of writing style techniques in OSN text across the stated factors.
- Presents common writing style techniques and datasets in OSN text across the stated factors.
- Presents speech variation (whereby people adapt their language to their conversational partners) as a writing style in OSN text across demographics and cybersecurity.
- Proposes a novel machine learning prediction model based on tense morphology, to classify age and gender from English blogs, and the PAN 2013 dataset [45].

In the next section, we define and categorize the writing styles that we will refer to throughout the paper. Section III compares the writing styles used in demographics. Section IV compares the writing styles in personality & behavior, while Section V compares them in cybersecurity. Next, we discuss

[1]https://www.grammarly.com/blog/grammarly-editor-document-type/

TABLE 1. The discipline of writing styles research from previous review papers.

Review paper	Demographics	Personality	Linguistic	Authorship
[36]	✓			
[37]	✓	✓		
[10]			✓	
[2]				✓
[38]			✓	✓
[39]				✓
[40]			✓	✓
This paper	✓	✓		✓

the experiments based on the morphology model in Section VI. Finally, we conclude the survey and discuss future research directions in Section VII.

II. WRITING STYLES IN LITERATURE

Writing, according to Merriam-Webster [46], refers to ''a style or form of composition'', while style refers to ''a distinctive quality, form, or type of something''. Daelemans [47] gave a definition of writing style as stylistic language variance that links factors that are psychological (e.g., personality, mental health, native speaker) and demographics (e.g., age, gender, education level, region of language acquisition) to content and text genre. However, it was yet to be seen at that time whether the link is consistent over time, and whether style features are unconscious and uncontrollable. The researchers further stated that a hypothesis arising from this definition recognizes style as being like a fingerprint, unique to an individual [48]. The same researchers used machine learning to demonstrate the existence of variance in writing styles.

This paper extends the above definition to include other demographic factors such as ethnicity; variance between age and gender as well as between gender and biological sex. We also include the variance from authors adapting their language to conversation partners and in deception detection, where features of OSN text that comes from scammers, frauds and even hackers.

A. WRITING STYLES IN PRE-INTERNET PERIOD

We will now briefly discuss the history of writing styles. Researchers have for centuries used writing styles to differentiate authors and have been quite successful. However, as their methods were more intuitive, contradictions existed from one philologist to another, and no one could verify or refute their findings; even the most successful philologist did not know all the style markers used. In 1851, Augustus De Morgan, a mathematician, was the first to use scientific methods when he suggested that each author's prose had a specific mathematical fingerprint. He proposed the use of word length to clarify the authorship of a biblical book, i.e., the Letter to the Hebrews [49]. Mendenhall extended this suggestion in 1887 and counted the frequency of the words of a given length occurring in 1000-word samples of a few authors [50].

However, a more through and convincing study in styles was done in 1964, when Mosteller and Wallace used the frequency of function words to clarify the authorship of the 12 disputed Federalist Papers [29]. In 1994, Holmes, who also considered style variables to be stylistic fingerprints, classified and defined 15 writing style techniques used by researchers. This includes word length, syllables, sentence length, distribution of parts of speech (POS), function words, type-token ratio (TTR), vocabulary distribution and word frequencies [51].

Variations in speech was first identified in 1984 by Bell [22] where people adapt their language to their audience. A similar phenomenon was seen in a Twitter study performed by Nguyen *et al.* [52] in 2014. The adaptation to audience phenomenon occurred due to the author's intention to convey information to the targeted audience on first glance perception. This method is similar to how a fraudster or perpetrator would design posts in OSN and on the web to draw the attention of potential victims [53]. In our opinion, a speech style can be considered as a writing style as people sometimes think aloud while writing. This is further strengthened with numerous speech to text applications where the dictated text can be shared as an email, or even as a OSN post.

Three rather simple but interesting conclusions can be drawn from the history of writing styles, and are applied in this paper. First, the researchers agree that writing style is a unique fingerprint. Second, the same writing style techniques used during the pre-Internet period are still used today, even when the medium has changed to online. Finally, the language variation in speech noticed is currently being seen as a writing style in OSN text studies.

B. WRITING STYLES IN POST-INTERNET PERIOD

We will now discuss the stylometry writing style technique that is still relevant today. This technique refers to statistical

TABLE 2. Feature categories of writing style, adapted from [2], [41]–[44].

Category	Description	Examples of Writing Style Techniques
Lexical	Captures vocabulary-related style.	Average sentence length in characters, average sentence length in words, average word length, total number of words, total unique words, vocabulary richness.
Syntactic	Captures organization style of sentences.	Punctuation frequency, function word frequency, POS word frequency, frequency of words with all capital letters.
Structural	Captures organization style of a document.	Paragraph length, indentions, use of greeting and farewell statements, No. of words/sentences/characters per paragraph.
Content-specific	Includes frequency of keywords or other content-specific information on the topic of the document.	
Additional features	To fit the needs of researchers; may span across categories or stand alone.	Character and word n-grams, morphology, idiosyncratic, LIWC, 20 factors.

analysis of literary studies that consists of feature categories that are able to differentiate authors or author groups. While there has never been complete agreement on the feature categories in stylometry, most researchers have categorized them into four main categories: lexical, syntactic, structural and content-specific. Stylometry has also been known as writeprints, a similar technique that uses the four feature categories often used for authorship analysis in forensics and cybersecurity [21], [54]. Both [55] and [21] introduced idiosyncratic (focusing on misspellings and grammatical mistakes that are often ignored or autocorrected) as a new category. In contrast, [2] recognized this to be a writing style technique in the additional features category and introduced a new semantic feature category.

While [44] and [41] separate character-level features into a separate category, [41] also separates vocabulary richness features from the lexical category to be a separate category. In contrast, [56] introduced the morphological as a new feature category in order to categorize frequencies of POS; it is categorized as a syntactic feature category in [43].

The term stylometry, as we see above, is overused. We follow the feature categorization by Neal *et al.* [2], as it is the most recent work. Where needed, subtle changes are made to suit our purpose (see Table 2), which requires the writing style techniques to be differentiated, as our focus is on categorizing them between three factors: demographics, personality & behaviour, and cybersecurity.

The first subtle change is moving character-level and word-level features from the lexical to the additional features category, as they are used across the categories. The character-level and word-level features use the n-grams writing style technique. This approach somewhat follows that of Ashraf *et al.* [41] and Reddy *et al.* [44], who separate character-level features into a separate feature category. In fact, the superiority of character n-grams is often attested in stylometry [57]. The reason is because n-grams is a technique borrowed from the field of information retrieval. In this technique, text is divided into series of consecutive overlapping groups of n characters. For example, 1-grams considers unigrams ($n = 1$), 2-grams considers bigrams ($n = 2$), and so forth. A similar dilemma was faced when categorizing the natural language processing (NLP) and machine learning techniques that were borrowed from the artificial intelligence field and used across categories that are currently placed under the additional features category. Therefore, our lexical category definition also follows Ashraf *et al.* [41] but with one major difference: we add the vocabulary richness feature, which is a separate category in their study, into the lexical category, following Neal *et al.* [2].

The second subtle change is to place POS words in the syntactic feature category. This change is made because Neal *et al.* [2] put function words in the syntactic category, and in doing so, we follow a somewhat similar direction to Neal *et al.* and Hollingsworth [43], who put frequency of POS words into the syntactic feature category.

The third subtle change is the preference for using the term content-specific instead of domain-specific as used in [2], as most literature [21], [54] uses the former. Last but not least, three writing style techniques (morphology, linguistic inquiry and word count [LIWC], and 20 factors) are also not part of stylometry, and we place them under the additional features category. Stankev *et al.* [42] separate the morphology feature category (study of the structure and formation of words) from stylometry. On the other hand, LIWC created by Pennebaker *et al.* [58] in 2001, uses an internal dictionary of more than 2,300 of the most common words and word stems and is categorized into linguistic and psychological categories.

Together with Pennebaker, Argamon and colleagues [30] created 20 factors from the 1000 most common words, comprising 323 function words and 677 content words. Moreover, following a similar concept to LIWC, function words were grouped into several categories, i.e., 20 factors, according to

their POS. Other additional features that are less disputed, and is specific to a particular factor is explained in their respective sections.

III. DEMOGRAPHICS

In the early 2000s, Srihari *et al.* [6] were among the pioneers to use rigorous scientific studies to show author identification for handwritten text. Their idea was that handwriting style varies from person to person, while, for a particular person, the style is somewhat consistent. At that time, such a hypothesis had not been conclusively tested. Handwriting style was studied in respect to demographics such as gender, age and ethnicity.

In this section, the same demographics with the sex factor, are explored in a modern context, i.e., OSN. We focus on Twitter, blogs, user reviews, and chat messages.

A. GENDER, SEX, AGE AND ETHNICITY

Within social science, the study of demographics and language is not new. Sex, together with social status, age and ethnicity, as explained by Cheshire [59], is one of the most widely used demographic factors. These factors are now recognized to be more complex than their labels.

Wodak and Benke [14] argue that, in their view, many empirical studies have neglected the context of language behavior and have often analyzed gender by merely looking at the speakers' biological sex. Instead, they stress that concepts such as gender and age are shaped differently depending on individuals' experiences and personality as well as the society and culture that they belong to.

To complicate things further, gender is often intertwined with age, where studying one often implies that the other is also studied [12]. In this paper, we do not differentiate between gender and biological sex, as we do not know the authors' implicit views when writing their papers.

1) TWITTER

Twitter is a OSN site where people interact using short messages known as tweets.

Word n-grams is the main writing style technique employed in most studies [34], [60]–[64]. Rao *et al.* [60] used word n-grams of 1-2 on tweet text to infer gender, age and ethnicity. The authors preserved the emoticons and other punctuation sequences that were traditionally deleted. The researchers highlighted the different kinds of emoticons used, where female users tend to use emoticons such as hearts (< 3), while male users prefer grins (:D) and winks (;)). However, emoticons are almost identical in both age groups. This observation may, however, be contributed to the gender pool, which targeted certain users. In contrast, North Indians tend to use more emoticons, alphabetical lengthening and excitement markers than South Indians. Though accuracy up to 73% (using support vector machine [SVM]) was seen for age, gender and ethnicity prediction achieved their best accuracy when augmented with 15 sociolingustic-based features.

In a separate study, [63] also preserved emoticons and punctuation to infer gender from the top 10,000 words using statistical hypothesis. There was one contradictory finding from [60]; namely, emoticons such as grins and winks were found to be female markers. In addition, gender was associated with function words such as the pronouns (*u*, *ur*, and *yr*) and prepositions and articles that included alternative spellings. One male marker is that the number '2' is often used as a homophone for *to*, while the abbreviated form of *with* appears in female markers (e.g., *w/a*, *w/the*, *w/my*). Moreover, despite swear and taboo words being male markers, the anti-swear *darn* is a female marker.

[65] also used the word n-grams feature but did so to extract the top topic-sensitive terms, in addition to using 383 function words, 10 punctuation marks and 13 other features (e.g., word frequency, character frequency, vocabulary richness). The accuracy reached approximately 83% when augmented with the multimedia feature, which denotes the number of tweets containing an image, video and/or music. One distinct finding was that gender detection accuracy began at 53.2% with the top 10,000 terms and reached its peak with 2.5 million terms. The accuracy, however, gradually dropped when more terms were used.

In addition to word unigrams, [62] used LIWC to infer age. They used 48 proposition words from the Dutch version of LIWC, in addition to adapting it to use alphabet lengthening, slang and English pronouns, as Dutch people often tweet in English as well. Intensifiers, tweet length, average tweet length and user tweeting behavior (including tweets containing a hyperlink or hashtag) were also used. The results revealed little difference after the age of approximately 30, which was attributed to insufficient data in the extremes of the age groups. The researchers then extended this work in their next study to find the reason with respect to age and gender.

The main reason, according to Nguyen *et al.* [52], is that more than 10% of Twitter users do not use language that is normally associated with their biological sex. Their gender association study suggests that Twitter users vary their language depending on the context and their conversation partners. A significant trend is observed in female users (Pearson's $r = 0.270$, $p < 0.001$), which seems to suggest that they emphasize aspects other than their gender in tweets. Echoing findings of their previous work, on average, older Twitter users' age is underpredicted, where the prediction errors start at the end of the 20s. The gap between actual and predicted age increases with age.

[66] is another study that, like [62], uses LIWC, but to infer gender. The LIWC dictionary produced 64 content-based features combined with six user tweeting behavior features. These content-based features, according to the authors, have proven to be more effective than other simple features such as k-top words. On its own, the LIWC feature set yielded accuracy of 80.43% for the last 1000 tweets, which increased to 82.26% when augmented with the tweeting behavior.

TABLE 3. Writing style features to identify author demographics on Twitter.

Literature	Data	Writing style features				
		Lexical	Syntactic	Structural	Content-specific	Additional features
[60] (2010)	English Twitter posts. Crawl seed for gender is sororities, fraternities, male, female hygiene products, which produced approximately 500 users per class. Crawl seed for age is baby boomers, young moms, junior, freshman in user description, which produced approximately 1,000 users per class.	✓	✓		✓	Word n-grams, 15 sociolinguistic-based features.
[61] (2011)	Multilingual (> 13 countries) Twitter posts. Crawl seed is users who have a URL to their blogger profile, which resulted in 184,000 users producing more than 4 million tweets.	✓	✓	✓		Character and word n-grams.
[34] (2011)	Flemish Dutch post from Netlog. Crawl seed is info on age, gender and location that is available for authors. This resulted in 1,537,283 posts and 18,713,627 tokens (words, emoticons and punctuation marks).	✓	✓			Morphology, character and token n-grams.
[67] (2012)	English Twitter posts. Crawl seed is to remove users whose gender is unclear or who could not write English. This resulted in 3,000 users.	✓				Character n-grams.
[69] (2013)	Multilingual Twitter posts. 943 French users, 3237 Indonesian users, 3609 Turkish users, 829 Japanese users.	✓	✓	✓	✓	Morphology, word n-grams, k-top words, k-top bigrams and trigrams, k-top hashtags.
[52], [62] (2013, 2014)	Dutch Twitter posts. Crawl seed is word 'het' contained in tweets and less than 5000 followers (to limit inclusion of celebrities and organizations), resulting in 2439 users.	✓	✓	✓	✓	Dutch version of LIWC, word unigrams.
[63] (2014)	American English Twitter posts. Crawl seed is authors who use at least 50 of the 1,000 most common words in the US sample overall and have between 4 and 100 mutual-mention friends, resulting in 14,464 users and 9,212,118 tweets.	✓	✓			Word n-grams, top 10,000 words.
[65] (2014)	Twitter posts. Crawl seed is 1000 active users (486 male and 514 female users) who have indicated their gender on their profile, which resulted in more than 3.6 million tweets.	✓	✓			Word n-grams.
[66] (2014)	Twitter posts. Crawl seed is first name and last name on Twitter, where the former was one of the top 100 most common names on record with US Social Security Department for baby boys/girls born in the year 2011, which resulted in 192 users.	✓	✓		✓	LIWC, word n-grams.

Character n-grams is another writing style technique seen in the following studies [34], [61], [67], [68]. [67] used every possible 1 through 5-gram for the 95 most-used ASCII characters to represent the tweet to infer gender. A tweet length of 25 words, for example, produces 15,000 features, where accuracy of more than 90% was seen.

In contrast to the previous study, [68] introduced 12 readability features to enhance accuracy. Together with three character-level features, five word-level features, one sentence richness feature and one vocabulary richness feature, with convolutional neural networks as the classifier, an accuracy of 97.7% and 90.1% was achieved for gender and age, respectively.

[34], on the other hand, not only combines word n-grams of 1-3 and character n-grams of 2-4 but also focuses on text with an average length of 12.2 tokens (i.e., words, emoticons and punctuation marks) per post, which, to the best of our knowledge, is the only Netlog study that focuses on very short text for age and gender classification. The researchers studied Flemish language usage in post and chat conversations. One interesting conclusion from [34] is that regional varieties and dialects make n-grams a better approach than POS tagging, e.g., *dak* (roof) is tagged as a noun, while its abbreviated/dialect form *dak ik* (that I) should be tagged as a combination of a relative and personal pronoun. The best accuracy was achieved with 50,000 features where word n-grams outperformed character n-grams. A confusion matrix indicated that varieties in chat language usage were attributable more to age than to gender, as there was greater confusion between gender classes of the same age groups compared to age classes of the same gender groups.

In a large multilingual study, [61] mainly relied on word n-grams for segmented languages and on character n-grams for unsegmented languages. Echoing the findings of [60], the researchers concluded that tweet text conveys more about the gender of the author than the profile's description, where the text only yielded an accuracy of 75.5% (using balanced Winnow2). One distinct stylistic finding is that the presence of *http* is a strong male indicator. Cursory examination seems to suggest that female users are more likely to use "bare links" (cf. nltk.org vs. http://nltk.org). In a separate study, [64] used a combination of word and character n-grams to achieve the best accuracy for the English, Spanish and Arabic datasets.

Morphology is a commonly used writing style technique for non-English texts. In contrast to the previous study, [69] uses unique features of other languages to improve inference accuracy. This study shows that prediction is easier in some languages not due to their conventions in word usage, but rather to their syntactic structure. Using k-top words, k-top bigrams and trigrams, k-top hashtags with three other features, and SVM as classifier, the accuracy of French (76%), Indonesian (83%), and Turkish (87%) was comparable to that of English. Japanese, in contrast, performed at 63%, even with white space tokenization that was inserted to break the words. One conclusion drawn by the researchers is that word n-grams carry little information in languages with complex orthography. French, on the other hand, has information about gender in its nouns. Thus, using morphology, an accuracy of 90% was achieved in French. Table 3 shows an overview of using writing styles to identify author demographics on Twitter.

2) BLOGS

A blog, according to Merriam-Webster [46], refers to "a website that contains online personal reflections, comments,

and often hyperlinks, videos, and photographs provided by the writer''.

The 20 factors [30] is main writing style technique seen in the following studies [30], [70]–[74]. In [30], Argamon *et al.* pioneered this technique to predict gender and age. Briefly, this study showed that POS often used by younger (cf. older) bloggers are also used often by female (cf. male) bloggers. The technique achieved gender accuracy of 80.5% (using multiclass balanced real-valued Winnow) and an age accuracy of 77.4% (using Bayesian multinomial logistic regression). Careful examination of the age confusion matrix revealed that distinguishing 20s from 10s and 30+ is difficult.

In a separate study to infer gender and age, the same researchers [70] explored some categories of LIWC, POS, function words, blog words and hyperlinks, totaling 502 stylistic features and 1,000 unigram content features. Their study shows that stylistic differences remain more telling than content differences, and regardless of gender, writing style becomes more male with age. Much like the previous study, 30s were misclassified as 20s. On its own, average sentence length in the same dataset gave little improvement in accuracy in the next study [71]. However, when average sentence length and 52 slang words were augmented with 35 content words (mostly from the previous study), a clear distinction between slang usage was observed for both genders.

In contrast to the three previous studies, [74] subdivided some of the 20 factors, added new factors, and introduced a frequency-based score to determine gender; their approach improved gender detection for the best result obtained by Argamon *et al.* [30] to 82%. Much like that study, [72] added three new factors. Furthermore, [72] introduced variable-length sequence mined from POS in order to capture stylistic characteristics of both genders. Writing styles were analyzed using words including blog words, 10 gender preferential features, 23 factors and POS sequence patterns as features. The best accuracy was achieved using the SVM (regression) classifier. [73] extended this study by focusing on words and punctuation that people use, average word/sentence length, and POS analysis as well as the 20 factors; the best accuracy of 72.10% was achieved with SVM (linear kernel) as the classifier.

Mimicking the experiments of [70] on a different dataset, [32] inferred gender using seven lexical features and interests mined from profile pages as content words. There was not a noticeable increase in slang and punctuation for younger bloggers, while the number of links and images varied across all ages. The accuracy, however, was improved to 67% (using the logistic regression classifier) when online behavior was added as a feature. Closer examination of the age confusion matrix showed that many aged 38-42 were misclassified as being aged 28-32.

Like Twitter studies, morphology is a commonly used writing style techniques for non-English blogs, where text segmentation is an issue [75]–[77]. Reference [75] is a multilingual study that uses four different types of mostly content-independent features, comprising 15 character-level features, 14 word-level features, two sentence-level features, and from 22 for French to 65 for English syntactic features. Monolingual gender identification yielded accuracies between 77% and 88% (using bagging as the classifier). However, the number of features is reduced to 27 language-independent features when the six individual datasets are merged into a single dataset, where an accuracy of 74.67% was achieved. In comparison to other languages, with the exception of quotation marks, deviations from reference features for German authors are small. This result can probably be attributed to ethnicity influences.

In addition to morphology, these two studies demonstrated that word n-grams contribute more to gender identification than character n-grams for non-English blogs, as they convey specific syntactical and morphological patterns. Reference [76] used 298 features that can be classified into character-level features, word-level features and other features to predict age, gender and ethnicity for Vietnamese, where the best accuracy was achieved using the IB1 classifier. On the other hand, [77] echoed this word n-grams vs. character n-grams finding when analyzing Greek blogs using five vocabulary richness, three word length, one letter frequency and five character-level and word-level features consisting of character and word grams. The sequential minimal optimization (SMO) classifier yielded an accuracy of over 82%. Table 4 shows an overview of using writing styles to identify author demographics on blogs.

3) USER REVIEWS

To review, according to Merriam-Webster [46], means ''to give a critical evaluation of''. Here, we will examine user opinions on products and services shared with other consumers.

Like Twitter and blog studies, morphology is a common writing style technique for non-English user reviews. To our best knowledge, [78] is the first study that analyzes morphosyntactic variation on a large scale for gender, age and ethnicity. One distinct finding was on intensifiers, where female users are said to be more restrained in their language and use hedges to soften what they say. However, this finding contradicts the previous Twitter findings of [69] and the next study by Hemphill and Otterbacher [79].

[79] used a range of features including rates of pronoun use and hedging (a list of 55 hedging phrases), complexity of words and sentence structure, and vocabulary richness. Two distinct findings were observed. First, though it was expected for female users to adjust their writing style patterns by reducing the use of some female markers, it was not expected for male users to adopt some female markers. Second, using roughly the same features and dataset with an additional 50 markers (50 top words consisting of 5 content words and 45 function words), [80] showed that female users exhibited a richer vocabulary, contradicting the findings of the aforementioned study indicating that female users exhibit less vocabulary richness.

TABLE 4. Writing style features to identify author demographics on blogs.

Literature	Data	Writing style features				
		Lexical	Syntactic	Structural	Content-specific	Additional features
[70] (2005)	English blogger.com blogs. Crawl seed is contained at least 200 common English words and author-provided gender. This resulted in 37,478 blogs (equal number of male and female blogs), 1,405,209 posts and over 295 million words.	✓	✓	✓	✓	Unigrams, LIWC.
[30], [71], [74] (2007, 2009, 2013)	English blogger.com blogs. Crawl seed is contained at least 500 total words including at least 200 occurrences of common English words and author-provided gender and age. This resulted in 19,320 blogs, composed of 681,288 posts, over 140 million words, and, on average, approximately 35 posts and 7,300 words per blog.	✓	✓	✓	✓	Factors.
[76] (2009)	Crawl seed is authors are native Vietnamese speakers and the main language in blogs is Vietnamese; blogs written in the last four years; each author has more than 10 posts; number of words is more than 150 words; blog posts must be authored by blog author. This resulted in 73 blogs, 3,524 posts, 74,196 words with an average of 1,016 words per blog.	✓	✓		✓	Morphology, character and word n-grams.
[72], [73] (2010)	Blogs from blog hosting sites and blog search engines, e.g., blogger.com and technorati.com. This resulted in 3,100 blogs, where 1588 blogs (1512) were written by male (female) users. Average blog length for male (female) users is 250 (330) words.	✓	✓	✓	✓	Factors.
[32] (2011)	English blogs from livejournal.com. Crawl seed is bloggers who have indicated age and listed the US as their country, where each blog entry is written by a unique person, contains a user profile and up to 25 recent posts written between 2000-2010, with the most recent post being written in 2009-2010. This resulted in 11,521 bloggers, comprising 256,000 posts and 50 million words.	✓	✓	✓	✓	
[77] (2012)	Greek blogs from Greek blogosphere. Crawl seed is 50 most recent posts of 10 male and 10 female blogs that share a common topic (e.g., personal affairs). This resulted in 1,000 posts and 406,460 words.	✓			✓	Morphology, character and word n-grams.
[75] (2015)	Multilingual blogs (Catalan, English, French, German, Italian and Spanish). Crawl seed is authors were known and the genders could be inferred. This resulted in 29,117 posts in a merged dataset: English (7148 posts, 51 authors); Spanish (5794 posts, 101 authors); German (3564 posts, 127 authors); French (4310 posts, 18 authors); Catalan (4078 posts, 33 authors); Italian (4265 posts, 43 authors).	✓	✓		✓	Morphology, character and word n-grams.

In another study, [81] used morphological gender and ethnicity markers to infer age, gender and ethnicity. As an example, *træls*, a swear word found in the Jutland dialect (Denmark) that has no exact translation in the standard language, was found to be used in 84% of people living in a very small portion of the Copenhagen region. It was noticed that female users (cf. younger people) tend to use this term more than male users (cf. older people). Moreover, their findings also showed that though the English language has influenced the official Danish language, e.g., single words in English (''today'') are multiwords in Danish (*i dag*), the older generation still wrote the prescribed spelling, while the younger generation did not. Table 5 shows an overview of using writing styles to identify author demographics on user reviews.

4) CHAT MESSAGES

Chat, according to Merriam-Webster [46], refers to ''online discussion in a chat room''. Here, we do not consider chat plugins, for example, Facebook's instant messaging services.

To our best knowledge, [82] and [83] are the only two studies that analyze the writing style of very short messages (6.2 words) for gender prediction. The researchers used writing style techniques to consider lexical and syntactic features consisting of average message length, average word length, frequency of stopwords, list of 78 stopwords, frequency of smileys, a list of 79 smileys, frequency of punctuation marks, a list of 37 punctuation marks, number of distinct words (i.e., vocabulary richness), and frequency of each character. The best accuracy was achieved using a feature set without stopwords and vocabulary richness measures. A surprising finding was that stopwords and punctuation marks vary for male users, who either use them heavily or very lightly.

In a separate study, the same researchers [83] extended their work to the day-night interval. An unforeseen finding is that during the day, chatters write shorter messages containing auxiliary elements (smileys and punctuation marks) with hedging. In contrast, during the night, they write longer messages containing many function words and punctuation marks.

Like the previous OSN studies in Sections III-A1- III-A3, morphology is also used in non-English chat messages [33]. This study tracked morphological gender markers such as nouns, adjectives, pronouns or determiners in English and Spanish chat rooms, as their participants are mainly from the US and Spain. Table 6 shows an overview of using writing styles to identify author demographics in chat messages.

B. SUMMARY

In this section, we reviewed previous studies on the use of writing styles to identify demographics, in particular, gender, sex, age and ethnicity for OSN: Twitter; blogs; user reviews; and chat messages. We summarize the studies as listed in Tables 3, 4, 5 and 6. In demographics studies, researchers do not always explicitly categorize the writing style techniques into the lexical, syntactic, structural, content-specific and additional features writing style feature categories. Therefore, it is our attempt to categorize them based on descriptions found in their papers.

Most demographics studies support the following definitions of style: as an expression of individuality, subjectivity or emotions of the author; as a deviation; and as a language choice. The latter is a speech variation hypothesis of [22] seen in [52], who further suggests that Twitter users vary their language depending on the context and conversation partners.

TABLE 5. Writing style features to identify author demographics on user reviews.

Literature	Data	Writing style features				
		Lexical	Syntactic	Structural	Content-specific	Additional features
[79], [80] (2010, 2012)	User reviews from 250 top movies from Internet Movie Database (IMDB.com). Comprises 31,300 reviews, 21,012 unique reviewers (10,394 male, 10,618 female), 8,022,196 words with vocabulary size of 35,783.	✓	✓		✓	Morphology, statistical analysis.
[78], [81] (2015)	Multilingual (24 countries) user reviews from 13 different languages.	✓	✓		✓	Morphology, semantic variation, statistical analysis, NLP

TABLE 6. Writing style features to identify author demographics on chat messages.

Literature	Data	Writing style features				
		Lexical	Syntactic	Structural	Content-specific	Additional features
[82], [83] (2006, 2008)	Turkish chat messages. Consisting of 1,500 users from Heaven BBS, 250,000 messages with 50,000 distinct words and 6.2 words per message on average.	✓	✓			
[33] (2008)	5 English-speaking chat rooms hosted by America Online and 4 Spanish-speaking chat rooms from mIRC channel#Hispanic.		✓		✓	Morphology.

In addition, [52] is the only example in the computer science literature that we are aware of that follows the social science perspective on gender by making an explicit distinction between users' biological sex and gender.

In a separate study, [79] showed that female users adjust their writing style patterns to become less feminine and that male users adopt some female markers. Table 7 summarizes some of the common female and male markers seen in this review.

The heart of research lies in the dataset; creating an unbiased dataset is a nontrivial task. Reference [60] remove the biases by excluding Twitter users who have high follower counts, as they tend to be celebrities and organizations. Reference [62], on the other hand, remove organizations from their Dutch Twitter corpus by sampling users with 10-200 tweets who have less than 5000 followers. In contrast, [63] remove spammers, dormant users and one-time conversation users from their American English Twitter dataset by creating a mutual network. These approaches will, of course, create various sizes of datasets; the difficulty of finding the correct size was pointed out by Lopez-Escobedo *et al.* [84].

In terms of using lexicons, there seems to be a preference to use LIWC [58] in Twitter studies, and 20 factors [30] in blog studies.

IV. PERSONALITY AND BEHAVIOR

With Web 2.0, people have more freedom to record and share their thoughts and activities in virtual communities such as websites and on social media. Therefore, it is reasonable to expect that individuals' online writings will also contain their personality-related residue. Previous research has shown that people inadvertently leave personality-related ''behavioural residue'' in their physical and virtual environments [95]. Due to this, research on language use has shown a connection between personality and writing style [4], [85].

Pennebaker *et al.* [4] investigated the degree to which language is reliable across time, and factor structure and established good constructs in comparison with self-report personality traits. Three-phase experiments were taken using handwritten materials from inpatients of addiction treatment and students and researchers from a psychology-related education background, similar to the work by [96]. The results show that language offers a perspective to detect individual differences of styles and personality. The study proved the work by [96], which found three distinct writing styles: one related to extraversion, another related to introversion, and a final one related to reflectivity.

While the first two studies focus on handwritten text, after 2007, the year of global communication, researchers started to explore OSN text [85], [87], [92]. OSN allows anyone to share, express and show their views on certain topics on the Internet. The researchers in [85] and [86] predicted Twitter authors' personality based not only on author profiles but also on their tweet texts. Three features in the LIWC tool—psychological processes (e.g., emotional, cognitive, sensory, and social processes), relativity (e.g., words about time, past, future), and personal concerns (e.g., occupation, health)—and word count, words per sentence, and swear word counts were used in their experiment to determine the text features related to personality. The features were then run through the MRC Psycholinguistic Database, a list of over 150,000 words with linguistic and psycholinguistic features, and word sentiment analysis to extract the score of word use by Twitter authors that relates to personality traits. After testing the technique on Twitter, [85] expanded the experiment with another OSN, Facebook, and they managed to detect the personality traits of the social media users [86]. Furthermore, LIWC has also been used by [15] to derive the personality of ten chatroom authors to act as a baseline for authorship identification and could successfully identify the personality traits of the chat authors.

LIWC was also used in the study by [87] with Twitter posts, focusing only on text messages. Emoticons in the posts were changed to the words ''PositiveEM'' and ''NegativeEM'' to allow LIWC to detect the emoticons. Forty-nine LIWC characteristics were used in this study. Using just the LIWC technique, the study was able to find the distinct characteristics associated with the author's personality traits.

TABLE 7. Summary of the gender and age markers for female and male users.

Markers	Male	Female	Older	Younger
Articles	[30], [69], [70]			
Prepositions	[30], [69], [70], [80]		[70]	
Nouns	[78]			
Numerals	[63], [78]			
Pronouns		[30], [60], [63], [69], [70], [78]–[80]		[70], [79], [80]
Verbs		[30], [78], [80]		
Blog words	[69]	[60], [63], [70]		[60], [62], [70], [71]
Assent / Negation	[63]	[63], [70]		[70]
Hesitation / Disfluency		[60], [63]		
Shorter sentences / posts		[80], [82], [83]	[31]	[32], [71]
Long messages	[82], [83]			
Short words	[82], [83]			
Long words		[82], [83]		
Emotion terms / emoticons		[33], [60], [63], [67], [81], [82]	[60], [63]	[32], [60], [81]
Capitalization				[31], [32], [62]
Hedges		[69], [79]		
Alphabetical lengthening		[60], [62]		[60], [62]
Excitement markers		[60], [63]		

The agreeableness personality trait has a negative correlation with negative words, and the authors with this trait use fewer exclusive and sexual words. The openness to experience personality trait has a negative correlation with second-person words, assent words and positive emotion words. The extraversion personality trait has a negative correlation with function words and a positive correlation with assent words. The study also shows that the agreeableness and neuroticism personality traits are easiest to detect and perceive by readers.

In addition to LIWC, which is a closed-vocabulary approach, an open-vocabulary approach using language-based assessments (LBAs) has been used to predict the personality of social media users. Reference [92] used LBAs on Facebook users. LBAs are based on normalized relative frequencies of words and phrases, binary representations of words and phrases, and topic usage. The external criteria used to correlate with LBAs using regression modeling are the convergences of personality self-reports at the domain level and facet level, the discriminant validity between predictions of distinct traits, the agreement with informant reports of personality, the patterns of correlations with external criteria (e.g., number of friends and political attitudes) and, lastly, the test-retest reliability over 6-month intervals. The results in [92] show that the LBA approach can constitute valid personality measures.

We also found that closed-vocabulary and open-vocabulary approaches can be combined together as proposed by [88]. Their study shows a positive correlation in predicting the personality of social media users. The study used 19 million Facebook status updates provided by 136,000 volunteer participants. All five personality traits had a correlation value of 0.3 to 0.4 when word phrases containing n-grams of size 1-3, a latent Dirichlet allocation (LDA) combination of topics within the documents and LIWC were combined together as features to predict the personality of the Facebook users, other than their age and gender.

Another study has also used writing styles to predict personality & behavior of authors in a language other than English. Reference [93] shows that bag-of-words (BOW), text segmentation and a weighted scheme can be used to identify the extroverted or introverted personality of 222 Facebook users who write their posts in Chinese. Since Chinese segmentation is different from English, the authors used the Jieba algorithm that specifies Chinese text segmentation and term frequency - inverse document frequency (TF-IDF) to eliminate noise in the texts.

Another method to predict personality based on writing styles is stylometric. The definition of stylometric has been discussed in Section II. However, rather than using the common approach of extracting stylometric features from a set of training texts, [90] applied the stylometric technique in chat conversations collected using a keylogger. Chat conversations are different from normal text posts on the web, as the texts in chats appear interchangeably between authors. Therefore, the researchers expanded the stylometric features to include turn-taking. Additionally, due to the privacy of the respondents and the objective of their study focusing more on basic behavior than on complex personality traits, [90] opted not to use content-specific, structural and idiosyncratic features. Using statistical analysis, they identified that authors who use more short words and in a particular rhythm are predicted to have positive affectivity behavior. This study shows that

writing styles in a chat environment can predict authors' psychological factors when interacting with another person in cyberspace.

The study by [91] offers further proof of this theory by examining when sexual predators in online chat rooms can be detected using the LIWC features and their online chat platform behavior. The LIWC features seen in the profile of sex predators are first-person pronouns, negative emotion words, affect words, sadness words, time, motion and location words. For the behavior, the style of opening conversation, time logging in and the use of online chat-room features are some of the behavior observed to differentiate sex predators from normal users. Combining the writing styles and chat platform behavior can help to perform automatic prediction of dangerous authors such as sex predators on online chat rooms and alert other users about them.

Rather than just predicting the personality of authors, writing styles could also detect behavior such as depression. In the study by [89], 22 linguistic styles of LIWC were used to determine positive and negative words. They further use the Affective Norms for English Words (ANEW) lexicon to determine the dominance and activation of each negative word to create the weight of the words. Particular sentences that show the usage of antidepressant medicine are also included as depression features, together with 1,000 depression lexicons that [89] compiled from Mental Health Yahoo Answer! discussions and compared with Wikipedia to identify the words with high TF-IDF. Another feature that is included in their prediction model is the ego-network—the number of followers-followees and social communication. Using Twitter posts, the researchers were able to obtain 70% depression classification accuracy for their prediction model. The researchers further studied a specific type of depression, postpartum depression. Based on the same features used in the depression prediction model but tested on Twitter posts from users who announced their child's birth, [89] were able to achieve 71% accuracy for the window of 3 months and 81.6% for posts in the window of 2-3 weeks after delivery. The study shows that postpartum depression has a better chance of being detected at an early stage within the first 2 to 3 weeks after giving birth based on the writing style of the mothers in their OSN posts.

The study by [94] has also proven that writing style analysis is useful to automate personality classification, specifically on neuroticism and extraversion personality traits. By using Facebook status updates, Twitter posts, YouTube speech transcripts and normal essays, the authors are able to extract semantic and sentiment words. While semantic and sentiment words are not a common writing style technique, the researchers used semantic words extracted from the lexical, POS, suffix and word n-gram areas on WordNet, in addition to sentiment words extracted from SentiWordNet focusing on positive and negative emotions.

Word sense disambiguation (WSD) is also extracted to sense the meaning of a word used in a sentence, especially when the word has multiple meanings. The study shows that the combination of WSD, semantic and sentiment words can improve personality classification for both the neuroticism and extraversion personality traits. Table 8 shows an overview of using writing styles to identify author personality & behavior.

A. SUMMARY

In this section, previous studies regarding the analysis of writing styles to identify authors' personality & behaviors have been reviewed. Most personality & behavior studies support the following definitions of style: as an expression of individuality, subjectivity or emotions of the author; and as a deviation. We summarize the studies as listed in Table 8. A high percentage of the literature uses the closed-vocabulary approach LIWC technique as the feature to assess author writing style. The approach is a favorite probably due to the massive numbers of positive and negative emotion word extracts that it is programmed to analyze. Regarding personality & behavior, positive and negative emotion words can narrow down the type of personality & behavior of the author. LIWC also appears to combine well with a content-specific approach. The type of topics or content help to describe what makes an author have a positive or negative reaction. The reaction shows the type of personality that the author has.

For the type of dataset used in the literature, most texts came from OSN posts rather personal websites or comments. This is probably due to the nature of social media privacy settings. Once authors publish their posts, they are considered public and anyone can capture them. Furthermore, the connectivity (network relationship) and tag features on social media platform allow researchers to gather the posts and easily filter them accordingly. The massive number of users on these OSN is also another factor explaining why most sources of datasets in personality analysis use the writing style technique.

Hence, we see that the number of datasets is in the thousands to millions of message posts, compared to the hundreds for chat texts. Readers can refer to the comparison table for other details.

V. CYBERSECURITY

Cybersecurity is the study of securing the cyberspace. It involves defending online systems against vulnerability exploitations that can lead to abuse, intrusion, and other dangers. Cybersecurity not only covers traditional information security but also includes the protection of information resources such as physical, digital, and human assets [97]. Many cybersecurity attacks include the deception of humans using texts to scam, defraud, and phish, among other schemes. In this section, we will show how writing styles analysis is used to detect and even mitigate the risk of cybersecurity attacks that use text as their main attack mechanism.

A. FAKE ONLINE INFORMATION

Since the implementation of Web 2.0, anyone can post about almost anything on the Internet, be it factual or fictional.

TABLE 8. Writing style features to identify author personality of OSN texts.

Literature	Data	Writing style features				
		Lexical	Syntactic	Structural	Content-specific	Additional features
[85], [86] (2011)	2,000 Twitter posts, 167 Facebook authors with min 10 words posts.				✓	LIWC.
[87] (2012)	142 Twitter authors (28,978 posts).					LIWC.
[15] (2012)	341 chat posts.					LIWC.
[88] (2013)	136,000 Facebook authors (19 million Facebook status updates).					LIWC, LDA.
[89] (2013)	376 Twitter accounts (have post-birth announcements).	✓			✓	LIWC, ego-network.
[90] (2014)	Chat texts from 50 participants.	✓	✓	✓	✓	
[91] (2014)	Chat containing sexual predators from the PAN 2012 dataset (only 150 and fewer messages).					LIWC, interface features.
[92] (2015)	66,732 Facebook authors; posts (715 million posts) for training and 4,824 Facebook posts for testing.				✓	LBA.
[93] (2015)	222 Facebook authors posting $>$ 10 posts in Chinese only.					BOW, Chinese text segmentation, TF-IDF.
[94] (2017)	Twitter, 9,917 Facebook status updates, 404 Youtube Vloggers - text transcripts (28 hours), 2,479 open essays from psychology students.	✓			✓	

The problem arises when the fictional information online cannot be distinguished by readers as fact or fiction. Distinguishing between real and fake information becomes harder specifically when the information is short, such as Twitter posts [98]. It is even harder to detect when more than one person shares the same information or message online, or the number of likes on the message increases, as is often the case in OSN. A viral message (shared by many) and endorsed by others (number of likes and shares) could lead a reader to believe that the message is credible even without knowing the truth of the message [99], [100]. Fake online information attacks that will be covered in this subsection are fraud scams, hoaxes and rumors, as well as misinformation. The reason is due to the use of text in these attacks.

Online scamming is a type of fraud whereby a computer is used to trick and deceive readers into believing the information that they shared online is true. The fraudsters use promised monetary gains to attract their victims, making them believe they have won a lottery, have a chance to receive high-value investments, and more [101]. The main purpose of scamming is money. To achieve this purpose, scammers can use a direct con of money scheming, fake commercials advertising that something is free when in fact there is a hidden fee, or methods to do further cyber attacks, such as

phishing and spreading viruses in order to gain confidential information such as online banking authentication information [98].

The first two techniques very much predate modern scam schemes that now use computer technology to assist scammers in reaching out further and across borders [102]. However, the last scamming technique is more modernized. Fraudsters use computer-oriented attacks to trick people into downloading malicious software or malware such as Trojan horses, keyloggers, viruses and worms [102]. These malwares collect private and confidential information of unsuspecting victims and send it over the Internet to their owner, the fraudster. Once the fraudster has this information, he will use it to log into the victim's bank account, or charge payment to the victim's credit card, or sell the information on the dark web for a substantial amount of money. These are just some of the acts fraudsters can do with all the information they gain about their victims with this computer-oriented scam scheme, all without their victims knowing about it.

Hoaxes or rumors are also focused on deceiving people, but the reason behind them is different. Rumors and deliberate disinformation are disseminated to cause panic and influence public opinion. On OSN, rumors propagate much faster, especially if they are news-related. OSN readers search for news information on the platform [103]; the more interesting (in terms of hope, fear, or hatred) the information, the higher the views and shares of the news post become [104]. Rumor virality and popularity thus become one of the drivers by irresponsible people to continue to create and post hoaxes or rumors online, especially on social media [105]. The rumors are designed to adapt to their target's preferences. However, some rumors shared online are due to misinformation; because the information seems true, others share it because they find it useful and want other people to know as well, without realizing that the information they shared is not verified [106], [107].

B. DETECTION METHOD

Detecting or identifying fraudulent scams and rumors, whether due to misinformation or disinformation, is not easy, especially when the authors use an authornym, a fake name [19]. Other than celebrities and book authors, online users, especially on OSN, often use authornyms. In addition to using fake names, certain social media users are also seen using fake pictures called avatars to represent themselves in the online community. Facebook, Twitter and other OSN have received many reports regarding identity theft on these social media platforms. This online identity theft is known as an identity cloning attack (ICA) [107], [108]. ICAs can either be two accounts with the same authornym on the same or a different OSN, where one is a fake and the other the true identity, or a single OSN account that is hijacked to allow the perpetrator to act as the owner of the account. Therefore, it is difficult to say if a OSN account is the perpetrator in the dissemination of fake information online. To overcome this situation, researchers have proposed the analysis of writing style to detect whether online information is fake and whether the author who wrote/disseminated the fake information is actually the real person and not a victim of ICA. Writing style could be a good indicator of the authenticity of the author, especially should the fake author try to adapt his or her writing style for the victim's audience to match the victim's own style.

1) AUTHORSHIP DETECTION

In previous studies, researchers have used authorship identification and similarity detection to identify whether a text posted online is fake or true [21], [53] based on who wrote them. In this section, we will look at both methods of authorship detection and the features proposed by researchers to perform the detection. First, we look at authorship attribution. [109] used style markers, structural features, and content-specific features in their authorship attribution experiment. Other than using the features from previous studies, they also added extra features in each of the feature categories: seven extra style markers, two extra structural features and nine extra content-specific features. The extra features were added due to the type of English and Chinese datasets that the researchers used to predict the authors. The messages consisted of different genres: personal interest emails, school work emails, research activities emails, information technology trading messages, and bulletin board messages. Only the bulletin board messages were in Chinese. Using a classifier algorithm, [109] managed to predict the authors of the online messages with average prediction accuracies reaching 90% for email messages, 97% for the trading messages, and up to 85% accuracy for the Chinese bulletin board messages. Structural features added with style markers gave a significantly higher performance improvement rather than style markers alone, especially on the Chinese messages, which use 68.3% fewer style markers than the 205 style markers for English messages.

The researchers later improvised their authorship attribution framework by looking into four stylometry feature categories: lexical, syntactic, structural, and content-specific features [110]. There were 87 lexical features, composed of character and word-level features, for English messages. For the syntactic features, they used 158 features of punctuation and function words, excluding POS features, as the POS features cannot cater to the Chinese language. For the structural features, they used 14 features focused on email layout.

Since the dataset that the researchers used was the same as in their 2003 publication, the content-specific features covered online trading and sale keywords. Some of the features were also used for the Chinese language messages, including some added features that are specifically for the Chinese language. The best accuracy for author attribution on online messages for both languages was when all the stylometry features were combined together, as they have found in their previous work [109]. The four stylometric features in [110] were also used in [15] with an addi-

tion of n-gram based features to do author attribution on 341 online chat-room posts. The results also showed that the combination of stylometric features yielded the best accuracy in different statistical and machine learning approaches for authorship attribution, even for short messages such as chat posts.

Another study on authorship detection for different languages is the study conducted by [111]. The researchers used Thai and English text datasets collected from webboards discussing entertainment and politics and two political fanpages. Overall, the datasets contained 150 articles. Fifty-three writing attributes including Thai and English characters, words, punctuation marks, emotions, structure features and content features were used in the experiment. Similar to the result from [109], the accuracy of authorship detection was high using writing style features even with a language other than English.

The second approach to authorship deception is authorship verification. [112] verifies an author by identifying the uncharacteristic writing behavior of an author using the combination of stylometry and content features. Nine stylometric feature categories cover the different types of characters, such as the alphabets, symbols and digits, punctuation marks, POS tags, grammatical features such as nouns, adjectives and verbs, first-person pronoun frequency, lexical diversity, average sentence length, average word length, and total number of words. The content features are topics and keywords related to the topics. Two types of document dataset are tested: blogs and the Brennan-Greenstadt attack corpus, where other authors conceal their own identity by imitating the writing style of writer Cormac McCarthy. In the blogs dataset, [112] was able to identify a deceptive author 89% of the time on average from a set of authors using both stylometric and content features, while on the imitation writing dataset, all of the deceptive authors were able to be verified using just the stylometric features.

Authorship verification can also be applied for detecting phishing emails. [113] outperformed the two common phishing email detection methods, PILFER and FSSPD, by 10%. The researchers combine stylometry (97 features), gender features (7 features), and personality features based on emotion words (15 features) in order to verify the author of a phishing attack, which is also known as spear-phishing. A spear-phishing is a targeted attack where the author disguises themselves as a trusted sender and sends the phishing email to the target victim. The combination features are used to verify and authenticate the email sender. Classification techniques are included in the authorship verification model to profile the traits of the original author. If the traits of the sender are not consistent with the real author's traits, the email will be tagged as a spear-phishing email.

Plagiarism is also another area where the author attribution detection can be applied as imposing other people's work as one's own is illegal and unethical. [114] has used 446 stylometric features, as well as some additional features. Those additional features include symbols such as mathematical symbols, combined-words, word endings, sentences with ''the'' at the start, and some unique words commonly used in formal writing such as, ''e.g.'' and ''etc.''). Machine learning techniques such as k-nearest neighbor (kNN) and SMO are used to identify the pattern of each author and help with the detection. Documents with 10,000 words give an accuracy of over 90% accuracy for both machine learning techniques with SMO performed the best at 98%. The researchers also find that the accuracy decreases when the document's number of words is less than 5,000, and the number of authors is more than five.

Other than kNN and SMO, Artificial Neural Networks (ANN) has also been applied in classification tasks using stylometric features to determine authorship of documents. Lexical and syntactic features have been used in the studies by [115]. Their study revealed that syntactic features outperform lexical features, where the accuracy was slightly higher when two stylometric features are combined. However, these results, when compared to other works in literature, are less convincing as a fairly small corpus of 168 fragments of text from novels by authors Henryk Sienkiewicz and Boleslaw Prus is used.

While [109], [110] and [15] focus on authorship attribution, [53] combines writing style features with authorship attribution to identify deceptive online documents. In this paper, the researcher first identifies a set of discriminating features that distinguish deceptive writing from regular writing. Three sets of features—writeprints (lexical, syntactic and content specific), lying-detection and authorship attribution detection—are used. Next, a supervised authorship recognition test is conducted where a classifier is trained on sample documents consisting of regular and deceptive documents from different authors to build a model for each author. Three datasets have been used: the Extended-Brennan-Greenstadt Corpus (regular, obfuscated and imitated writing samples); articles from the International Imitation Hemingway Competition and Faux Faulkner contest; and lastly Thomas-Amina Houx blog posts from ''A Gay Girl in Damascus''. The [53] study shows that the most effective features for detecting deceptive writing are function words such as 'I', 'there', 'are', and 'you'. The research also find regular authorship recognition to be more effective than deception detection to find indication of stylistic deception in large size deception documents. Although the study was not able to identify the author of a document, the large, content-independent feature set used in this study could detect the deceptive documents.

In addition to the previous study, [116] also used the authorship detection method to detect a cybersecurity attack. The cybersecurity attack in focus is called sockpuppet. A sockpuppet attack is where a real social media account has been forged or imitated by perpetrators in order to deceive other social media account users within their social network to believe that the forged account owner has supported a product or a piece of information posted online. [116] proposed

a combination of the authorship attribution technique and social network analysis to perform the sockpuppet detection. The authorship attribution techniques that they studied cover the four stylometric features: lexical, syntactic, structural, and content-specific.

A combination of features to perform authorship detection was also conducted by [117]. Rather than just combining stylometry features, [117] combined three different elements: personality, tone and writing style. Eight hundred tweets from 200 users were used as the dataset in this study. Pretrained word vectors for Twitter were used to train a convolutional neural network for the personality model and tone model. The personality covered is the Big 5 personality traits, and the tones include anger, fear, sadness, disgust, surprise, joy, humor, sarcasm and neutral. LIWC was used to identify the writing style of the users. The combination of features was able to accurately identify the Twitter authors, as an author will not be able to imitate all three features. This technique could be used to detect the different authorship when a OSN account has been hacked.

2) DECEPTION DETECTION

Online reviews have been a target by scammers and spammers to trick readers into believing that a product is good. A reader's honesty in writing a review can be a difficult task to prove, as spammers always try to write wise reviews, following the structure of an honest review. To detect which reviews are honest and true review and which are fake, deception detection is needed. Not just online reviews need to be assessed for their truthfulness; online news is also as important. In the study by [118], three types of fake news were found online: fabricated news, hoaxes and humorous fakes such as parody news and satire. For readers who have no idea of the real news situation or the type of news agency website, for example, tabloids or known satire news sites such as The Onion, these readers may easily believe fake news that they read online. It is important to detect these online deceptive texts. In this section, we will discuss the deception detection method using writing styles proposed by previous researchers.

Other than online platforms, emails are also prone to be targeted by scammers to deceive people through an attack known as phishing. There are two types of phishing: deceiving people in order to collect personal information and spreading malwares [119]. In phishing attacks, victims will receive an email with texts that influence them to click on a link. The website link is where cybercrime occurs i.e., spreading malware and collecting personal information [120]. The body of the phishing email, i.e., text, sometimes acts as a threat, a reward, or an offer to deceive the victim. Most phishing detection methods focus on layout, web hyperlink embedded in the email, keywords, and email address [121]–[123] than writing styles. In this section, we will discuss the deception detection method of various cyber attacks using writing styles proposed by previous researchers.

Reference [124] used syntactic stylometry features for deception detection within essay texts such as product reviews and trip reviews by considering unigram, bigram, and their combination as features. To strengthen the unigram features, the researchers combined POS tags with the unigram. The next features that they used were the probabilistic context-free grammar (PCFG) parse trees that included unlexicalized and lexicalized production rules. The PCFGs were also combined with the unigram features. The features driven from the PCFG combination with unigram features gave a consistent improvement in the deception detection of essays.

Reference [125] and [126] also used hotel reviews as their dataset to detect deceptive reviews from truthful reviews. While the reviews in [125] are for hotels in Chicago, USA, in [126], they are for 15 hotels from five popular tourist destinations within Asia, ranging from luxury to mid-range and budget hotels. The number of reviews, deception detection aims and techniques are also different between the two studies. Reference [125] use 80 hotel reviews that are divided between positive and negative reviews, with each category consisting of the same number of deceptive and truthful reviews. Bag-of-characters (BOC) n-grams and BOW n-grams are used as comparisons to perform deception detection on the hotel reviews. The aim of the study is to identify which technique works best to capture deceptive content. Their results show that between the two techniques, the BOC n-grams technique was able to capture the deceptive opinion content better than BOW n-grams.

While [125] used n-grams, [126] used the common LIWC and POS for their deception detection for online reviews across different categories of hotels. Forty-four variables—six on comprehensibility features, 16 on specificity features, nine on negligence features and 13 on exaggeration features—were used in the study to detect deceptive opinions from 1,800 hotel reviews of different hotel categories. However, the results showed inconsistencies in features between the deceptive and truthful reviews, therefore making it difficult to ascertain the authenticity of the reviews.

Another deception detection method for hotel reviews, but using a stylometric lexical and syntactic approach, has been proposed by [127]. Seventy-seven lexical and 157 syntactic features and two different classifiers, i.e., the SVM with SMO and naive Bayes for deceptive review classification, are used. The lexical and syntactic features are used separately and combined. Their results show that stylometric features are useful in detecting deceptive reviews online.

The lexical and syntactic features such as POS, TTR, lexical diversity (distribution of content words such as nouns, verbs, adjectives and adverbs) and others have also been found to be useful for checking spam reviews. The work by [128] focuses on 25 syntactic features from the syntactic development in language learners. The syntactic complexity is measured by sentential clause and T-unit (measuring the smallest word group in a grammatical sentence) length of unit. The results from [128] show that the combination of

both lexical and syntactic features managed to detect deceptive reviews better than separate features. The researchers' proposed method could correctly identify 93.3% of truthful reviews and 89.5% of fake reviews.

Although most studies focus on online reviews, we have found that the Potthast *et al.* [129] dataset is different, as they look into fake news. Using 1,627 political news articles from mainstream publishers, right-wing and left-wing political bias was annotated by five journalists. The dataset came from the BuzzFeed-Webis Fake News Corpus. Overall, the corpus contains 299 fake news articles. Character n-grams, stop words, POS with n-grams, readability and dictionary features, and domain features such as hyperlinks and quotes are used to detect the fake news. While the study is able to distinguish satire news, detecting fake news using writing style alone is harder.

On phishing attacks, [130] combines both the stylometrics technique on the email title, body and attachment, with publicly available content on the victim's LinkedIn profile to find the connection between social footprint and phishing target (specifically spear-phishing attacks). However, the social footprint gives little significant outcome in the detection method, but it would probably give better detection if the targeted receiver has a rich and strong social footprint.

While previous studies focus on English deceptive texts, other studies have looked at non-English texts [16], [131], [132]. [131] use texts written in English and Spanish. The English texts come from two different countries, US and India, and the Spanish texts from Mexico. Two hundred statements (for each topic) on three different topics (abortion, death penalty and best friend) are collected for the English texts from each country, as well as 78 statements for abortion, 84 statements for the death penalty, and 188 statements for the best friend topic in the Spanish language. The statements contain equal numbers of deceptive and truthful statements. The study aims to address cross-cultural deception detection. LIWC and unigrams are used to do the detection, and LIWC is found to be able to serve as a bridge for cross-cultural classification. For deceptive detection, 60% to 70% accuracy rates are achieved.

Focusing on web fraud and scams, [16] use computational linguistic features from NLP and psycholinguistic features from LIWC to detect the mentioned cybercrimes. The fraud dataset used in this study comes from the publicized Enron email messages, including the 89 fraudulent email messages from the Enron chairman. The scam dataset consists of Facebook post messages from 1,036 email account holders of publicly leaked Nigerian cybercriminals and their friends on Facebook. The Enron email messages are in English, while the Facebook post messages are in a combination of Nigerian languages, Spanish and French. It is found that fraud messages were detected with 60% accuracy using the predictive model from the scam dataset, while scams on Facebook were detected with 50% accuracy using the predictive model from the fraud email dataset. Fraud and scam messages both contain verbose words, but fraud messages are more expressive and use high lexical diversity words, while scam messages are more redundant and less complex, despite the difference in languages.

Another non-English dataset is the Russian one used in the study by [132]. The study uses 113 pairs of deceptive and truthful narrative statements from 113 people compiled in the Russian Deception Bank Corpus. The dataset is marked into three groups: psycholinguistic (from the MRC Psycholinguistic Database) and sentiment (positive and negative emotion words, including 36 different emotions and attitudes words); 11 POS tags and POS tag bigram frequencies, excluding the white space and punctuation marks; and lastly, syntactic (18 parameters in Russian) and readability (Automated Readability Index and the Coleman–Liau readability formula) features. POS tags and POS tag bigrams features give the best implication for binary text classification on automated deception detection for the Russian language. Table 9 provides an overview of the writing style techniques and the datasets discussed in this section.

C. SUMMARY

This section reviewed previous studies regarding the use of writing styles to detect cybersecurity issues of authorship and fake online information. We summarize the studies in Table 9. Most cybersecurity studies support the following definitions of style: as an author-function and as a language choice. A high percentage of the literature uses similar online review datasets, either for hotels or products, and website board systems such as forums and chat rooms. In regard to detecting deceptive texts, most studies in the literature created the deceptive dataset rather than using real-world data. The approach is designed to ensure that the texts are actually fake and deceptive. However, there is a possibility that the people asked to write the deceptive texts do not have the same deceptive style as the cybercriminals, thus making it harder for the detection model to distinguish between what is true and fake [126]. Even with real-world data, it is not an easy task to detect fake news [129]. We also see that there is an interest in deception detection in non-English texts, but these studies are few compared to the English dataset. This topic could be further explored in future research. Furthermore, the intention of the writer for disseminating deceptive and fake online information needs to be taken into account as not all such information is derived from the ill-intention act, as it may also be due to the sharers' negligence or naivety. Moreover, combined with the feeling of social responsibility that either made them share the information as it is or edit the texts to show their support. There is a social and juristic dilemma that is faced here. For example, should they be condemned as accomplices to the cybercrime, or should more focus be given on the detection methods that would be able to judge the information as fake? Such dilemma of drawing the line between the right and the wrong for machines and their engineers is discussed by [133].

TABLE 9. Writing style features to detect cybersecurity issues.

Literature	Data	Writing style features				
		Lexical	Syntactic	Structural	Content-specific	Additional features
[109] (2003)	70 email (English), 153 newsgroup messages (English), 70 Bulletin Board System (BBS) messages (Chinese).		✓	✓	✓	
[110] (2006)	20 authors' newsgroup messages (English), 20 authors' Chinese BBS messages.	✓	✓	✓		LIWC.
[115] (2008)	Documents of Henryk Sienkiewicz and Bolesław Prus	✓	✓			
[15] (2012)	341 chat posts.	✓	✓	✓	✓	LIWC, n-grams.
[112] (2012)	28,500 Spinn3r Blog posts from 2,194 authors.	✓	✓	✓	✓	
[53] (2012)	68 authors' samples (Brennan-Greenstadt Corpus & Amazon Mechanical Turk (AMT) workers), 18 articles (The International Imitation Hemingway Competition) and 15 articles (The Faux Faulkner Contest), 20,500-word posts from the Amina and Thomas Yahoo! group and 248,500-word articles from the "A Gay Girl in Damacus" blog.	✓	✓		✓	Lying-detection feature, 9-feature set.
[124] (2012)	400 truthful reviews from TripAdvisor and 400 deceptive reviews from AMT workers for 20 hotels in Chicago, 400 filtered and displayed reviews for 35 Italian restaurants from Yelp, truthful and deceptive essays from AMT workers on abortion, best friend, and death penalty topics.		✓	✓		Unigram, bigram, PCFG.
[116] (2013)	539 users, 11 threads and 4,951 replies of Tianya Dataset, Taobao Dataset, 127 home pages and 9,854 comments.				✓	Social network model.
[127], [128] (2013, 2015)	400 truthful positive reviews from both TripAdvisor and AMT workers, 400 truthful negative reviews from Expedia, Hotels.com, Orbitz, Priceline, TripAdvisor, and Yelp and 400 deceptive negative reviews from AMT workers.	✓	✓			
[111] (2014)	75 entertainment and 25 political messages, 50 messages from political fanpages (all in Thai and English).		✓	✓	✓	Word analysis, emotion (emoticons and Internet slang).
[131] (2014)	100 English deceptive statement and 100 truthful statements on 3 different topics (abortion, death penalty and best friend) from US and India origin AMT workers, 78 Spanish statements (abortion topic), 84 Spanish statements (death penalty topic), and 188 Spanish statements (best friend topic) from Mexico origin authors.					LIWC, Unigram.
[130] (2014)	4,742 targeted spear phishing emails sent to 2,434 unique victims, 9,353 non-targeted attack emails sent to 5,912 unique non-victims, and 6,601 normal emails from the Enron dataset sent to 1,240 unique Enron employees.	✓	✓	✓	✓	Social media features (LinkedIn).
[125] (2015)	1,600 hotel reviews for 20 hotels in Chicago.					BOC n-grams, BOW n-grams.
[126] (2016)	1,800 hotel reviews from 15 hotels in Asia (luxury, mid-range and budget hotel categories).			✓		LIWC.
[16] (2016)	500,000 Enron email messages, Facebook messages of 1036 Nigerian cybercriminals.					LIWC, NLP.
[114] (2016)	10 thesis by 10 authors	✓	✓	✓	✓	n-grams, symbol, specific words.
[117] (2017)	800 tweets (200 authors).					LIWC, Big 5 personality traits, emotion tone.
[129] (2017)	1,627 political articles from BuzzFeed-Webis Fake News Corpus					Characters n-grams, content-specific features.
[132] (2017)	113 pairs of deceptive and truthful statements from 113 people compiled from the Russian Deception Bank Corpus.		✓			MRC Psycholinguistic Database, sentiment words, Automated Readability Index and the Coleman-Liau Readability Formula.
[113] (2019)	More than 50 emails per sender of 10 senders from the Enron Email datasets.	✓	✓	✓	✓	Idiosyncratic, gender and personality features.

VI. EXPERIMENTS

All three disciplines use demographics, such as gender, in their studies. As an example, with respect to personality / behavior and cybersecurity studies, the interest is finding out the demographics of sexual predators. To the best of our knowledge, morphology has not been explored in English texts as it is a gender neutral language. In this section, we will explore the use of tense morphology in gender and age studies in English blogs, and the PAN 2013 dataset. Morphology is the study of words, how they are formed and their relationship to other words in the same language [134]. It is commonly used in the non-English texts, as many non-English languages have morphological gender markers that do not exist in the English language.

The study in [135] shows that past tense morphology has a different mental representation in aphasic patients. Inspired by this work, we focus on the connection between syntactic patterns of "to be" verbs in simple past, simple present, present participle and past participle tenses, with age and gender.

A. DATA

For the purpose of comparing the experimental results of Schler *et al.* [70] and Goswami *et al.* [71], we used their blog dataset that is publicly available (see Table 10).

TABLE 10. Blog distribution over age and gender.

Category	Age	Male	Female	Total
10	13-17	4120	4120	8240
20	23-27	4043	4043	8086
30	33-42	1226	1234	2460
	Total	9389	9397	18786

B. DISTINGUISHING FEATURES

In this section, we consider differences among bloggers of age categories from Table 10 and gender. We analyze the syntactic patterns of the above tenses using POS classes, summarized in Table 11. For each pattern (known as feature), we measured the frequency it appears in the dataset, by age categories and by gender. Extreme gradient boosting (XGBoost) classifier and spaCy [136] are used to develop a prediction model to extract relevant features based on the highest information gain. We then used these features to predict age categories and gender. Stratified K-Folds, a commonly accepted cross validation technique, with 10-folds is used here.

In Table 12, we show the frequency of syntactic patterns with their highest information gain based on the age categories. All differences are statistically significant at $p < 0.00002$. Similarly, in Table 13, we show the same statistics,

TABLE 11. Main POS classes used in the experiments.

POS class	Penn Tree Tagset
Nouns	NN - Singular noun NNS - Plural noun NNPS - Plural proper noun
Verbs	VB - Verb base form VBD - Verb past tense VBG - Verb, gerund or present participle VBN - Verb, past participle VBP - Verb, non-3rd person singular present VBZ - Verb, 3rd person singular present
Adverbs	RB - Adverb WRB - Wh- adverb
Pronouns	PRP - Personal pronoun WP - Wh- pronoun

TABLE 12. The features for age categories are ordered based on highest information gain.

	Feature	10(%)	20(%)	30(%)
1	PRP+RB+VBD	43.86	38.90	34.26
2	NNS+VBG	11.54	15.93	16.91
3	NNS+VBD	25.81	28.24	33.27
4	VBG+NNPS	0.31	0.43	0.56
5	NN+RB+VBN	1.04	1.22	1.25
6	VBN+WP	0.51	0.58	0.61
7	PRP+RB+VBG	1.29	1.24	0.87
8	MD+PRP	15.51	13.33	12.10
9	VBZ+NNPS	0.13	0.14	0.16

TABLE 13. The features for gender are ordered based on highest information gain.

	Feature	Male	Female
1	PRP+RB+VBG	0.71	0.73
2	PRP+EX+VBP	0.02	0.02
3	NNS+RB+VBD	1.30	1.01
4	NN+WRB+VBG	0.19	0.12
5	VB+PRP	96.24	97.42
6	PRP+EX+MD	0.01	0.01
7	NNPS+VBD	1.53	0.68

but by gender. All differences are statistically significant at $p < 0.007$.

C. RESULTS

Age experiments were run on the combined dataset of the three age categories. Compared to Schler *et al.* (76.2%) and Goswami *et al.* (80.4%), a higher accuracy is obtained (98.2%) using our approach. The confusion matrix is shown in Table 14. Similarly, gender experiments were run on the combined dataset of the male and female categories. Compared to Schler *et al.* (80.1%) and Goswami *et al.* (89.3%), a higher accuracy is obtained (97.4%) using our approach. The confusion matrix is shown in Table 15.

D. CONCLUSION

As an alternative dataset, we used PAN 2013 [45]. This dataset consist of 413,555 documents comprising of blog posts, OSN posts such as Netlog, and short conversations including conversations of sexual predators (see Table 16).

TABLE 14. Confusion matrix for the age classifiers.

Classified as →	10	20	30
10	8127	151	8
20	28	7869	2
30	85	66	2450

TABLE 15. Confusion matrix for the gender classifiers.

Classified as →	Male	Female
Male	9278	405
Female	111	8992

TABLE 16. Various document distribution over age and gender.

Category	Age	Male	Female	Total
10	13-17	13836	13815	27651
20	23-27	76796	79057	155853
30	33-47	118912	111139	230051
	Total	209544	204011	413555

TABLE 17. The features for age categories are ordered by highest information gain.

	Feature	10(%)	20(%)	30(%)
1	WP+RB+VBZ	5.94	12.01	2.86
2	VBG+WP	4.40	3.16	8.27
3	PRP+RBR+VBG	3.42	2.85	4.34
4	NN+WRB+VBP	4.73	8.05	3.41
5	VBP+EX	7.55	6.36	9.72
6	NNPS+VB	5.04	11.71	6.86
7	VBG+NN	6.49	3.42	3.44
8	PRP+RB+VBN	7.48	6.31	11.30
9	EX+MD	6.63	4.04	7.61
10	NN+RBR+VBZ	10.21	5.28	9.75
11	PRP+VB	4.73	2.08	0.62
12	VBD+EX	1.63	1.50	0.77
13	NNS+RB+VBN	2.69	4.66	10.98
14	NNS+WRB+VB	3.43	3.18	1.18
15	NNPS+MD	2.46	0.80	1.10
16	EX+RBS+VBN	3.61	3.80	0.74
17	NNS+WRB+VBP	6.25	8.68	8.48
18	EX+VBZ	6.14	3.32	3.47
19	WP+WRB+MD	3.56	6.21	2.07
20	WP+MD	3.60	2.58	3.0

Like the previous dataset, we show the frequency of syntactic patterns with their highest information gain based on age categories in Table 17 and based on gender in Table 18. All differences are statistically significant at $p < 0.001$.

We achieved an accuracy of 94.0% for age experiments, and the confusion matrix is depicted in Table 19. In contrast, we obtained an accuracy of 95.5% for gender experiments, and the confusion matrix is depicted in Table 20.

To conclude, though both the datasets have about similar accuracies, the selected syntactic patterns are different, which is expected as it will depend on the type of dataset, including its statistics such as length. This further goes to show that different datasets will have different syntactic patterns. Given the high prediction results, we emphasize that our novel machine learning prediction model that is based on tense morphology, is very promising in age and gender classification from English texts.

TABLE 18. The features for gender are ordered based on highest information gain.

	Feature	Male	Female
1	VBG+NN	10.11	2.35
2	EX+MD	3.64	7.24
3	NNP+VB	11.52	14.63
4	WP+RB+VBZ	5.29	1.97
5	VBG+WP	7.65	2.84
6	PRP+VB	1.89	0.39
7	VBG+NNP	8.26	15.19
8	NNPS+MD	1.95	0.58
9	NNPS+VB	5.07	8.37
10	NN+WRB+VBP	2.99	2.31
11	PRP+RBR+VBG	3.46	2.13
12	PRP+RB+VBN	5.76	9.59
13	NNS+RBR+VBN	2.11	3.68
14	NNS+WRB+VBN	3.40	4.02
15	PRP+WRB+VBZ	4.77	2.53
16	NN+RBR+VBZ	4.97	5.04
17	VBP+EX	6.84	5.59
18	EX+VBZ	3.50	2.30
19	NNS+WRB+VBP	6.17	8.17
20	WP+VBZ	0.64	1.09

TABLE 19. Confusion matrix for the age classifiers.

Classified as →	10	20	30
10	24359	63	34
20	2168	155291	20739
30	1124	499	209278

TABLE 20. Confusion matrix for the gender classifiers.

Classified as →	Male	Female
Male	208015	17278
Female	1529	186733

VII. CONCLUDING REMARKS AND FUTURE WORK

In this survey paper, we have 1) provided definitions of writing styles used across three multidisciplinary factors: demographics, personality & behavior, and cybersecurity; 2) presented common writing styles during pre- and post-Internet periods; 3) presented an overview of writing style feature categories across the stated factors; 4) provided an in-depth comparison of writing style techniques in OSN text across the stated factors; 5) presented common writing style techniques and datasets in OSN text across the stated factors; 6) presented variation in speech (whereby people adapt their language to their conversational partners) as a writing style in OSN text in demographics and cybersecurity; and 7) proposed a novel machine learning prediction model based on tense morphology, to classify age and gender from English blogs, and the PAN 2013 dataset.

In a nutshell, the information derived from writing style can be used to derive specific information. Writing style has features that range from stylometry to additional features. The stylometry features relevant to this survey paper are lexical, syntactic, structural, and content-specific. Additional features include LIWC, 20 factors, LBA, LDA, BOW, BOC, TF-IDF, n-grams, semantic, morphology and idiosyncratic.

This paper has shown that writing styles study has been used across a multidisciplinary field where similar datasets can be employed for different disciplines. Take the PAN 2013 dataset as an example, where the dataset contains online sexual predator documents, which is a cybersecurity issue. However, it can also be utilized to study the demographics and behavior of the perpetrator. Other than that, similar features are also seen across disciplines. Lexical and syntactic features are the most common writing style features used as they represent the writing characteristics of the author.

Another pattern that has been discovered is that the combination of features from different disciplines. In the cybersecurity and personality & behavior disciplines, researchers are seen to use keywords that represent demographics such as gender and age as part of the features to identify the author. There are also some papers in the cybersecurity discipline that either use personality features or combine both demographics and personality features into their detection methods [111], [113], [117], [132]. Therefore, future studies regarding writing styles will also benefit by combining multidisciplinary domains as people are complex and not one dimensional.

Numerous future research directions are possible.

Exploring writing styles on dictated text Numerous speech to text applications exists where one can easily share the dictated text as an email or even, as a tweet. It offers the flexibility of a word processing editor where an individual can edit the text to perfection. Some work has been done to differentiate spontaneously and dictated speech in an Indonesian dataset [137]. It would be interesting to see if writing style differences exist between dictated text and OSN text.

Reranking pages based on writing styles In relation to information retrieval, implementing writing styles for reranking pages based on readers' personality or writing style preferences could be an option. This could be part of the personalization research area, where a person's preferences are included as part of the information retrieval process and writing preferences can be part of that feature. Currently, some web search personalization techniques use individual search history and user demand [138], trust-based hubs and authorities [139] and features presented on a page to render heterogeneous results onto the search result page [140].

Writing style as an authentication mechanism There have been multiple techniques in authentication, including textual, graphical and biometric [141] methods. As biometrics are a unique feature in every human being [142], and writing style is also considered a unique personal feature, similar to fingerprints, we believe that writing style can also be considered a security mechanism [143].

Incorporating writing style from handwritten allographs Characteristics of handwritten allographs (e.g., middle zone height, middle zone breadth, upper zone height, lower zone height), including left- or right-handedness, have been related to personality [96] and to

forensics [144]. An attempt to relate left- or right-handedness to demographics was made in [6], but it achieved little success due to the limited quantity of data. A review study can be undertaken to relate common writing styles in handwritten allographs with OSN text across multidisciplinary factors.

Incorporating unexplored writing styles As common writing style techniques are seen across multidisciplinary factors studies, the unexplored techniques can be applied across disciplines. An example is morphology, which has been explored in an Indonesian dataset by [69], but it has not been explored on the modern Malay dataset even though they share somewhat similar origins. The modern Malay alphabet is a Latin alphabet consisting of 26 alphabets, similar to the English and Indonesian language alphabet. Likewise, idiosyncratics is applied in cybersecurity, but not in the demographics and personality & behavior.

Exploring writing styles on Manglish-type datasets To the best of our knowledge, writing styles have not been employed on English-based pidgin datasets such as Manglish (Malaysian English) in OSN text. Pidgin, according to Merriam-Webster, is ''a simplified speech used for communication between people with different languages''. In contrast, Manglish is a different sort of pidgin as it is mainly made up of (at least 99%) English words, where some elements are borrowed from at least the three main languages of Malaysia (Malay, Chinese and Indian). One excerpt of Manglish is from [145]: ''The school are so many teacher and friend. I can read the book in this school.'' Though this sentence may use English vocabulary, it does not follow the grammar rules for a sentence construction. Further, the study of [146] on the handwritten text of Malay undergraduate students supports this observation, showing that the similarities and differences in the English and Malay languages give rise to substitution (''use of native language forms in the target language'') and calques (''errors that reflect very closely a native language structure'').

ACKNOWLEDGMENT

Many thanks to Mark Sanderson for fruitful discussions and helpful comments.

REFERENCES

[1] E. L. Moerk, ''Quantitative analysis of writing styles,'' *J. Linguistics*, vol. 6, no. 2, pp. 223–230, Sep. 1970.

[2] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, ''Surveying stylometry techniques and applications,'' *ACM Comput. Surveys*, vol. 50, no. 6, pp. 1–36, Jan. 2018.

[3] D. Hovy, ''Demographic factors improve classification performance,'' in *Proc. 53rd Annu. Meeting Assoc. Comput. Linguistics, 7th Int. Joint Conf. Natural Lang. Process.*, 2015, pp. 752–762.

[4] J. W. Pennebaker and L. A. King, ''Linguistic styles: Language use as an individual difference.,'' *J. Personality Social Psychol.*, vol. 77, no. 6, pp. 1296–1312, 1999.

[5] R. Plamondon and S. N. Srihari, ''Online and off-line handwriting recognition: A comprehensive survey,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, pp. 63–84, Jan. 2000.

[6] S. N. Srihari, S.-H. Cha, H. Arora, and S. Lee, ''Individuality of handwriting,'' *J. Forensic Sci.*, vol. 47, no. 4, Jul. 2002, Art. no. 15447J.

[7] C. Ramaiah, U. Porwal, and V. Govindaraju, ''Accent detection in handwriting based on writing styles,'' in *Proc. 10th IAPR Int. Workshop Document Anal. Syst.*, Mar. 2012, pp. 312–316.

[8] E. F. Kotzé, ''Author identification from opposing perspectives in forensic linguistics,'' *Southern Afr. Linguistics Appl. Lang. Stud.*, vol. 28, no. 2, pp. 185–197, Oct. 2010.

[9] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, ''Automatically profiling the author of an anonymous text,'' *Commun. ACM*, vol. 52, no. 2, pp. 119–123, Feb. 2009.

[10] J. B. Herrmann, K. van Dalen-Oskam, and C. Schöch, ''Revisiting style, a key concept in literary studies,'' *J. Literary Theory*, vol. 9, no. 1, pp. 25–52, Jan. 2015.

[11] T. M. Arthur, *The Oxford Companion to the English Language (A Bridge Edition)*. Oxford, U.K.: Oxford Univ. Press, 1996.

[12] P. Eckert, ''Age as a sociolinguistic variable,'' in *The handbook Sociolinguistics*. Oxford, U.K.: Blackwell, 1997.

[13] P. Eckert, ''The whole woman: Sex and gender differences in variation,'' *Lang. Variation Change*, vol. 1, no. 3, pp. 245–267, 1989.

[14] R. Wodak and G. Benke, ''Gender as a sociolinguistic variable: New perspectives on variation studies,'' in *The Handbook Sociolinguistics*, F. Coulmas, Ed. Oxford, U.K.: Blackwell, 1997, pp. 127–150.

[15] F. Amuchi, A. Al-Nemrat, M. Alazab, and R. Layton, ''Identifying cyber predators through forensic authorship analysis of chat logs,'' in *Proc. 3rd Cybercrime Trustworthy Comput. Workshop*, Oct. 2012, pp. 28–37.

[16] A. Mbaziira and J. Jones, ''A text-based deception detection model for cybercrime,'' in *Proc. Int. Conf. Technol. Manag*, 2016, pp. 1–8.

[17] M. H. Altakrori, F. Iqbal, B. C. M. Fung, S. H. H. Ding, and A. Tubaishat, ''Arabic authorship attribution: An extensive study on Twitter posts,'' *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 18, no. 1, pp. 1–51, Jan. 2019.

[18] R. A. Posner, ''Judges' writing styles (and do they matter?)'' *Univ. Chicago Law Rev.*, vol. 62, no. 4, pp. 1421–1449, 1995.

[19] L. A. Heymann, ''The birth of the authornym: Authorship, pseudonymity, and trademark law,'' *Notre Dame L. Rev.*, vol. 80, p. 1377, May 2004.

[20] M. Foucault, ''What is an author,'' in *The Foucault Reader: An Introduction to Foucault's Thought*, P. Rabinow, Ed. London, U.K.: Penguin, 1991.

[21] A. Abbasi and H. Chen, ''Writeprints: A stylometric approach to identity-level identification and similarity detection in cyberspace,'' *ACM Trans. Inf. Syst.*, vol. 26, no. 2, pp. 1–29, Mar. 2008.

[22] A. Bell, ''Language style as audience design,'' *Lang. Soc.*, vol. 13, no. 2, pp. 145–204, 1984.

[23] R. S. Campbell and J. W. Pennebaker, ''The secret life of pronouns: Flexibility in writing style and physical health,'' *Psychol. Sci.*, vol. 14, no. 1, pp. 60–65, Jan. 2003.

[24] Q. A. Bui, M. Visani, S. Prum, and J.-M. Ogier, ''Writer identification using TF-IDF for cursive handwritten word recognition,'' in *Proc. Int. Conf. Document Anal. Recognit.*, Sep. 2011, pp. 844–848.

[25] A. M. M. M. Amaral, C. O. D. A. Freitas, and F. Bortolozzi, ''Combination and analysis of features from forensic letters,'' in *Proc. Int. Conf. Artif. Intell. (ICAI)*, 2014, pp. 1–6.

[26] M. Bulacu and L. Schomaker, ''Text-independent writer identification and verification using textural and allographic features,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 4, pp. 701–717, Apr. 2007.

[27] S. Argamon, M. Koppel, J. Fine, and A. R. Shimoni, ''Gender, genre, and writing style in formal written texts,'' *Text-Interdiscipl. J. Study Discourse*, vol. 23, no. 3, pp. 321–346, Jan. 2003.

[28] V. Ganjigunte Ashok, S. Feng, and Y. Choi, ''Success with style: Using writing style to predict the success of novels,'' in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2013.

[29] F. Mosteller and D. Wallace, *Applied Bayesian and Classical Inference: The Case of the Federalist Papers*, 2nd ed. New York, NY, USA: Springer-Verlag, 1984.

[30] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, ''Mining the blogosphere: Age, gender and the varieties of self-expression,'' *1st Monday*, vol. 12, no. 9, 2007.

[31] J. D. Burger and J. C. Henderson, ''An exploration of observable features related to blogger age,'' in *Proc. AAAI Spring Symp., Comput. Approaches Analyzing Weblogs*, Stanford, CA, USA, Mar. 2006, pp. 15–20.

[32] S. Rosenthal and K. R. McKeown, "Age prediction in blogs: A study of style, content, and online behavior in pre- and post-social media generations," in *Proc. 49th Annu. Meeting Assoc. Comput. Linguistics, Hum. Lang. Technol. Conf.*, Portland, OR, USA, Jun. 2011, pp. 763–772.
[33] M. Del-Teso-Craviotto, "Gender and sexual identity authentication in language use: The case of chat rooms," *Discourse Stud.*, vol. 10, no. 2, pp. 251–270, Apr. 2008.
[34] C. Peersman, W. Daelemans, and L. Van Vaerenbergh, "Predicting age and gender in online social networks," in *Proc. 3rd Int. Workshop Search Mining User-Generated Contents (SMUC)*, 2011, pp. 37–44.
[35] C. A. MacArthur, "The impact of computers on the writing process," *Exceptional Children*, vol. 54, no. 6, pp. 536–542, Apr. 1988.
[36] D. Nguyen, A. S. Doğruöz, C. P. Rosé, and F. de Jong, "Computational sociolinguistics: A survey," *Comput. Linguistics*, vol. 42, no. 3, pp. 537–593, Sep. 2016.
[37] J. Hinds and A. N. Joinson, "What demographic attributes do our digital footprints reveal? A systematic review," *PLoS ONE*, vol. 13, no. 11, 2018, Art. no. e0207112.
[38] S. Elmanarelbouanani and I. Kassou, "Authorship analysis studies: A survey," *Int. J. Comput. Appl.*, vol. 86, no. 12, pp. 22–29, 2014.
[39] E. Stamatatos, "A survey of modern authorship attribution methods," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 60, no. 3, pp. 538–556, Mar. 2009.
[40] K. Luyckx and W. Daelemans, "The effect of author set size and data size in authorship attribution," *Literary Linguistic Comput.*, vol. 26, no. 1, pp. 35–55, Apr. 2011.
[41] S. Ashraf, H. R. Iqbal, and R. M. A. Nawab, "Cross-genre author profile prediction using stylometry-based approach," in *Proc. CLEF Working (CEUR)*, vol. 1609, 2016, pp. 992–999.
[42] M. Stankevich, V. Isakov, D. Devyatkin, and I. Smirnov, "Feature engineering for depression detection in social media," in *Proc. 7th Int. Conf. Pattern Recognit. Appl. Methods*, 2018, pp. 426–431.
[43] C. D. Hollingsworth, "Syntactic stylometry: Using sentence structure for authorship attribution," Ph.D. dissertation, Univ. Georgia, Athens, Georgia, 2012. Accessed: Mar. 25, 2019.
[44] T. R. Reddy, B. V. Vardhan, M. GopiChand, and K. Karunakar, "Gender prediction in author profiling using ReliefF feature selection algorithm," in *Intelligent Engineering Informatics*, V. Bhateja, C. A. C. Coello, S. C. Satapathy, and P. K. Pattnaik, Eds. Singapore: Springer, 2018, pp. 169–176.
[45] (2013). *Pan Author Profiling*. Accessed: Dec. 22, 2019. [Online]. Available: https://pan.webis.de/clef13/pan13-web/author-profiling.html
[46] Merriam-Webster. (2019). *Merriam-Webster Since 1828*. [Online]. Available: https://www.merriam-webster.com
[47] W. Daelemans, "Explanation in computational stylometry," in *Proc. 14th Int. Conf. Comput. Linguistics Intell. Text Process. (CICLing)*, vol. 2. Berlin, Germany: Springer-Verlag, 2013, pp. 451–462.
[48] H. van Halteren, H. Baayen, F. Tweedie, M. Haverkort, and A. Neijt, "New machine learning methods demonstrate the existence of a human stylome," *J. Quant. Linguistics*, vol. 12, no. 1, pp. 65–77, Apr. 2005.
[49] A. Morton and J. McLeman, *The Genesis John*. Edinburgh, U.K.: St Andrew's Press, 1980.
[50] P. Grzybek, "History and methodology of word length studies," in *Contributions to Science Text Lang.*, P. Grzybek, Ed. Dordrecht, The Netherlands: Springer, 2007, ch. 10, pp. 15–90.
[51] D. I. Holmes, "Authorship attribution," *Comput. Humanities*, vol. 28, no. 2, pp. 87–106, 1994.
[52] D. Nguyen, D. Trieschnigg, A. S. Dogruöz, R. Gravel, M. Theune, T. Meder, and F. de Jong, "Why gender and age prediction from tweets is hard: Lessons from a crowdsourcing experiment," in *Proc. 25th COLING*, 2014, pp. 1950–1961.
[53] S. Afroz, M. Brennan, and R. Greenstadt, "Detecting hoaxes, frauds, and deception in writing style online," in *Proc. IEEE Symp. Secur. Privacy*, May 2012, pp. 461–475.
[54] J. Li, R. Zheng, and H. Chen, "From fingerprint to writeprint," *Commun. ACM*, vol. 49, no. 4, pp. 76–82, Apr. 2006.
[55] M. Baj and T. Walkowiak, "Computer based stylometric analysis of texts in polish language," in *ICAISC* (Lecture Notes in Computer Science), vol. 10246. Cham, Switzerland: Springer, 2017, pp. 3–12.
[56] A. Sboev, T. Litvinova, D. Gudovskikh, R. Rybka, and I. Moloshnikov, "Machine learning models of text categorization by author gender using topic-independent features," *Procedia Comput. Sci.*, vol. 101, pp. 135–142, 2016.
[57] J. Grieve, "Quantitative authorship attribution: An evaluation of techniques," *Literary Linguistic Comput.*, vol. 22, no. 3, pp. 251–270, 2007.
[58] J. W. Pennebaker, M. E. Francis, and R. J. Booth, "Linguistic inquiry and word count: LIWC 2001," *Linguistic Inquiry and Word Count*, vol. 71, no. 2001, p. 2001, 2019. Accessed:Apr. 28, 2019.
[59] J. Cheshire, "Sex and gender in variationist research," in *The Handbook of Language Variation and Change*, J. Chambers, P. Trudgill, and N. Schilling-Estes, Eds. Oxford, U.K.: Blackwell, 2002, pp. 423–443.
[60] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying latent user attributes in Twitter," in *Proc. 2nd Int. Workshop Search Mining User-Generated Contents (SMUC)*, 2010, pp. 37–44.
[61] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating gender on Twitter," in *Proc. EMNLP*, New York, NY, USA: ACL, 2011, pp. 1301–1309.
[62] D. Nguyen, R. Gravel, D. Trieschnigg, and T. Meder, "'How old do you think I am?' A study of language and age in Twitter," in *Proc. 7th ICWSM*. Palo Alto, CA, USA: AAAI Press, Jun. 2013, pp. 439–448.
[63] D. Bamman, J. Eisenstein, and T. Schnoebelen, "Gender identity and lexical variation in social media," *J. Sociolinguistics*, vol. 18, no. 2, pp. 135–160, Apr. 2014.
[64] S. Daneshvar and D. Inkpen, "Gender identification in Twitter using n-grams and LSA," in *Proc. 9th Int. Conf. CLEF Assoc. (CLEF)*, 2018, pp. 1–10.
[65] F. Huang, C. Li, and L. Lin, "Identifying gender of microblog users based on message mining," in *WAIM* (Lecture Notes in Computer Science), vol. 8485. Cham, Switzerland: Springer, 2014, pp. 488–493.
[66] P. S. Ludu, "Inferring gender of a Twitter user using celebrities it follows," 2014, *arXiv:1405.6667*. [Online]. Available: https://arxiv.org/abs/1405.6667
[67] Z. Miller, B. Dickinson, and W. Hu, "Gender prediction on Twitter using stream algorithms with N-Gram character features," *Int. J. Intell. Sci.*, vol. 2, no. 4, pp. 143–148, 2012.
[68] K. Surendran, O. P. Harilal, P. Hrudya, P. Poornachandran, and N. K. Suchetha, "Stylometry detection using deep learning," in *Computational Intelligence in Data Mining* (AISC), H. S. Behera and D. P. Mohapatra, Eds. Singapore: Springer, 2017, pp. 749–757.
[69] M. Ciot, M. Sonderegger, and D. Ruths, "Gender inference of Twitter users in non-English contexts," in *Proc. EMNLP*, 2013, pp. 1136–1145.
[70] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging," in *Proc. AAAI, Symp., Comput. Approaches Analyzing Weblogs*. Bethel Island, CA, USA: Springer, 2005, pp. 199–205.
[71] S. Goswami, S. Sarkar, and M. Rustagi, "Stylometric analysis of bloggers' age and gender," in *Proc. 3rd ICWSM*, E. Adar, M. Hurst, T. Finin, N. S. Glance, N. Nicolov, and B. L. Tseng, Eds. San Jose, CA, USA: AAAI Press, 2009, pp. 214–217.
[72] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proc. EMNLP*, 2010, pp. 207–217.
[73] C. Zhang and P. Zhang, "Predicting gender from blog posts," Univ. Massachusetts Amherst, Amherst, MA, USA, Tech. Rep., 2010.
[74] F. Belbachir, K. Henni, and L. Zaoui, "Automatic detection of gender on the blogs," in *Proc. ACS Int. Conf. Comput. Syst. Appl. (AICCSA)*, May 2013, pp. 1–4.
[75] J. Soler-Company and L. Wanner, "Multiple language gender identification for blog posts," in *Proc. 37th Annu. Meeting COGSCI*, 2015, pp. 1–6.
[76] D. D. Pham, G. B. Tran, and S. B. Pham, "Author profiling for vietnamese blogs," in *Proc. Int. Conf. Asian Lang. Process.*, Dec. 2009, pp. 190–194.
[77] G. K. Mikros, "Authorship attribution and gender identification in Greek blogs," *Methods Appl. Quant. Linguistics*, vol. 21, pp. 21–32, Apr. 2012.
[78] A. Johannsen, D. Hovy, and A. Søgaard, "Cross-lingual syntactic variation over age and gender," in *Proc. 19th Conf. Comput. Natural Lang. Learn.*, 2015, pp. 103–112.
[79] L. Hemphill and J. Otterbacher, "Learning the lingo?: Gender, prestige and linguistic adaptation in review communities," in *Proc. ACM Conf. Comput. Supported Cooperat. Work (CSCW)*, 2012, pp. 305–314.
[80] J. Otterbacher, "Inferring gender of movie reviewers: Exploiting writing style, content and metadata," in *Proc. 19th ACM Int. Conf. Inf. Knowl. Manage. (CIKM)*, 2010, pp. 369–378.
[81] D. Hovy, A. Johannsen, and A. Søgaard, "User review sites as a resource for large-scale sociolinguistic studies," in *Proc. 24th Int. Conf. World Wide Web (WWW)*, 2015, pp. 452–461.
[82] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining for gender prediction," in *Advances in Information Systems* (Lecture Notes in Computer Science), vol. 4243. Berlin, Germany: Springer, 2006.

[83] T. Kucukyilmaz, B. B. Cambazoglu, C. Aykanat, and F. Can, "Chat mining: Predicting user and message attributes in computer-mediated communication," *Inf. Process. Manage.*, vol. 44, no. 4, pp. 1448–1466, Jul. 2008.

[84] F. López-Escobedo, C.-F. Méndez-Cruz, G. Sierra, and J. Solórzano-Soto, "Analysis of stylometric variables in long and short texts," *Procedia-Social Behav. Sci.*, vol. 95, pp. 604–611, Oct. 2013.

[85] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting personality from Twitter," in *Proc. IEEE 3rd Int. Conf. Privacy, Secur., Risk Trust IEEE 3rd Int. Conf. Social Comput.*, Oct. 2011, pp. 149–156.

[86] J. Golbeck, C. Robles, and K. Turner, "Predicting personality with social media," in *Proc. Extended Abstr. Human Factors Comput. Syst. (CHI)*, 2011, pp. 253–262.

[87] L. Qiu, H. Lin, J. Ramsay, and F. Yang, "You are what you tweet: Personality expression and perception on Twitter," *J. Res. Personality*, vol. 46, no. 6, pp. 710–718, Dec. 2012.

[88] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, 2013, Art. no. e73791.

[89] M. De Choudhury, S. Counts, and E. Horvitz, "Predicting postpartum changes in emotion and behavior via social media," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst. (CHI)*, 2013, pp. 3267–3276.

[90] G. Roffo, C. Giorgetta, R. Ferrario, W. Riviera, and M. Cristani, "Statistical analysis of personality and identity in chats using a keylogging platform," in *Proc. 16th Int. Conf. Multimodal Interact. (ICMI)*, 2014, pp. 224–231.

[91] J. Parapar, D. E. Losada, and A. Barreiro, "Combining psycho-linguistic, content-based and chat-based features to detect predation in chatrooms," *J. UCS*, vol. 20, no. 2, pp. 213–239, 2014.

[92] G. Park, H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, M. Kosinski, D. J. Stillwell, L. H. Ungar, and M. E. Seligman, "Automatic personality assessment through social media language," *J. Personality Social Psychol.*, vol. 108, no. 6, p. 934, 2015.

[93] K.-H. Peng, L.-H. Liou, C.-S. Chang, and D.-S. Lee, "Predicting personality traits of Chinese users based on Facebook wall posts," in *Proc. 24th Wireless Opt. Commun. Conf. (WOCC)*, Oct. 2015, pp. 9–14.

[94] X.-S. Vu, L. Flekova, L. Jiang, and I. Gurevych, "Lexical-semantic resources: Yet powerful resources for automatic personality classification," 2017, *arXiv:1711.09824*. [Online]. Available: http://arxiv.org/abs/1711.09824

[95] S. D. Gosling, S. J. Ko, T. Mannarelli, and M. E. Morris, "A room with a cue: Personality judgments based on offices and bedrooms," *J. Personality Social Psychol.*, vol. 82, no. 3, pp. 379–398, 2002.

[96] M. Williams, G. Berg-Cross, and L. Berg-Cross, "Handwriting characteristics and their relationship to Eysenck's extraversion-introversion and Kagan's impulsivity-relfectivity dimensions," *J. Personality Assessment*, vol. 41, no. 3, pp. 291–298, Jun. 1977.

[97] R. von Solms and J. van Niekerk, "From information security to cyber security," *Comput. Secur.*, vol. 38, pp. 97–102, Oct. 2013.

[98] X. Chen, R. Chandramouli, and K. P. Subbalakshmi, "Scam detection in Twitter," in *Data Mining for Service*. Berlin, Germany: Springer, 2014, pp. 133–150.

[99] M. J. Metzger, A. J. Flanagin, and R. B. Medders, "Social and heuristic approaches to credibility evaluation online," *J. Commun.*, vol. 60, no. 3, pp. 413–439, Aug. 2010.

[100] S. M. Shariff, X. Zhang, and M. Sanderson, "On the credibility perception of news on Twitter: Readers, topics and features," *Comput. Hum. Behav.*, vol. 75, pp. 785–796, Oct. 2017.

[101] M. Button, C. M. Nicholls, J. Kerr, and R. Owen, "Online frauds: Learning from victims why they fall for these scams," *Austral. New Zealand J. Criminol.*, vol. 47, no. 3, pp. 391–408, Dec. 2014.

[102] D. Wall, *Cybercrime: The Transformation of Crime in the Information Age*, vol. 4. Cambridge, U.K.: Polity Press, 2007.

[103] M. Hu, S. Liu, F. Wei, Y. Wu, J. Stasko, and K.-L. Ma, "Breaking news on Twitter," in *Proc. ACM Annu. Conf. Hum. Factors Comput. Syst. (CHI)*, 2012, pp. 2751–2754.

[104] R. H. Knapp, "A psychology of rumor," *Public Opinion Quart.*, vol. 8, no. 1, pp. 22–37, 1944.

[105] S. C. Pendleton, "Rumor research revisited and expanded," *Lang. Commun.*, vol. 18, no. 1, pp. 69–86, Jan. 1998.

[106] V. Qazvinian, E. Rosengren, D. R. Radev, and Q. Mei, "Rumor has it: Identifying misinformation in microblogs," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2011, pp. 1589–1599.

[107] S. M. Shariff and X. Zhang, "A survey on deceptions in online social networks," in *Proc. Int. Conf. Comput. Inf. Sci. (ICCOINS)*, Jun. 2014, pp. 1–6.

[108] M. Huber, M. Mulazzani, and E. Weippl, "Who on earth is 'Mr. Cypher': Automated friend injection attacks on social networking sites," in *Proc. IFIP Int. Inf. Secur. Conf.* Berlin, Germany: Springer, 2010, pp. 80–89.

[109] R. Zheng, Y. Qin, Z. Huang, and H. Chen, "Authorship analysis in cybercrime investigation," in *Proc. Int. Conf. Intell. Secur. Inform.* Berlin, Germany: Springer, 2003, pp. 59–73.

[110] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," *J. Amer. Soc. Inf. Sci. Technol.*, vol. 57, no. 3, pp. 378–393, Feb. 2006.

[111] R. Marukatat, R. Somkiadcharoen, R. Nalintasnai, and T. Aramboonpong, "Authorship attribution analysis of thai online messages," in *Proc. Int. Conf. Inf. Sci. Appl. (ICISA)*, May 2014, pp. 1–4.

[112] L. Pearl and M. Steyvers, "Detecting authorship deception: A supervised machine learning approach using author writeprints," *Literary Linguistic Comput.*, vol. 27, no. 2, pp. 183–196, Jun. 2012.

[113] W. Xiujuan, Z. Chenxi, Z. Kangfeng, T. Haoyang, and T. Yuanrui, "Detecting spear-phishing emails based on authentication," in *Proc. IEEE 4th Int. Conf. Comput. Commun. Syst. (ICCCS)*, Feb. 2019, pp. 450–456.

[114] H. Ramnial, S. Panchoo, and S. Pudaruth, "Authorship attribution using stylometry and machine learning techniques," in *Intelligent Systems Technologies and Applications*. Cham, Switzerland: Springer, 2016, pp. 113–125.

[115] U. Stanczyk and K. A. Cyran, "Application of artificial neural networks to stylometric analysis," in *Proc. 8th Conf. Syst. Sci. Comput.* Singapore: World Scientific, 2008, pp. 25–30.

[116] Z. Bu, Z. Xia, and J. Wang, "A sock puppet detection algorithm on virtual spaces," *Knowl.-Based Syst.*, vol. 37, pp. 366–377, Jan. 2013.

[117] A. Usha and S. M. Thampi, "Authorship analysis of social media contents using tone and personality features," in *Proc. Int. Conf. Secur., Privacy Anonymity Comput., Commun. Storage*. Cham, Switzerland: Springer, 2017, pp. 212–228.

[118] V. L. Rubin, Y. Chen, and N. J. Conroy, "Deception detection for news: Three types of fakes," in *Proc. 78th ASIS T Annu. Meeting, Inf. Sci. Impact, Res. Community*, 2015, p. 83.

[119] M. Chawla and S. Singh Chouhan, "A survey of phishing attack techniques," *Int. J. Comput. Appl.*, vol. 93, no. 3, pp. 32–35, 2014.

[120] Q. Ma, "The process and characteristics of phishing attacks-a small international trading company case study," *J. Technol. Res.*, vol. 4, p. 1, Jul. 2013.

[121] S. Afroz and R. Greenstadt, "Phishzoo: An automated Web phishing detection approach based on profiling and fuzzy matching," in *Proc. 5th IEEE Int. Conf. Semantic Comput. (ICSC)*, Sep. 2009, pp. 1–11.

[122] A. Aggarwal, A. Rajadesingan, and P. Kumaraguru, "PhishAri: Automatic realtime phishing detection on Twitter," in *Proc. eCrime Researchers Summit*, Oct. 2012, pp. 1–12.

[123] Y. Han and Y. Shen, "Accurate spear phishing campaign attribution and early detection," in *Proc. 31st Annu. ACM Symp. Appl. Comput. (SAC)*, 2016, pp. 2079–2086.

[124] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proc. 50th Annu. Meeting Assoc. Comput. Linguistics*, 2012, pp. 171–175.

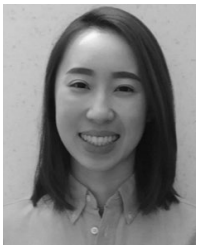
[125] D. H. Fusilier, M. Montes-y Gómez, P. Rosso, and R. G. Cabrera, "Detection of opinion spam with character n-grams," in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*. Cham, Switzerland: Springer, 2015, pp. 285–294.

[126] S. Banerjee and A. Y. K. Chua, "Authentic versus fictitious online reviews: A textual analysis across luxury, budget, and mid-range hotels," *J. Inf. Sci.*, vol. 43, no. 1, pp. 122–134, Feb. 2017.

[127] S. Shojaee, M. A. A. Murad, A. B. Azman, N. M. Sharef, and S. Nadali, "Detecting deceptive reviews using lexical and syntactic features," in *Proc. 13th Int. Conf. Intellient Syst. Design Appl.*, Dec. 2013, pp. 53–58.

[128] X. Wang, X. Zhang, C. Jiang, and H. Liu, "Identification of fake reviews using semantic and behavioral features," in *Proc. 4th Int. Conf. Inf. Manage. (ICIM)*, May 2018, pp. 115–119.

[129] M. Potthast, J. Kiesel, K. Reinartz, J. Bevendorff, and B. Stein, "A stylometric inquiry into hyperpartisan and fake news," 2017, *arXiv:1702.05638*. [Online]. Available: http://arxiv.org/abs/1702.05638
[130] P. Dewan, A. Kashyap, and P. Kumaraguru, "Analyzing social and stylometric features to identify spear phishing emails," in *Proc. APWG Symp. Electron. Crime Res. (eCrime)*, Sep. 2014, pp. 1–13.
[131] V. Pérez-Rosas and R. Mihalcea, "Cross-cultural deception detection," in *Proc. 52nd Annu. Meeting Assoc. Comput. Linguistics*, vol. 2, 2014, pp. 440–445.
[132] D. Pisarevskaya, "Deception detection in news reports in the russian language: Lexics and discourse," in *Proc. EMNLP Workshop, Natural Lang. Process. Meets Journalism*, 2017, pp. 74–79.
[133] M. Perc, M. Ozer, and J. Hojnik, "Social and juristic challenges of artificial intelligence," *Palgrave Commun.*, vol. 5, no. 1, pp. 1–7, Dec. 2019.
[134] (2020). *What is Morphology*. Accessed: Feb. 15, 2020. [Online]. Available: https://findwords.info/term/morphology
[135] L. K. Tyler, P. de Mornay-Davies, R. Anokhina, C. Longworth, B. Randall, and W. D. Marslen-Wilson, "Dissociations in processing past tense morphology: Neuropathology and behavioral studies," *J. Cognit. Neurosci.*, vol. 14, no. 1, pp. 79–94, Jan. 2002.
[136] M. Honnibal and M. Johnson, "An improved non-monotonic transition system for dependency parsing," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 1373–1378.
[137] C. B. Vista, C. H. Satriawan, D. P. Lestari, and D. H. Widyantoro, "Specific acoustic models for spontaneous and dictated style in indonesian speech recognition," *J. Phys., Conf. Ser.*, vol. 978, Mar. 2018, Art. no. 012059.
[138] A. Sadhwani and N. Saxena, "A new approach to ranking algorithm-custom personalized searching," in *Proc. 2nd Int. Conf. Comput. Sustain. Global Develop. (INDIACom)*, 2015, pp. 130–133.
[139] S. Chawla, "Personalised Web search using trust based hubs and authorities," *Int. J. Eng. Res. Appl.*, vol. 4, no. 7, pp. 157–170, 2014.
[140] Y. Wang, D. Yin, L. Jie, P. Wang, M. Yamada, Y. Chang, and Q. Mei, "Beyond ranking: Optimizing whole-page presentation," in *Proc. 9th ACM Int. Conf. Web Search Data Mining (WSDM)*, New York, NY, USA: ACM, 2016, pp. 103–112.
[141] I. Velásquez, A. Caro, and A. Rodríguez, "Authentication schemes and methods: A systematic literature review," *Inf. Softw. Technol.*, vol. 94, pp. 30–37, Feb. 2018.
[142] J. Wayman, A. Jain, D. Maltoni, and D. Maio, "An introduction to biometric authentication systems," in *Biometric Systems* London, U.K.: Springer, 2005, pp. 1–20.
[143] M. L. Brocardo, I. Traore, and I. Woungang, *Continuous Authentication Using Writing Style*. Cham, Switzerland: Springer, 2019, pp. 211–232.
[144] V. S. Suneet Kumar, "Differentiation of handedness of writer based on their strokes and characteristic features," *J. Forensic Res.*, vol. 4, no. 5, pp. 1–3, 2013.
[145] R. Tan. (2017). *Manglish Getting More Mangled*. Accessed: May 31, 2019. [Online]. Available: https://www.thestar.com.my/news/
[146] A. Hashim, "Crosslinguistic influence in the written English of Malay undergraduates," *J. Mod. Lang.*, vol. 12, no. 1, pp. 60–76, 2017.



KAH YEE TAI received the B.Sc. degree in computer science and the M.Sc. degree in business information systems from the University of Monash Malaysia, in 2015 and 2017, respectively. She is currently pursuing the M.Sc. degree with Monash University Malaysia. Her research interests include big data and machine learning.

JASBIR DHALIWAL received the B.Sc. degree in computer science (Hons.) from the University of Malaya, in 2004, and the M.Sc. degree in applied science and the Ph.D. degree from the Royal Melbourne Institute of Technology (RMIT) University, Australia, in 2008 and 2013, respectively. She is currently a Lecturer with Monash University Malaysia. In between, she worked as a Software Engineer at Motorola, Malaysia, Postdoctoral Researcher with IBM Research Australia, and as a Data Scientist at FTI Consulting Australia. Her research interests include big data, machine learning, and string related algorithms.

SHAFIZA MOHD SHARIFF is currently pursuing the Ph.D. degree from RMIT University, Australia and a Lecturer with Universiti Kuala Lumpur, Malaysia, focusing on credibility perception and human information behaviour. Other topics of interest include the application of cybersecurity and on deception detection.