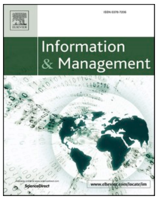
Information & Management 62 (2025) 104153

Contents lists available at ScienceDirect

Information & Management

journal homepage: www.elsevier.com/locate/im



Stylometric characteristics of code-switched offensive language in social media

Lina Zhou [a,*], Zhe Fu [b]

[a] Department of Business Information Systems and Operations Management, The University of North Carolina at Charlotte, Charlotte, NC, USA
[b] Department of Software and Information Systems, The University of North Carolina at Charlotte, Charlotte, NC, USA

ARTICLE INFO

Keywords:
Offensive language
Code-switching
Stylometric characteristics
Social media
Detection model

ABSTRACT

Offensive language is a significant detriment to social media environments. Existing research predominantly assumes monolingual expression, overlooking the prevalent behavior of code-switching (CS). To address this critical knowledge gap, this study identifies and empirically validates the distinct stylometric characteristics of code-switched (CSed) offensive language. Additionally, we developed methods to construct the first social media dataset specifically for CSed offensive content. Our analysis of this dataset reveals that CSed offensive language exhibits unique stylometric characteristics; moreover, these characteristics vary between the language segments involved in the CS. Furthermore, incorporating these features significantly enhances the performance of offensive language detection models. These findings offer significant research and practical implications for social media researchers, platforms, moderators, and users.

1. Introduction

While social media platforms are designed to facilitate communication, information sharing, and community building, they are increasingly plagued by offensive language. Recent data highlight the severity of this issue: nearly half (46%) of teens aged 13-17 have experienced online bullying or harassment [1], with Instagram, Facebook, and Snapchat identified as the most common social media platforms for such abuse [2]. In addition, a 2021 report [3] indicates that 41% of American adults have personally encountered online harassment, with an alarming trend since 2017 toward more severe forms, including physical threats, stalking, and sustained harassment. This rise in the use of offensive language was particularly noticeable during the pandemic [4,5].

Offensive language, specifically hate speech, personal attacks, and cyberbullying represents a critical concern for online safety [6]. This issue is increasingly prevalent due to the surge in user-generated content, especially on social media platforms [7]. Cyberbullying, for example, is now recognized as a major public health issue [8] with well-documented detrimental effects. These effects range from behavioral and health-related problems [9] to significant psychological and physiological harm, including panic, anxiety, depression, and even self-harm or suicide [10–12]. Despite growing awareness, a 2018 survey revealed that 66% of youth believed social media platforms are failing to adequately address online harassment and bullying [13]. Consequently, a deeper, more nuanced understanding of offensive language is imperative to develop effective detection methods and safeguard online communities.

Code switching (CS) describes the act or phenomenon of switching from one language to another during the same communicative event [14,15]. CS has become increasingly common in social media [16,17], especially in bilingual communities. The U.S. has experienced significant growth in its foreign-born population over the past half-century [18], reaching an estimated 13.9 percent of the total population in 2022. This demographic shift is accompanied by a rapid increase in bilingualism [19], and a rising trend of CS among second-generation immigrants [20]. For instance, a study reports that over 80% of students and educators engage in this practice on social networks [21]. However, existing work predominantly focuses on spoken communication [22,23] or the general presence of CS online [24,25]. Consequently, a significant gap remains in our understanding of nuanced characteristics of CS within specific online contexts [26] such as offensive language. While CS is known to serve various sociolinguistic functions in both spoken and written discourses such as conveying social identity, expressing solidarity with a particular group, expressing emotions, or emphasizing a particular point [17–27], the interplay of external and internal factors

* Corresponding author.
E-mail addresses: lzhou8@charlotte.edu (L. Zhou), zfu2@charlotte.edu (Z. Fu).

https://doi.org/10.1016/j.im.2025.104153
Received 14 November 2023; Received in revised form 6 April 2025; Accepted 7 April 2025
Available online 24 April 2025

influencing CS behavior requires further investigation [28]. Importantly, there is a lack of empirical research examining CS behavior in the context of offensive language on social media platforms.

Social media platforms increasingly employ machine learning techniques to develop automated moderation systems for offensive language. However, these systems predominantly rely on supervised classification models (e.g., [29,30]]), which are trained on limited labeled datasets, hindering their ability to identify previously unseen offensive content, particularly “unstandard” language (e.g., CS) [31,32]. This limitation can be actively exploited by users who employ adversarial techniques like CS to intentionally modify offensive language and bypass algorithmic moderation filters, as in the case of “Algospeak” [33,34]. For instance, in response to platform moderation of antisemitic language (framed by related legislation in the United States), users have utilized CS, such as replacing “Jewish” with the Chinese term “鱿鱼” (squid). This practice effectively obfuscates offensive intent, hindering detection by automated systems and potentially human moderators or general users. This underscores a critical need to explore and identify novel features capable of more effectively detecting offensive language within code-switched (CSed) contexts. Therefore, understanding the role of CS in online offensive language can have both theoretical significance and practical values to inform the design of effective intervention strategies or measures.

While a substantial body of research has explored offensive language detection by developing machine-learning or deep learning models [35–38]], these studies typically assume that offensive language is monolingual. However, offensive language can manifest in CSed contexts, as illustrated in Table 5, where CSed refers to a specific instance or occurrence of CS. Compared to monolingual offensive language, effective CSed detection faces several notable challenges. First, CS introduces significant contextual complexity. Offensive language detection models often rely heavily on contextual cues to discern offensiveness. In CSed content, the mixing of multiple languages, within and across sentences, disrupts monolingual grammatical and structural coherence. This linguistic hybridization can obscure the contextual signals and structural patterns typically found in monolingual text, thereby complicating the analysis. Second, semantic nuances and cross-lingual differences pose difficulties. Offensive language frequently employs subtle nuances, such as sarcasm, innuendo, and culturally-specific expressions. In CSed contexts, these nuances may be conveyed differently depending on the language combination, complicating the identification of offensive patterns. For instance, while the English term “bitch” may be directly used to denote a dissolute and scheming individual, a CSed context alternating between English and Chinese might employ the Chinese term “绿茶” (green tea) to convey a similar offensive connotation. Third, the availability of annotated datasets for developing models for CSed offensive language detection is severely limited or non-existent, contrasting sharply with the abundance of resources for monolingual tasks (e.g., Hatespeech [39], Hateval [40], Toxkaggle [41]). Moreover, the inherent complexities of CS render data collection and annotation substantially more difficult and resource-intensive compared to monolingual data.

The strategy of translating the CSed language into monolingual equivalents presents three critical limitations, as partially recognized by prior work [42]. First, the fidelity of the resulting monolingual text is contingent upon the translation models’ efficacy. Variations in translator performance, particularly across diverse language pairs, introduce inconsistencies and potential inaccuracies in the translated output. Second, the linguistic proficiency of multilingual social media users in each constituent language within CSed data often varies. This variability manifests as non-standard grammatical structures, orthographic errors, and idiosyncratic language usage, all of which pose substantial challenges for accurate translation. Third, semantic shifts during translation can inadvertently neutralize or alter the offensiveness of certain terms. For example, the Chinese term “支那”, widely regarded as a highly derogatory reference to China or the Chinese people, is frequently rendered as the neutral English term "China" by many translation systems (e.g., Google Translate, Baidu AI). This loss of semantic nuance underscores the inherent difficulty in preserving the offensive valence of language during translation. Consequently, a significant knowledge gap persists regarding the intricacies of CS behavior in online offensive language, particularly concerning the stylistic and linguistic features of CSed offensive language.

To address the identified knowledge gaps, this study investigates the following interconnected research questions: 1) How to characterize the linguistic style of CSed social media text? 2) What are the unique stylometric characteristics of CSed offensive language? 3) Are there cross-language differences in the stylometric features of CSed offensive language?

By answering the proposed research questions regarding offensive language in CSed social media text, this study makes distinct contributions to the related fields. First, to our knowledge, this is the first research to examine CS contexts in the analysis of online offensive language. Second, it extends the scope of stylometry analysis to CSed discourse by introducing novel stylometric features tailored to the unique characteristics of CSed text, such as switch type, switch location, and content similarity. Third, this research develops a comprehensive methodology for the simultaneous analysis of CSed text across constituent languages, integrating language identification, deep learning, and other natural language processing techniques. This advancement goes beyond traditional sociolinguistic approaches, which are often manual and lack scalability, and overcomes the limitations of existing stylometric methods that do not account for CSed data. Fourth, it identifies distinct stylometric characteristics of CSed offensive language for the first time, and further validates these findings through robustness analysis tasks, thereby addressing a critical gap in the literature. These findings offer valuable insights for enhancing existing offensive language detection models without requiring the development of entirely new frameworks. Fifth, this study sheds light on the interlingual stylometric differences within CSed offensive text, contributing to a deeper understanding of linguistic nuances associated with offensive language. Last but not least, it introduces a two-stage approach for constructing a CSed offensive language dataset from social media, directly addressing the data scarcity in this domain and advancing related studies on CSed social media discourse.

2. Background and Related Work

This section provides background on offensive language and CS, and reviews related work in online offensive language detection and stylometry analysis.

2.1. Online Offensive Language

Offensive language, defined as “hurtful, derogatory or obscene comments made by one person to another person” [7], manifests in various forms, including toxic comments, aggressive content, cyberbullying, and hate speech [43]. Online offensive language is the dissemination of harmful, abusive, or derogatory language through some form of digital media or electronic communication technologies. The inherent anonymity or weak identity verification [44,45], coupled with the rapid dissemination capabilities of social media platforms [46], amplify the propagation of offensive language. Its detrimental effects range from individual harm, such as insult, derogation, harassment, hatred, and humiliation, with potential consequences for physical and mental well-being, and may even contribute to societal violence or instability [46].

2.2. Code-switching in Offensive Language

Code-switching (CS) is usually used to describe the literary style of using two different languages or language varieties in the context of a single conversation or written text [47,48]. While the precise delineation of CS remains a subject of linguistic inquiry, it is widely understood to involve the process of alternating between two or more languages, which can occur at either the sentence (inter-sentential) or sub-sentence (intra-sentential) level [49]. Inter-sentential CS involves switching languages at sentence boundaries, whereas intra-sentential CS occurs within a single sentence [50,51]. According to a study by Astani et al. [52], intra-sentential CS is the most frequent type on YouTube.

The intersection of offensive language and CS remains a largely unexplored territory. To the best of our knowledge, the only study aligned with the scope of this research is Agarwal et al. [56]. This study explored CS between Hindi and English among Indian users in tweets containing swear words, analyzing the correlations between the use of such tweets and user demographic characteristics, including gender (inferred from usernames), topics (manually grouped), and language preferences (based on the frequency of Hindi and English swear words in tweets). However, this study did not explore the stylometric features of CSed language, a central focus of our current research. Similarly, while Yadav et al. [53] addressed multilingual hate speech detection using deep learning models, such as CNN, LSTM, BiLSTM with word embeddings, they did not investigate CSed stylometric features. Furthermore, it is crucial to distinguish between swearing, hate speech, and offensive language. Swearing, while often offensive, can occur in non-offensive contexts [54,55]. Hate speech, though a form of offensive language, is distinguished by its intent to discriminate, provoke hatred, or encourage violence against target groups. Therefore, these studies, while relevant, do not directly address the nuanced complexities of CSed offensive language.

A significant challenge in this field is the scarcity of annotated CSed datasets. Yadav et al. [53] either developed nor utilized CSed datasets. Agarwal et al. [56] first identified CSed English-Hindi tweets based on word-level language labels from Gella et al. [57], and then employed a lexicon-based approach for swearing classification. However, this method presents notable limitations. Specifically, it struggles with contextual ambiguities inherent in lexicon-based analysis, such as the interpretation of words like “pig” or “die”). Consequently, there is a clear need for a more robust and context-aware approach for detecting CSed offensive language on social media.

2.3. Online Offensive Language Detection

To mitigate the dissemination of offensive language online, many social media platforms (e.g., X, Facebook, YouTube) have integrated automated offensive language detection technologies. Existing methods for offensive language detection fall into two main categories: lexicon-based and machine learning-based approaches. Lexicon-based methods identify offensive language by comparing textual content against a precompiled lexicon of offensive words or phrases. For example, Rizwan et al. [58] compiled a lexicon of hateful words through online keyword searches and interviews, which consists of abusive and derogatory terms as well as slurs or terms pertaining to religious hate and sexist language. Machová et al. [59] developed a Slovakian lexicon of toxic words by annotating the toxicity levels within an existing list of negative words, subsequently applying it to detect offensive content on Facebook. Albania et al. [60] adopted a two-step process to create a Tamazight offensive language lexicon: they first invited two native speakers and Facebook administrators to identify seed words frequently used by online offenders and abusers; and then asked these participants to produce orthographic variations of the identified seed words.

Machine learning-based methods for offensive language detection typically rely on text-based features and classification techniques. The textual features include character n-grams, sentiments, sentence representation, BoW, stylistic features, emotion words, most common word bigrams, TF-IDF, Word2Vec, and word embeddings [61–65]. For example, Chen et al. [6] first analyzed the grammatical structures of sentences using a natural language parser before selecting word sets as features for offensiveness detection. To detect offensive content on Twitter, Davidson et al. [66] leveraged three types of text-based features, including Flesch-Kincaid Grade Level and Flesch Reading Ease scores, sentiment scores, and other linguistic features, such as the number of characters, words, and syllables in each tweet. A variety of machine learning and deep learning techniques have been used to build offensive language detection models, including SVM, logistic regression, Naïve Bayes, J48, RNN, CNN, BiLSTM, BiGRU, and transformer models, such as BERT, mBERT, and XLM-RoBERTa [35,37,38,66–72]. It is worth highlighting that the aforementioned detection methods address monolingual offensive language. There are emerging studies on offensive language detection from multilingual text [73–75]. Nevertheless, the focus of those studies was on improving offensive content detection by leveraging state-of-the-art transformer models but not on characterizing the features of offensive language. More importantly, their analyses were conducted in different languages separately, while overlooking the CS behavior.

2.4. Stylometry Analysis and Features

Stylometry analysis provides a quantifiable evaluation of the distinctive qualities of a text [76]. The characterization of stylometry features occurs at multiple linguistic levels, including lexical, syntactic, semantic, and structural [76,77]. Lexical features, the most fundamental form of feature representation, focus on word-level characteristics. By quantifying the frequency of lexical item occurrences, these features reflect the vocabulary richness inherent in texts produced by specific authors or author groups [78–80]. Syntactic features capture sentence-level structural patterns, reflecting an author’s unique writing style. For instance, Markov et al. [81] leveraged part-of-speech features to capture the morpho-syntactic patterns, while Belvisi et al. [82] extracted syntactic features, such as punctuation and function word usage, for authorship analysis. Unlike syntactic features, semantic features address the interpretation and meanings of words, phrases, and sentences. For example, Clark and Hannon [83] employed synonym-based features, including synonym count and total occurrences, for author recognition. Finally, structural features represent document-level organization, capturing the macro-level arrangement of extended texts.

Despite the established history of stylometry research, existing analyses predominantly focus on monolingual texts. The study most relevant to this research focused on cross-lingual authorship identification using data from online book archives [84]. In that study, documents were first partitioned into smaller equal-length fragments (30,000 tokens), which were then further segmented into chunks of 1,500 tokens each. From each of these chunks, 16 language-independent features (e.g., average number of words per sentence, number of sentences, and frequency of punctuations) were extracted, and the authorship predictions from all fragments were combined by computing the average probability mass function to determine the overall authorship of the text. This approach demonstrated strong performance in cross-lingual authorship identification; however, it was limited to languages within the Indo-European family (e.g. English, French, Spanish, Portuguese, German), which likely share certain stylometric traits. Whether these stylometric features extend to languages from distinct families, such as the Sino-Tibetan family (e.g., Chinese), remains uncertain. In addition, the study did not explore offensive language in cross-lingual social media contexts, highlighting a significant gap in the literature. Most critically, the objective of the research [84] was to determine whether documents in one language could be used to accurately identify the authorship of a document in another language by the same author. The cross-language framework differs fundamentally from the dynamics of CS, rendering the characterization and extraction of stylometric features of CSed text a nascent area of inquiry. This study seeks to address these identified knowledge gaps by introducing, extracting, and evaluating the stylometric characteristics of CSed offensive language in social media.

3. Stylometric Characteristics of Online Code-switched Text

Our literature review indicates that previous research on offensive language has largely overlooked CS. The stylometric characteristics of online CS, particularly in the context of offensive language, remain unexplored. Against the backdrop of growing public health concerns surrounding offensive language and the increasing prevalence of CS on social media, there are compelling theoretical and practical motivations to examine the role of CS in offensive language within these platforms.

In this section, we address the first research question by introducing a set of stylometric features to characterize CSed text in social media. These features build on, enrich, extend, and integrate existing literature on CS and stylometry, as reviewed in Section 2. We focus specifically on lexical, syntactic, and semantic features, excluding structural features due to their limited relevance to the informal and concise nature of social media text. Moreover, we categorize the selected stylometric features into two main groups: overall features and language-specific features (see Table 1). While language translation could convert CSed text into monolingual text, such as English, it often fails to preserve the nuanced, context-dependent aspects of CS and may introduce errors or inconsistencies due to linguistic ambiguity and differences. This study directly addresses the complexities of CS rather than circumventing them through translation.

3.1. Overall Features

Overall stylometric features characterize the CSed text in its entirety, reflecting the text's holistic linguistic properties. Specifically, we expand existing stylometric feature categories (see Section 2.4) by introducing novel features designed for the analysis of CS, addressing a notable gap in the current literature. We also propose new semantic features (e.g., content similarity) and lexical features for analyzing CSed social media text (e.g., URLs). Additionally, we adapt and extend traditional lexical and syntactic stylometric features to the context of CS.

3.1.1. CS features

To address the unique characteristics of CS in social media, we develop and employ novel stylometric features and illustrate them with Chinese-English CS data.

Switch type. There are two main types of CS (see Section 2): *inter-switch* (inter-sentential *CS*), where language changes between sentences, as in "My mama bought me a new face mask. It's so cute! 😍 谢谢, 妈! 我爱你!" (My mama bought me a new face mask. It's so cute! 😍 Thank you, mom! I love you!), and *intra-switch* (*intra*-sentential *CS*), where language changes within sentences, such as "The COVID is 恐怖." (The COVID is horrendous). In this study, we first measured the frequency of each CS type and then normalized the *inter-switch* frequency by sentence count, and that of *intra-switch* frequency by word count.

Table 1
Stylometric Characteristics of Online CSed Text.

Category	Sub-category	Specific features
Overall features	CS features	Switch type
		Switch location
	Lexical features	URLs
		Lexical diversity
		Exclamation mark
	Syntactic features	Pausality
	Semantic features	Content similarity
Language-specific features	Complexity	Readability
		Segment length
	Expressivity	Emotion words

Switch location. We define *switch location* as the position within a social media discourse where CS occurs. Prior research has shown that bilinguals tend to *exhibit* consistent *switch locations* when describing visual stimuli [85]. Following the work of Calvillo et al. [86], we categorize switch location into three types: *beginning, middle,* and *end*, corresponding to the first 10%, middle 80%, and last 10% of each sentence, respectively. For each switch location, we computed the frequency and then normalized it by sentence count.

3.1.2. Lexical features

URLs. It is calculated as *the* count of URLs within a text [36], normalized by the total number of words in that text.

Lexical diversity. It is *defined* as the ratio of the number of unique words or terms to the total number of words [87].

Exclamation mark. We use the ratio of exclamation mark count to the total word count in a text as the measure, as exclamation marks are commonly *associated* with emotional comments [88].

3.1.3. Syntactic features

Pausality. Punctuation marks segments sentences into short units, potentially reducing sentence complexity and emphasizing textual intent [89]. We select two pausality measures. *Sentence pausality* is defined as the ratio of punctuation mark count to sentence count [87]. Given the often ambiguous sentence boundaries in social media texts, we introduce *word pausality*, the ratio of punctuation mark count to word count, as an alternative complexity measure.

Among the above lexical and syntactic features, all serve as indicators of complexity except for exclamation mark, which reflects expressivity.

3.1.4. Semantic features

Content similarity. It assesses the semantic overlap between language segments within a CSed text. We used cosine similarity based on word vector encodings of the two languages. CS can serve to reiterate or emphasize messages for enhanced clarity or opinion expression [90]. To facilitate comparison, we translated text segments from one language into the other prior to calculating similarity.

3.2. Language-specific Features

The language-specific features analyze individual language segments within the CSed text, identifying language-specific characteristics. Importantly, these features are derived directly from the original languages used in CSed text without translation, enabling authentic cross-language comparisons. To ensure generalizability, we selected features applicable across diverse languages. Consequently, our focus was primarily on lexical features because semantic features are defined based on the entire CSed text (see Section 3.1). We further categorized lexical features into two sub-groups: complexity and expressivity.

3.2.1. Complexity

Word length. It is a common measure of monolingual text complexity [87]. We extend word length to analyze language segments for individual languages within CSed text, defining it as the average number of characters per word in those segments.

Readability. As a direct and common measures of text complexity, readability indicates the ease with which a text is understood by readers [91]. In this study, we employed the Flesch-Kincaid Grade

Level [92] for English text segments and the Chinese readability index (the average version) [93] for Chinese segments in CSed text. Lower scores indicate easier text, while higher scores suggest greater difficulty for both measures. As these indices rely on distinct linguistic features—syllables and sentence length for English, and clause length and function word proportion for Chinese—their value ranges differ. To enable cross-language segment comparisons of text difficulty, we normalized the raw scores of both indices into z-scores, reflecting relative difficulty within each language rather than absolute values.

Segment length. It is defined as the average word count within language-specific text segments, which extends the traditional measures of word count and sentence length from monolingual text [87],[94] to the analysis of CSed text.

3.2.2. *Expressivity*

Emotion words. Based on the premise that offensive language expressions are often associated with negative emotions, we analyzed emotion words within language-specific segments across various levels of granularity: *overall emotion, negative emotion*, and specific negative emotions, such as *sadness, anger*, and *anxiety*. We measured these different emotion features for their respective language segments using Linguistic Inquiry and Word Count [95].

4. Dataset Construction Methods

As discussed earlier, one primary challenge in CSed offensive language research is the lack of datasets. To address this gap, we developed a dataset construction methodology. We illustrate the proposed methods using Chinese-English CS as a case study. Importantly, our approach can be extended to analyze offensive language involving CS between other languages.

Our dataset construction method consists of two main components: 1) CS dataset collection and 2) CSed offensive language identification.

4.1. *Code-switching Dataset Collection*

Previous studies suggest that social media platforms, such as Twitter/X, Facebook, and Weibo, are ideal venues for collecting CS data [56],[96]. We chose Twitter as our data source for three main reasons. First, as previously mentioned, Twitter is a common platform for studying CS behavior. Second, it supports a total of 34 languages for its widgets and buttons[1], including English and Chinese, which facilitates CS expressions. Third, at the time of our data collection, the platform allowed developers to use its API for free, and additionally, provided a streaming API service that delivers new, relevant tweets as they occur. Furthermore, this study focuses on tweets related to COVID-19 due to several key considerations. The pandemic accelerates and intensifies the interactions between different social groups, including people with diverse cultural backgrounds, which generates opportunities for language change by borrowing terms from another language [97]. Given the extensive and profound impact of the pandemic, people faced increasing financial, psychological, and emotional stress, which can potentially fuel the use of offensive language. Based on the temporal analysis results of a large-scale Twitter dataset, a very recent study indicates that various COVID-19 pandemic events led to an increase in offensive speech [4]. The study demonstrates an increase in offensive tweets with abusive language targeting various individuals or groups over a short period, coinciding with worsening pandemic and the use of disparaging language by prominent politicians. For example, numerous posts involving anti-Asian racism and xenophobia have appeared on major social media sites such as Twitter/X during pandemic [5],[98], and these anti-Asian sentiments had an effect on consumer discrimination against businesses associated with Asian Americans [99].

Fig. 1 outlines the process of CS data collection, which consists of three stages: initial data collection and filtering, candidate CS identification, and candidate CS cleaning.

Initial data collection and filtering. Our identification and extraction of tweets in the selected languages, specifically English and Chinese, leveraged the language codes embedded in the tweet metadata. We then filtered the data by removing retweets, identical tweets, and tweets with less than 20 characters. Finally, we collected a total of 157,146,871 tweets for further processing.

Candidate CS identification. We adapted language identification techniques to detect Chinese characters within tweets initially coded as English, and conversely, English characters within tweets coded as Chinese. Tweets exhibiting such cross-linguistic features would be considered candidate CS.

Candidate CS cleaning. Candidate CSed tweets deemed trivial or representing common usage were excluded. This filtering process was informed by *patterns* identified through extensive manual analysis of the candidate CS dataset. Examples of trivial CS instances included single-word CS involving proper nouns, such as personal names (e.g., Fauci) and COVID-19 related terms (e.g., CDC, COVID-19, Pfizer), and platform reserved terms (e.g., retweet, RT, 转发). To compile a list of COVID-related terms, we first extracted named entities from COVID-related Wikipedia pages using the NLTK package. These entities subsequently went through manual screening, resulting in 447 terms, encompassing person names, organizations, locations, vaccines, COVID variants, and related diseases. We further eliminated tweets containing repetitive or nearly identical content across different contituent languages using the following procedure.

We first segmented each tweet T_k by language into two distinct groups: segments T_k^{cn} containing only Chinese words and segments T_k^{en} containing only English words. Next, we translated the Chinese segments T_k^{cn} into English $T_k^{cn\rightarrow en}$ using the Google Translate API. We then encoded words from both the original English text segment T_k^{en} and the translated English text $T_k^{cn\rightarrow en}$ into vector embeddings. With various embedding models exist for text representation, concerns have emerged regarding the non-smooth anisotropic distribution in the transformer-based language models [100], such as BERT [101], RoBERTa [102], and Llama-3 [103]. Notably, the TF-IDF algorithm does not suffer from this limitation. Other embedding models like GloVe [104] and

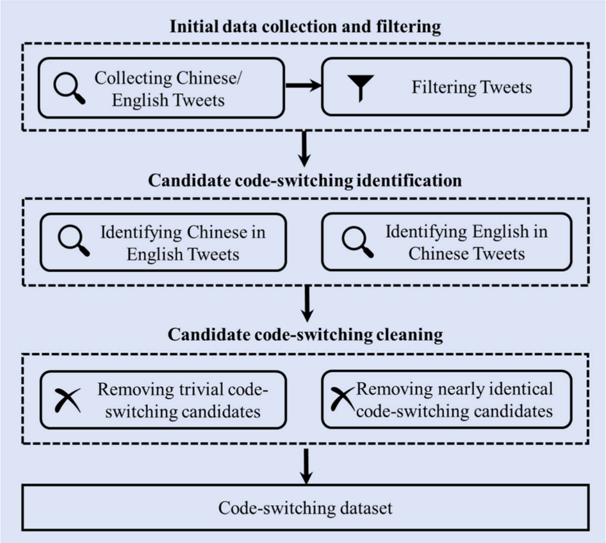

Initial data collection and filtering
Collecting Chinese/ English Tweets
Filtering Tweets
Candidate code-switching identification
Identifying Chinese in English Tweets
Identifying English in Chinese Tweets
Candidate code-switching cleaning
Removing trivial code-switching candidates
Removing nearly identical code-switching candidates
Code-switching dataset

Fig. 1. Code-switching Dataset Collection.

[1] Supported languages and browsers | Docs | Twitter Developer Platform

skip-gram [105] fail to provide a unified word-level representation across different language segments. Additionally, aggregating word embeddings within language segments can result in the loss of fine-grained information and contextual nuances. Therefore, we employed the TF-IDF algorithm as our text encoder for the Chinese and English segments, enabling identification of nearly identical content. The embedding vectors of the original English segments $T_k^{en}$and the English segments translated from Chinese counterparts $T_k^{cn\rightarrow en}$ are represented in equations (1)-(3), respectively.

$$e_k^{en} = \left[tf_{1k}^{en}, tf_{2k}^{en}, \dots, tf_{lk}^{en}\right] \tag{1}$$

$$e_k^{ch\rightarrow en} = \left[tf_{1k}^{ch\rightarrow en}, tf_{2k}^{ch\rightarrow en}, \dots, tf_{lk}^{ch\rightarrow en}\right] \tag{2}$$

$$tf_{ik} = freq(w_i, T_k) \cdot log\left(\frac{|T|}{|\{k : w_i \in T_k\}|}\right) \tag{3}$$

where tf_{ik} is the TF-IDF score of word w_i in tweet T_k, $freq(w_i, T_k)$ denotes the occurrence frequency of word w_i in T_k, l is the number of features, $|\{k : w_i \in T_k\}|$ denotes the number of tweets containing word w_i, $T = \{T_1, T_2, \dots, T_m\}$ is a set of tweets, and $|T|$ is the total number of tweets in the dataset.

To avoid extremely high-dimensional vectors that are noise-prone, we considered only the top l features, sorted in descending order of frequency across all tweets in the embedding representations. We measure the semantic similarity between T_k^{en} and $T_k^{cn\rightarrow en}$by calculating the cosine similarity between their embedding vectors of e_k^{en} and $e_k^{cn\rightarrow en}$ (see equation (4)).

$$\text{Sim}(T_k)^{ch\rightarrow en} = \frac{e_k^{en} \cdot e_k^{ch\rightarrow en}}{\left|e_k^{en}\right|\left|e_k^{ch\rightarrow en}\right|} \tag{4}$$

Similarly, we translated the English segment T_k^{en} into Chinese $T_k^{en\rightarrow cn}$, and computed the similarity score $\text{Sim}(T_k)^{en\rightarrow ch}$ between the original Chinese segment T_k^{cn} and the translated Chinese segment $T_k^{en\rightarrow cn}$ based on the cosine similarity of their vector embeddings.

Finally, we selected the higher of the two similarity scores as the final similarity value between the Chinese and the English segments in tweet T_k (see equation (5)).

$$\text{Sim}(T_k) = \max\left(\text{Sim}(T_k)^{ch\rightarrow en}, \text{Sim}(T_k)^{en\rightarrow ch}\right) \tag{5}$$

The higher the similarity score, the greater the likelihood that the CSed text represents a reiteration. Fig. 2 displays the distribution of similarity scores across our selected tweets. To determine a threshold for identifying reiterations, we applied the Inter-Quartile Range measure and empirically set the upper bound at 0.8. After removing trivial and reiterated CSed candidates, our dataset contained 13,603 CSed tweets.

4.2. Code-switched Offensive Language Identification

Since the expression of offensive language might be language-dependent, it presents new challenges to its identification in CSed text. To this end, we developed an approach that integrates lexicon-based and deep learning-based techniques within a human-in-the-loop framework for candidate CSed offensive tweet extraction, followed by manual curation for the final dataset. The approach is illustrated in Fig. 3.

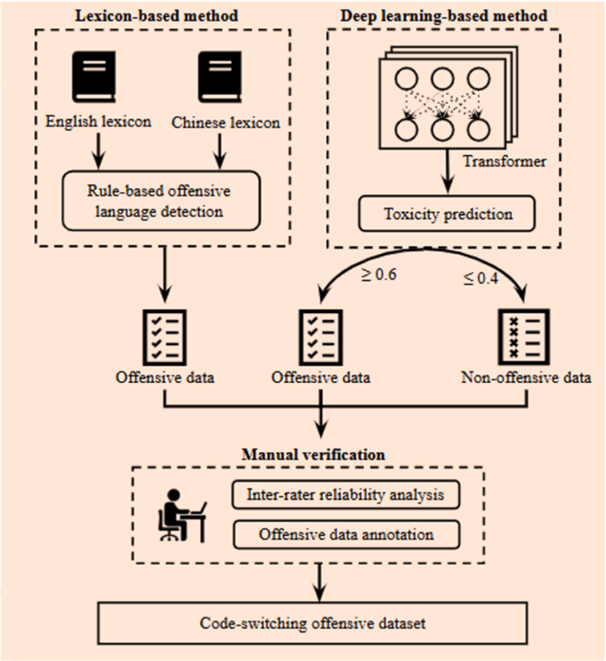


Fig. 3. Code-switched Offensive Language Identification.

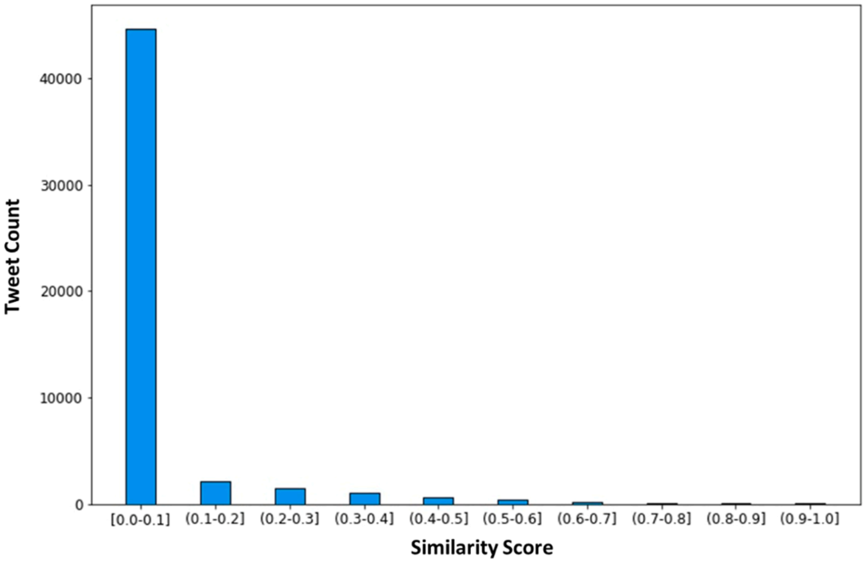


Fig. 2. Distribution of Similarity between Chinese and English Segments in CSed Tweets.

Lexicon-based method. The lexicon-based component uses a rule-based approach, comparing tweet content against predefined offensive language lexicons for both English and Chinese. Following previous studies on multilingual offensive content detection [106], [107], we constructed language-specific lexicons by merging existing resources [7,108], resulting in 1,415 English and 274 Chinese terms after redundancy removal. Applying this method to lemmatized CS tweets yielded 839 candidate offensive instances.

Deep learning-based method. The lexicon-based method, while straightforward, may be overly simplistic and thus struggle to deliver robust performance. For example, the meanings of terms such as “die”, “black”, and “日” (Chinese) are highly context-dependent, posing challenges for lexicon-based methods. To overcome this, we incorporated a deep learning approach, specifically a pre-trained multilingual character-level transformer [64], which excels at learning contextualized term embeddings. Based on our manual review of the model’s outputs, we empirically established a toxicity score of 0.6 as the minimum threshold for identifying offensive language and a score of 0.4 as the maximum threshold for classifying non-offensive language. These threshold choices were further corroborated by relevant literature (e.g., [109–111]). As a result, we identified 207 CSed offensive and 5,127 CSed non-offensive candidate tweets.

Manual verification. We manually validated the candidate tweets extracted using the two aforementioned methods. From the pool, we randomly selected 50 CSed offensive and 50 CSed non-offensive candidate tweets. Two coders, both proficient in English and Chinese and familiar with the Twitter platform, independently coded the tweets as either offensive or non-offensive. The inter-rater agreement and reliability assessed using Cronbach’s alpha, both reached 0.99, demonstrating exceptional reliability. Discrepancies were resolved through discussion, and these final resolutions consistently aligned with model predictions, validating the model’s performance. Given the near-perfect inter-rater agreement, one coder proceeded to manually verify the remaining candidate CSed offensive tweets, while the second coder cross-checked the results. Ultimately, we confirmed 377 CSed offensive tweets and 5,716 CSed non-offensive tweets.

5. Data Analyses and Results

In this section, we analyze the overall and language-specific stylometric characteristics of CSed offensive language. Additionally, we perform robustness tests to evaluate the impact of the stylometric features.

5.1. *Effects of Offensiveness on Stylometric Features*

5.1.1. *Overall Stylometric Features*

We report the descriptive statistics of the overall stylometric features in Table 2. We compared the stylometric characteristics of offensive and non-offensive language using an independent-samples t-test and the test results are also reported in the last column of the table.

Table 2
Descriptive Statistics of the Overall Stylometric Features of CSed Tweets.

Stylometric Features		Offensive		Non-offensive		t-test	
Type	Features	Mean	SD	Mean	SD	*t*	*p*-value
CS	Inter-switch	0.124	0.182	0.109	0.180	2.346	>.05
	Intra-switch	0.017	0.032	0.014	0.026	5.257	<.05
	Beginning switch	0.230	0.352	0.262	0.349	14.890	<.001
	Middle switch	0.738	0.363	0.712	0.356	13.077	<.001
	End switch	0.032	0.138	0.026	0.107	4.665	<.05
Lexical	URL	0.000	0.004	0.003	0.014	15.962	<.001
	Lexical diversity	0.831	0.174	0.850	0.163	4.766	<.05
	Exclamation mark	0.027	0.050	0.012	0.038	57.062	<.001
Syntactic	Sentence pausality	3.809	2.740	4.877	3.506	33.646	<.001
	Word pausality	0.162	0.084	0.175	0.085	8.392	<.01
Semantic	Content similarity	0.102	0.170	0.144	0.190	17.532	<.001

The analysis of lexical features reveals that *URLs* ($p < 0.001$) and *lexical diversity* ($p < 0.05$) are significantly lower, and *exclamation mark* is significantly higher ($p <.001$) in CSed offensive tweets compared to its non-offensive counterparts. The analysis results of syntactic features show that both *sentence pausality* ($p < 0.001$) and *word pausality* ($p < 0.01$) are significantly lower in CSed offensive text compared to non-offensive text. Furthermore, the comparison results of semantic features show that content similarity is significantly lower ($p < 0.001$) in CSed offensive than it non-offensive counterparts.

The analysis results of switch type show that *intra-switch* is more frequent in offensive than non-offensive text ($p < 0.01$), while *inter-switch* shows no significant difference ($p > 0.05$). The analysis results of switch location show that CSed offensive language is more likely to involve *middle* ($p < 0.001$) and *end switches* ($p < 0.05$), and less likely to use a *beginning switch* ($p < 0.001$), compared to CSed non-offensive tweets.

5.1.2. *Language-specific Stylometric Features*

We report the descriptive statistics of language-specific stylometric features in Table 3. Since these features incorporate both language and offensiveness dimensions, we conducted a two-way mixed model ANOVA on these features, using offensiveness (offensive vs. non-offensive) as a between-subjects factor and language (English vs. Chinese segment) as a within-subjects factor. The analysis results are reported in Table 4.

Table 3
Descriptive Statistics (Mean [SD]) of the Stylometric Features of Language Segments.

Stylometric Features		English		Chinese	
		Offensive	Non-offensive	Offensive	Non-offensive
Complexity	Word length	8.274 [2.847]	8.945 [8.945]	1.699 [0.220]	1.776 [0.246]
	Segment length	2.255 [2.154]	2.441 [2.175]	4.572 [1.794]	4.264 [1.836]
	Readability	-.379 [0.889]	.025 [1.002]	-.247 [0.914]	.016 [1.003]
Expressivity	Overall emotion	11.506 [16.187]	4.514 [9.498]	9.169 [8.215]	6.870 [7.244]
	Negative emotion	8.637 [14.678]	1.449 [5.216]	5.541 [6.149]	3.114 [4.958]
	Anxiety	0.505 [3.770]	0.233 [2.020]	0.258 [1.099]	0.388 [1.700]
	Anger	6.689 [12.443]	0.407 [2.588]	2.112 [4.336]	0.563 [2.164]
	Sadness	0.581 [5.713]	0.228 [1.885]	0.405 [1.509]	0.277 [1.498]

Table 4
ANOVA Results of Offensiveness and Language.

Stylometric Features		Offensiveness F(1,6091)=	Language F (1,6091)=	Offensiveness × Language F(1,6091)=
Complexity	Word length	20.389***	6758.2***	12.666***
	Segment length	.581	869.55***	12.412***
	Readability	67.76***	3.261†	4.247*
Expressivity	Overall emotion	179.45***	.001	56.336***
	Negative emotion	471.909***	12.246***	135.513***
	Anxiety	.881	.435	8.268**
	Anger	957.767***	342.942***	392.981***
	Sadness	10.392***	.783	2.45

Notes: ***: $p<.001$; **: $p<.01$; *: $p<.05$.

The results show that offensiveness has a significant effect on *average word length, readability*, and all *expressivity* features ($p<.001$) except *anxiety*. Specifically, CSed offensive language exhibits shorter *average word length* and greater *readability (i.e., lower scores)*, yet higher levels of *overall emotion, negative emotion, anger*, and *sadness*, compared to its non-offensive counterpart.

The analysis results show that the simple effects of offensiveness on *average word length, readability, overall emotion, negative emotion,* and *anger* (see Table 5) are consistent with its main effects. Interestingly, despite a lack of the main effects of offensiveness on *segment length* and *anxiety* ($p>.05$) and no interaction effect between offensiveness and language for *sadness* ($p>.05$), *segment length* is significantly shorter in offensive than non-offensive text for Chinese segments only ($p<.01$). Conversely, *anxiety* ($p<.05$) and *sadness* ($p<.01$) are significantly higher in offensive than non-offensive language for English segments only.

5.2. Effects of Language on Stylometric Features

Table 4 shows that language has a strong main effect on two of the complexity features: *word length* and *segment length* ($p<.001$), and a marginal effect on *readability* ($p<.1$). Additionally, among the expressivity features, language has a strong effect on *negative emotion* and *anger* ($p<.001$), but shows no effect on the remaining features ($p>.05$).

Given the significant interaction effects of language and offensiveness on all stylometric features ($p<.05$ or stronger) except sadness, we also analyzed the simple effects of language for offensive and non-offensive texts separately. The results are reported in Table 6, showing that the simple effects of language on *word length, segment length,* and *sadness* are consistent with its main effects reported earlier. Interestingly, the effect of language on *readability* is significant only for CSed offensive text ($p<.05$). Furthermore, although language significantly affects *overall emotion, negative emotion,* and *anger* regardless of offensiveness, the direction of these effects in CSed offensive text was opposite to that in non-offensive text. Specifically, English segments display higher levels of *overall emotion* ($p<.001$), *negative emotion* ($p<.001$), and *anger* ($p<.001$) than Chinese counterparts in CSed offensive text, whereas English segments show lower levels of *overall emotion* ($p<.001$), *negative emotion* ($p<.001$), and *anger* ($p<.01$) than Chinese counterparts in CSed non-offensive text. Additionally, language affects *anxiety* only in CSed non-offensive text ($p<.001$).

Table 5
Simple Effects of Offensiveness on Stylometric Features.

Stylometric Features		English		Chinese	
		Mean difference (N-O)	Std. Error (p-value)	Mean difference (N-O)	Std. Error (p-value)
Complexity	Word length	.672	.166***	.077	.013***
	Segment length	.186	.116	-.309	.098**
	Readability	.404	.053***	.263	.053***
Expressivity	Overall emotion	-6.992	.534***	-2.299	.389***
	Negative emotion	-7.188	.331***	-2.427	.268***
	Anxiety	-.272	.115*	.130	.089
	Anger	-6.281	.212***	-1.549	.125***
	Sadness	-.353	.123**	-.128	.08

Notes: ***: $p<.001$; **: $p<.01$; *: $p<.05$; *O*: offensive language; *N:* non-offensive language.

Table 6
Simple Effects of Language on Stylometric Features.

Stylometric Features		Offensive		Non-offensive	
		Mean difference (E-C)	Std. Error (p-value)	Mean difference (E-C)	Std. Error (p-value)
Complexity	Word length	6.575	.162***	7.17	.042***
	Segment length	-2.318	.136***	-1.823	.035***
	Readability	-.132	.066*	.009	.017
Expressivity	Overall emotion	2.337	.606***	-2.357	.156***
	Negative emotion	3.096	.396***	-1.665	.102***
	Anxiety	.247	.135†	-.155	.035 ***
	Anger	4.576	.231***	-.156	.059**
	Sadness	.176	.139	-.049	.036

Notes: ***: $p<.001$; **: $p<.01$; *: $p<.05$; *E*: English segment; *C:* Chinese segment.

5.3. Robustness Tests

5.3.1. Detection in Original CSed Text

To assess the robustness of the proposed stylometric features, we incorporated them as inputs into a downstream task: offensive language detection. Following the related literature [37,66], we selected the logistic regression technique (see Equation (6)) to construct models for detecting offensive language. In addition, we employed two state-of-the-art multilingual deep learning techniques, namely mBERT [107] and xlm-RoBERTa [112], in our model development.

$$\hat{y} = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}} \tag{6}$$

where $\boldsymbol{x} = \{x_1, x_2, \ldots, x_d\}$ represents our proposed d stylometric features, $\boldsymbol{w} = \{w_1, w_2, \ldots, w_d\}$ denotes the learnable weights associated with these features in the logistic regression models.

To better understand the impact of the stylometric features on offensive language detection, we built models with and without incorporating those features and used the latter as baseline models. For baseline models using logistic regression, we considered features that are commonly used in the literature for detecting offensive language, including dictionary words [113], bag-of-words [114], and n-grams [6]. N-grams achieved better performance in offensive language detection compared with dictionary words and bag-of-words [6] because the former features can take into consideration words' immediate context. Thus, we selected n-grams, and specifically unigrams and bigrams, as input features.

We evaluated model performance in offensive language detection using widely adopted metrics, including accuracy, macro precision, macro recall, and macro F-1 score. Macro metrics were defined as the average performance across both offensive and non-offensive classes, ensuring balanced evaluation without considering the class distribution.

To address the challenges arising from an imbalanced dataset (which contains a substantially larger number of non-offensive than offensive

CSed tweets), it is common practice to implement a sampling strategy on the training data [115,116]. To this end, we adopted an undersampling strategy. Specifically, we first applied stratified sampling to the original dataset by randomly selecting 80% of the data as the training set and the remaining 20% as the test data. We then performed undersampling on the training data by randomly selecting a number of non-offensive CSed tweets equal to the number of offensive ones, thereby achieving a balanced dataset. Moreover, the above process was repeated 100 times, and the results from these runs were averaged for reporting purposes.

The performance of the offensive language detection models, both with and without incorporating the proposed stylometric features, is reported in Table 7. The table shows that all evaluation metrics of the detection models improved dramatically when the proposed stylometric features were incorporated. The impact of those features is particularly pronounced in the state-of-the-art multilingual deep learning models, which improved the F1-score of mBERT in detecting offensive language by over 48% and that of xml-RoBERTa by over 25%, respectively. These findings highlight that integrating the stylometric features significantly enhances the model's ability to identify patterns in CSed offensive language.

5.3.2. Detection in Translated Monolingual Text

Since language translation can be an alternative approach for processing CSed text, we conducted an additional set of experiments as robustness tests. In these experiments, we first translated the CSed text into English using Google Translator API, then extracted stylometric features from the translated English text, and finally developed monolingual models (i.e., BERT and RoBERTa) by incorporating these features. To enable a fair comparison, we applied an identical set of features to the translated texts as we did to the CSed texts, particularly including CS type and location. The results are reported in Table 8. Two key observations emerge from the table. First, the stylometric features significantly boost model performance in offensive language detection. Second, the translation approach using monolingual text consistently underperforms compared to our proposed method across all models, regardless of whether the stylometric features are incorporated. Specifically, compared to their CSed counterparts, the F1 scores of the monolingual models in detecting offensive language dropped by 4.54% for BERT-based models, 6.03% for RoBERTa-based models, and 6.15% for LR models when the stylometric features were not included, and the F1 scores of the monolingual models in detecting offensive language dropped by 5.72% for BERT-based models, 1.19% for RoBERTa-based models, and 3.47% for LR models when the stylometric features were included. Our results suggest that directly processing CSed text preserves critical contextual nuances for offensive language detection, which are lost during translation.

6. Discussion

6.1. Major Findings and Explanations

The primary objective of this study was to identify the stylometric characteristics of CSed offensive language in social media. In this section, we summarize and discuss our main findings.

First, we examine complexity at both the overall CSed text and language-specific segment levels. At the CSed text level, offensive language exhibits lower complexity in terms of URLs and lexical diversity compared to the non-offensive text. This finding is corroborated by the analysis results of sentence pausality and word pausality, both are indicators of complexity [87]. We observe similar patterns at the individual language segment level, where CSed offensive text demonstrated shorter words and greater readability than non-offensive text.

Second, offensive CSed text displayed higher expressivity, as evidenced by increased exclamation mark usage. The finding is consistent with our observations regarding overall emotion, negative emotion, anger, anxiety, and sadness at the individual language segment level. They indicate that offensive language is more frequently associated with an emotional, particularly negative, emotional state. In addition, anxiety and sadness were more pronounced in offensive than non-offensive language in English segments but not in Chinese segments. Thus, these emotion-related effects of offensiveness seem more consistent across English than Chinese segments, potentially reflecting cultural differences in language choice during online CS and the specific context of our case study.

Third, the content similarity of CSed offensive language is significantly lower than that of non-offensive language. This suggests that non-offensive language tends to repeat similar content across different languages more frequently than its offensive counterpart. We provide the following explanations for this observation. The choice of language for expressing offensive content by a multilingual speaker may be influenced by factors such as social norms, cultural background, and linguistic proficiency. For example, a speaker might select a language prevalent within their cultural or social group to deliver offensive content effectively. Additionally, individuals may prefer a language in which they feel most at ease when articulating offensive content. Moreover, certain languages may carry stronger emotional weight, making them better suited for conveying offensive content and provoking the intended response. The findings in Table 4 substantiate that language choice significantly influences complexity and expressivity features. Consequently, offensive language may appear exclusively in a single language segment or vary distinctly across different language segments within CSed text.

Fourth, offensive language shows a higher level of intra-switch than non-offensive language. Additionally, offensive language is more likely to involve CS in the middle or at the end of text (as seen in tweets T1 and T2 from Table 9), and less likely to switch in the beginning, than non-offensive language. For instance, tweet T1 demonstrates intra-sentential and end-position CS, with a longer Chinese segment than the English one. The user starts by expressing frustration using offensive language in Chinese and ends with anger conveyed through offensive language in English. In contrast, tweets T3 and T4 in Table 9 illustrate inter-switch. Tweet T3 extensively employs exclamation marks and repetitions, beginning with a short offensive English phrase along with an exclamation mark repeated five times to explicitly express intense negative emotions, and ending with an implicit ironic statement in Chinese that differs entirely from the English content. Similarly, tweet

Table 7
Offensive Language Detection Performances on CSed text with and without the Proposed Stylometric Features.

Detection models	Stylometric features	Accuracy	Precision		Recall		F1-score	
			Offensive	Non-off	Offensive	Non-off	Offensive	Non-off
N-grams	w/o	64.76%	50.78%	79.10%	43.28%	72.93%	46.28%	75.66%
+LR	with	69.03%	51.81%	79.65%	58.42%	74.18%	53.75%	76.42%
mBERT	w/o	51.89%	18.74%	45.57%	51.08%	50.61%	25.57%	46.63%
	with	85.09%	65.81%	95.92%	88.52%	88.29%	73.86%	93.06%
xlm-	w/o	49.85%	26.18%	63.68%	59.55%	49.87%	40.99%	55.04%
RoBERTa	with	79.92%	62.51%	93.75%	80.11%	89.91%	66.27%	90.11%

Notes: Non-off: Non-offensive content

Table 8
Offensive Language Detection Performance on Translated Text.

Detection models	Stylometric features	Accuracy	Precision		Recall		F1-score	
			Offensive	Non-off	Offensive	Non-off	Offensive	Non-off
N-grams	w/o	60.25%	43.79%	71.24%	36.69%	64.08%	40.13%	67.30%
+LR	with	67.95%	47.22%	74.62%	53.70%	70.61%	50.28%	72.83%
BERT	w/o	42.11%	15.97%	43.46%	47.74%	48.07%	21.03%	44.61%
	with	78.69%	58.37%	92.00%	84.66%	81.80%	68.14%	87.18%
RoBERTa	w/o	43.08%	23.67%	56.18%	49.78%	44.58%	34.96%	49.76%
	with	75.13%	61.92%	91.98%	78.47%	85.00%	65.08%	87.25%

Notes: Non-off: non-offensive content

Table 9
Sample CSed Offensive Tweets and their Translated Versions.

Case	Original (CSed)	Translated Version
T1	幹你媽武漢肺炎快點結束 我真的想要看到大家能夠開開心心的做自己想做的without 這些shit thing.	‘Fuck you, Wuhan pneumonia will end soon. I really want to see everyone be able to do what they want without ‘these’ shit things.
T2	There is no credible evidence that you are serving the interest[s] of Canadians. Or I should say 去地狱吧, 你这个全球化的猪	There is no credible evidence that you are serving the interest of Canadians. Or I should say ‘Go to hell, you globalized pig’.
T3	FUCK NEWS! FUCK NEWS! FUCK NEWS! FUCK NEWS! FUCK NEWS! ((((新冠干得好, 为新中国71周年献礼了	Fuck News! Fuck News! Fuck News! Fuck News! Fuck News! ((((Well done to COVID-19, a gift for the 71st anniversary of New China.
T4	Making money from the pandemic! 黑心钱∝ 没屁眼ര Well, he is not the only one.	Making money from the pandemic! Dirty money? No asshole! Well, he is not the only one.

T4 shows low content similarity between its two language segments, further highlighting the distinct expression across languages in offensive CSed text.

Fifth, language affects several stylometric features of CSed text, including word length, segment length, negative emotion, and anger. Although the difference in word length may not be interesting due to inherent linguistic differences, the finding that Chinese segments are longer than English segments suggests that some social media users may prefer Chinese over English in a CSed context. Interestingly, compared to Chinese segments, English segments show greater readability (i.e., are easier to read) only in CSed offensive language, while displaying lower levels of anxiety solely in CSed non-offensive language. Furthermore, in CSed offensive language, English segments show higher levels of overall emotion, negative emotion, and anger than their Chinese counterparts, whereas these expressivity features are lower in English segments than in Chinese ones for the CSed non-offensive language. One potential explanation for these differences is that many English-Chinese speakers may consider Chinese as their primary language and accordingly have a greater proficiency in Chinese than in English. As a result, they may employ Chinese with greater complexity or difficulty when expressing offensive content and display more intense emotional expressions in Chinese segments for non-offensive content, compared to English. Another explanation is that “processing emotional information in a second language is less emotional than in a first language,” and “such a decrease in emotionality results in the neutralization of offense taken” ([117] p.395). Additionally, offensiveness has a positive effect on segment length only in Chinese segments. Taken together, these findings suggest that multilingual social media users tend to express themselves more extensively and emotionally in their first language than in their second language in a CSed non-offensive context, while they are more likely to convey emotions in their second language than their first when producing offensive text. To further explore this, we conducted an additional experiment comparing emotion intensity between first and second languages within the same CSed text. The results support our explanations, confirming that users are more inclined to express emotions in their second language rather than their first when crafting offensive content.

We can find several alternative explanations for the finding that the overall complexity of offensive language is lower than that of non-offensive language. First, offensive language typically draws from a narrow pool of words and phrases, often centered on vulgar or taboo subjects, which may create an impression of simplicity due to its restricted vocabulary. Second, offensive language often employs direct and unembellished communication, prioritizing provocation or shock over nuanced or elaborate expression, thus reducing the need for linguistic sophistication. Third, the strong emotional impact of offensive language may overshadow any potential complexity, as its affective intensity tends to take precedence over complex linguistic structures.

6.2. Research Contributions and Implications

This study offers multiple novel research contributions. First, it introduces stylometric characteristics of CS in social media text, extending traditional sociolinguistic theories of CS to online discourse and enriching the understanding of CS behavior by introducing the new dimension of switch location. Second, this study identifies a wide array of stylometric features specific to offensive language in CSed social media text for the first time. Despite existing discussions on the functions of CS, including its use for strategic purposes, emphasis, and emotion expressions, there remains a notable gap in understanding its role in offensive language. The study addresses the gap by shedding light on the role of CS in online offensive language, thereby extending existing CS theories. Third, it is the first to reveal the effect of language on the stylometric features of CSed offensive social media text. Fourth, it proposes a two-stage dataset construction method for CSed offensive language and develops the first dataset on Chinese-English CSed offensive language. Finally, it substantiates the proposed stylometric characteristics with empirical evidence, demonstrating their significantly positive effects on the performance of offensive language detection models.

Our findings have broad implications for research. Offensive content often transcends multiple languages and incorporates cultural references, and CS represents distinctive features of multilingual speakers. The findings of this study suggest that offensive language detection must account for these nuances to improve its effectiveness. Additionally, while deep learning models have shown impressive performance across various tasks, their opaque black- or gray-box nature limits their ability to directly enhance human understanding of offensive language. The stylometric features identified in this study pave the way for developing explainable algorithms, equipping users with insights into how offensive language is detected and flagged. This, in turn, empowers users to play an active role in mitigating online offensive language. Moreover, the proposed stylometric characteristics of CSed text provide a foundation for investigating online CS behavior in other types of contexts. For example, global brands might strategically incorporate CS elements in advertising campaigns to engage multilingual audiences, while political candidates could employ CS to tailor campaign messages that resonate with varied ethnic communities. The dataset construction methods introduced in this study can also be extended to develop CSed datasets for other language pairs or user behaviors.

The findings of this study also have practical implications for various stakeholders. The stylometric characteristics of online CSed offensive language identified in this research offer valuable tools for social media community moderators to enhance content moderation effectiveness. By integrating these features, social media platforms can improve their offensiveness detection models, enabling effective and efficient intervention against offensive language to foster civil discourse. Additionally, these insights can assist platforms in complying with legal regulations related to issues such as hate speech. For social media users, our findings raise awareness and deepen understanding of how CS is employed to express offensive language within their networks. This research contributes to cultivating a positive and respectful online environment, thereby enhancing the safety of social media users. This is particularly relevant for younger generations, who spend considerable time on social media but may lack the experience or awareness to navigate these spaces or filter out harmful content effectively.

6.3. Limitations and Future Work

This study has several limitations that highlight avenues for future research. First, the analysis treated offensive language as a single category. Given that offensive language manifests in diverse forms (e.g., abuse, insult, and profanity), a more fine-grained investigation is warranted. Future work could develop contextualized stylometric features tailored to these specific types within different contexts. Additionally, this study excluded CSed data samples exhibiting high content similarity between different language segments. Since this still represents a form of CS behavior, examining its underlying situational and motivational factors would be beneficial. Furthermore, the reliance on fixed toxicity score thresholds (0.6 for offensive, 0.4 for non-offensive) presents another limitation. Sensitivity analyses exploring the impact of varying these thresholds on the identified stylometric features would deepen the understanding of their robustness.

Second, our investigation focused on the general stylometric characteristics of CSed Chinese-English offensive language, without accounting for individual user differences, such as cultural background, linguistic proficiency, and specific linguistic variations. Exploring how such individual differences shape the stylometric characteristics of CSed text among multilingual speakers constitutes a significant research direction for future. This could further extend to authorship attribution studies based on multilingual users' CS stylometric profiles. The complex interplay of factors influencing language choice in CS for conveying offensive versus non-offensive content also merits more extensive study, particularly concerning cultural and contextual variables. Furthermore, the observed stylometric patterns of CSed offensive language may be language-pair dependent; and therefore, empirical investigations into CS between other language pairs (e.g., Spanish-English) are crucial for comprehensive generalizability assessment.

Third, this study did not consider different social media platforms, which may foster different CS styles. Platform-specific analyses are needed, because environments like Twitter/X (potentially favoring intra-switch) might differ from YouTube (potentially exhibiting more inter-switch) regarding offensive language expression. Finally, the use of COVID-19 as a specific case study limits generalizability. Future research employing broader datasets across various topics and contexts is helpful to validate and potentially refine current findings.

Declaration of Generative AI and AI-assisted technologies in the writing process: None

CRediT authorship contribution statement

Lina Zhou: Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation, Funding acquisition, Data curation, Conceptualization, Resources. **Zhe Fu:** Writing – review & editing, Writing – original draft, Validation, Methodology, Data curation, Investigation, Software.

Declaration of competing interest

None.

Acknowledgments

This work is partially supported by a Belk College Summer Research Grant, a Truist Business Research and Innovation Grant, and a Summer Seed Grant for Data Science at UNC Charlotte. Any opinions, findings or recommendations expressed here are those of the authors and are not necessarily those of the sponsors of this research.

References

[1] E.A. Vogels, Teens and cyberbullying, Pew Research Center, December 15 (2022). Available: https://www.pewresearch.org/wp-content/uploads/sites/20/2022/12/PI_2022.12.15_teens-cyberbullying-2022_FINAL.pdf.
[2] EIE, Cyberbullying statistics, 2024. Available: https://enough.org/stats_cyberbullying.
[3] Pew Research Center, The state of online harassment, 2021, January 13. Available: https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/.
[4] S. Liao, E. Okpala, L. Cheng, M. Li, N. Vishwamitra, H. Hu, F. Luo, M. Costello, Analysis of COVID-19 offensive tweets and their targets, in: In Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Long Beach, CA, USA, 2023, https://doi.org/10.1145/3580305.3599773.
[5] M. Vazquez, Calling COVID-19 the "Wuhan Virus" or "China Virus" is inaccurate and xenophobic, March 16, 2020. Available: https://medicalxpress.com/news/2020-03-covid-wuhan-virus-inaccurate-xenophobic.html#google_vignette.
[6] Y. Chen, Y. Zhou, S. Zhu, H. Xu, Detecting offensive language in social media to protect adolescent online safety, in: International Conference on Privacy, Security, Risk and Trust and International Confernece on Social Computing, 2012, pp. 71–80, https://doi.org/10.1109/SocialCom-PASSAT.2012.55.
[7] M. Wiegand, J. Ruppenhofer, A. Schmidt, C. Greenberg, Inducing a lexicon of abusive words – a feature-based approach, In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, New Orleans, Louisiana, 2018, pp. 1046–1056, https://doi.org/10.18653/v1/N18-1095.
[8] F. Bonetti, S. Tonelli, An analysis of abusive language data collected through a game with a purpose, In Proceedings of the 9th Workshop on Games and Natural Language Processing within the 13th Language Resources and Evaluation Conference, pp. 1–6, Marseille, France. European Language Resources Association, 2022.
[9] S. Hu, W. Lei, H. Zhu, C. Hsu, Cyberbullying perpetration on social media: A situational action perspective, Information & Management 61 (6) (2024) 104013, https://doi.org/10.1016/j.im.2024.104013.
[10] M.J. Matsuda, C.R. Lawrence III, R. Delgado, K.W. Crenshaw, Words that wound: Critical race theory, assaultive speech, and the First Amendment, Westview, Boulder, CO, 1993.
[11] M. Sullaway, Psychological perspectives on hate crime laws, Psychology, Public Policy, and Law 10 (2004) 250–292, https://doi.org/10.1037/1076-8971.10.3.250.
[12] S.K. Schneider, L. O'Donnell, A. Stueve, R.W. Coulter, Cyberbullying, school bullying, and psychological distress: a regional census of high school students, Am J Public Health 102 (1) (2012) 171–177, https://doi.org/10.2105/ajph.2011.300308.
[13] Pew Research Center, A majority of teens have experienced some form of cyberbullying, September 2018, Available: https://www.pewresearch.org/internet/2018/09/27/a-majority-of-teens-have-experienced-some-form-of-cyberbullying/.
[14] J.J. Gumperz, Discourse Strategies, Oxford University Press, 1982.
[15] C. Myers-Scotton, Common and uncommon ground: Social and structural factors in codeswitching, Language in Society 22 (4) (1993) 475–503, https://doi.org/10.1017/S0047404500017449.
[16] B. Danet, S.A. Herring, The Multilingual Internet: Language, Culture, and Communication Online, Oxford University Press, New York, 2007.
[17] S.N.A. Nazri, A. Kassim, Issues and functions of code-switching in studies on popular culture: A systematic literature review, International Journal of Language Education and Applied Linguistics 13 (2) (2023) 7–18, https://doi.org/10.15282/ijleal.v13i2.9585.
[18] S.S. Azari, V. Jenkins, J. Hahn, L. Medina, The foreign-born population in the United States: 2022, U.S. Census Bureau, ACSBR-019, 2024. Available: https://www.census.gov/library/publications/2024/acs/acsbr-019.html.
[19] S. Dietrich, E. Hernandez, Language use in the United States: 2019, U.S. Census Bureau, ACSBR-019, 2022. Available: https://www.census.gov/library/publications/2024/acs/acsbr-019.html.
[20] M. Heller, B. McElhinny, Language, capitalism, colonialism: Toward a critical history, University of Toronto Press, Toronto, 2017.

[21] M. Al-Emran, N. Al-Qaysi, Code-switching usage in social media: A case study from Oman, International Journal of Information Technology and Language Studies 1 (2017) 25–38.

[22] B. Migge, Code-switching and social identities in the Eastern Maroon community of Suriname and French Guiana1, Journal of Sociolinguistics 11 (1) (2007) 53–73, https://doi.org/10.1111/j.1467-9841.2007.00310.x.

[23] P. Piccinini, A. Arvaniti, Voice onset time in Spanish-English spontaneous code-switching, Journal of Phonetics 52 (2015) 121–137, https://doi.org/10.1016/j.wocn.2015.07.004.

[24] A. Georgakopoulou, Self-presentation and interactional alliances in e-mail discourse: the style- and code-switches of Greek messages, International Journal of Applied Linguistics 7 (2) (1997) 141–164, https://doi.org/10.1111/j.1473-4192.1997.tb00112.x.

[25] H. Liu, Intra-speaker variation in Chinese-English code-switching: The interaction between cognitive and contextual factors, International Journal of Bilingualism 22 (6) (2018) 740–762, https://doi.org/10.1177/1367006917698586.

[26] Z. Zhong, L. Fan, Worldwide trend analysis of psycholinguistic research on code switching using Bibliometrix R-tool, Sage Open 13 (4) (2023), https://doi.org/10.1177/21582440231211657.

[27] L. Barnes, The role of code-switching in the creation of an outsider identity in the bilingual film, Communicatio. 38 (3) (2012) 247–260, https://doi.org/10.1080/02500167.2012.716764.

[28] M. Deuchar, Code-switching in linguistics: A position paper, Languages. 5 (2) (2020) 22.

[29] P. Sheth, R. Moraffah, T.S. Kumarage, A. Chadha, H. Liu, Causality guided disentanglement for cross-platform hate speech detection, in: In Proceedings of the 17th ACM International Conference on Web Search and Data Mining, Merida, Mexico, 2024, https://doi.org/10.1145/3616855.3635771.

[30] K. Wang, Z. Fu, L. Zhou, D. Zhang, How does user engagement support content moderation? A deep learning-based comparative study, in: Americas Conference on Information Systems (AMCIS), 2023. Available: https://aisel.aisnet.org/amcis2023/sig_aiaa/sig_aiaa/3.

[31] J. Fillies, A. Paschke. Simple LLM based approach to counter Algospeak In Proceedings of the 8th Workshop on Online Abuse and Harms, Harms,Mexico City, Mexico, 2024, pp. 136–145, https://doi.org/10.18653/v1/2024.woah-1.10.

[32] D. Klug, E. Steen, K. Yurechko, How algorithm awareness impacts Algospeak use on TikTok, in: In Companion Proceedings of the ACM Web Conference 2023, Austin, TX, USA, 2023, https://doi.org/10.1145/3543873.3587355.

[33] E. Steen, K. Yurechko, D. Klug, You can (not) say what you want: Using Algospeak to contest and evade algorithmic content moderation on TikTok, Social Media + Society 9 (3) (2023), https://doi.org/10.1177/20563051231194586.

[34] P. Bhat, O. Klein, Covert hate speech: White nationalists and dog whistle communication on Twitter, in: G. Bouvier, J.E. Rosenbaum (Eds.), Twitter, the Public Sphere, and the Chaos of Online Deliberation, Springer International Publishing, Cham, 2020, pp. 151–172.

[35] I. Kwok, Y. Wang, Locate the hate: detecting tweets against blacks, in: In Proceedings of the 27th AAAI Conference on Artificial Intelligence, Bellevue, Washington, 2013.

[36] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, Y. Chang, Abusive language detection in online user content, in: In Proceedings of the 25th International Conference on World Wide Web, Montréal, Québec, Canada, 2016, https://doi.org/10.1145/2872427.2883062.

[37] G. Xiang, B. Fan, L. Wang, J. Hong, C. Rose, Detecting offensive tweets via topical feature discovery over a large scale twitter corpus, in: In Proceedings of the 21st ACM International Conference on Information and Knowledge Management, Maui, Hawaii, USA, 2012, https://doi.org/10.1145/2396761.2398556.

[38] A.A. Khan, M.H. Iqbal, S. Nisar, A. Ahmad, W. Iqbal, Offensive language detection for low resource language using deep sequence model, IEEE Transactions on Computational Social Systems (2023) 1–9, https://doi.org/10.1109/TCSS.2023.3280952.

[39] Z. Waseem, D. Hovy, Hateful Symbols or Hateful People?. Predictive features for hate speech detection on Twitter In Proceedings of the NAACL Student Research Workshop, San Diego, California, 2016, pp. 88–93, https://doi.org/10.18653/v1/N16-2013.

[40] V. Basile, et al., SemEval-2019 Task 5: Multilingual detection of hate speech against immigrants and women in Twitter, In Proceedings of the 13th International Workshop on Semantic Evaluation, Minneapolis, Minnesota, USA, 2019, pp. 54–63, https://doi.org/10.18653/v1/S19-2007.

[41] C. Adams, J. Sorensen, J. Elliott, L. Dixon, M. McDonald, N. Nithum, W. Cukierski. Toxic Comment Classification Challenge [Online]. Available: https://kaggle.com/competitions/jigsaw-toxic-comment-classification-challenge.

[42] J. Barnes, R. Klinger, S. Schulte im Walde. Bilingual sentiment embeddings: Joint projection of sentiment across languages, *Association for Computational Linguistics*, , Melbourne, Australia, 2018, pp. 2483–2493, https://doi.org/10.18653/v1/P18-1231.

[43] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, In Proceedings of the 5th International Workshop on Natural Language Processing for Social Media, Valencia, Spain, 2017, pp. 1–10, doi:10.18653/v1/W17-1101.

[44] R. Fredheim, A. Moore, J. Naughton, Anonymity and Online Commenting: The Broken Windows Effect and the End of Drive-by Commenting, in: In Proceedings of the ACM Web Science Conference, Oxford, United Kingdom, 2015, https://doi.org/10.1145/2786451.2786459.

[45] M. Mondal, L.A. Silva, F. Benevenuto, A measurement study of hate speech in social media, In, in: Proceedings of the 28th ACM Conference on Hypertext and Social Media, Prague, Czech Republic, 2017, https://doi.org/10.1145/3078714.3078723. Available:.

[46] F. Husain, O. Uzuner, A survey of offensive language detection for the Arabic language, ACM Transactions on Asian and Low-Resource Language Information Processing, 20 (1) (2021) 12, https://doi.org/10.1145/3421504. Article.

[47] E.J. Benson, The neglected early history of codeswitching research in the United States, Language & Communication 21 (1) (2001) 23–36, https://doi.org/10.1016/S0271-5309(00)00012-4.

[48] J.D. Takeuchi, Code-switching as linguistic microaggression: L2-Japanese and speaker legitimacy, Multilingua 42 (2) (2023) 249–283, https://doi.org/10.1515/multi-2021-0069.

[49] N. Jose, B.R. Chakravarthi, S. Suryawanshi, E. Sherly, J.P. McCrae, A survey of current datasets for code-switching research, in: In Proceedings of the 6th International Conference on Advanced Computing and Communication Systems (ICACCS), 2020, pp. 136–141, https://doi.org/10.1109/ICACCS48705.2020.9074205.

[50] S. Poplack, Sometimes I'll start a sentence in Spanish Y TERMINO EN ESPAÑOL: toward a typology of code-switching, Linguistics 18 (1980) 581, https://doi.org/10.1515/ling.1980.18.7-8.581.

[51] Y. Li, Y. Yu, P. Fung, in: A Mandarin-English code-switching corpus Proceedings of the8th International Conference on Language Resources and Evaluation (LREC), European Language Resources Assoc-Elra, Istanbul, Turkey, Paris, 2012, pp. 2515–2519.

[52] W. Astani, D. Rukmini, D. Sutopo, The impact of code switching in conversation of "Nebeng Boy" YouTube vlogs towards communication in English among the participants, English Education Journal 10 (2) (2020) 182–189.

[53] A.K. Yadav, M. Kumar, A. Kumar, Kusum Shivani, D. Yadav, Hate speech recognition in multilingual text: hinglish documents, International Journal of Information Technology 15 (3) (2023) 1319–1331, https://doi.org/10.1007/s41870-023-01211-z.

[54] M.A. Thelwall, Fk yea I swear: cursing and gender in MySpace, Corpora 3 (2008) 83–107.

[55] T. Jay and K. Janschewitz, The pragmatics of swearing, 4 (2), 267-288, (2008), doi:10.1515/JPLR.2008.013.

[56] P. Agarwal, A. Sharma, J. Grover, M. Sikka, K. Rudra, M. Choudhury, I may talk in English but gaali toh Hindi mein hi denge: A study of English-Hindi code-switching and swearing pattern on social networks, In Proceedings of the 9th International Conference on Communication Systems and Networks (COMSNETS), 2017, pp. 554–557, doi:10.1109/COMSNETS.2017.7945452.

[57] S. Gella, J. Sharma, K. Bali, Query word labeling and back transliteration for indian languages: Shared task system description, FIRE Working Notes 3 (2013) 89–105.

[58] H. Rizwan, M.H. Shakeel, A. Karim, Hate-speech and offensive language detection in Roman Urdu, Online, In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, (2020) 2512–2522, https://doi.org/10.18653/v1/2020.emnlp-main.197.

[59] K. Machová, M. Mach, K. Adamišín, Machine learning and lexicon approach to texts processing in the detection of degrees of toxicity in online discussions, Sensors 22 (17) (2022) 6468.

[60] K. Abainia, K. Kara, T. Hamouni, A new corpus and lexicon for offensive Tamazight language detection, in: In Proceedings of the 7th International Workshop on Social Media World Sensors, Barcelona, Spain, 2022, https://doi.org/10.1145/3544795.3544852.

[61] H. Khan, F. Yu, A. Sinha, S.S. Gokhale, A parsimonious and practical approach to detecting offensive speech, In Proceedings of the 2021 International Conference on Computing, Communication, and Intelligent Systems (ICCCIS), 2021, pp. 688–695, doi:10.1109/ICCCIS51004.2021.9397140.

[62] R. Lumbantoruan, R.U. Siregar, I. Manik, N. Tambunan, H. Simanjuntak, Analysis comparison of FastText and Word2vec for detecting offensive language, in: In Proceedings of the 2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM), 2022, pp. 1–8, https://doi.org/10.1109/ICOSNIKOM56551.2022.10034886.

[63] A.I. Alharbi, M. Lee, Combining character and word embeddings for the detection of offensive language in Arabic, In Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection, 2020, pp. 91–96, Marseille, France. European Language Resource Association, 2020.

[64] A. Lees, V.Q. Tran, Y. Tay, J. Sorensen, J. Gupta, D. Metzler, L. Vasserman, A new generation of perspective API: Efficient multilingual character-level Transformers, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, 2022, pp. 3197–3207, https://doi.org/10.1145/3534678.3539147.

[65] B. Alrashidi, A. Jamal, I. Khan, A. Alkhathlan, A review on abusive content automatic detection: approaches, challenges and opportunities, PeerJ Computer Science 8 (2022) e1142, https://doi.org/10.7717/peerj-cs.1142.

[66] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, in: In Proceedings of the International AAAI Conference on Web and Social Media 11, 2017, https://doi.org/10.1609/icwsm.v11i1.14955.

[67] G.A.D. Souza, M.D. Costa-Abreu, Automatic offensive language detection from Twitter data using machine learning and feature selection of metadata, in: International Joint Conference on Neural Networks (IJCNN), 2020, pp. 1–6, https://doi.org/10.1109/IJCNN48605.2020.9207652.

[68] V. Balakrishnan, S. Khan, H.R. Arabnia, Improving cyberbullying detection using Twitter users' psychological features and machine learning, Computers & Security 90 (2020) 101710, https://doi.org/10.1016/j.cose.2019.101710.

[69] B. Gambäck, U.K. Sikdar, Using convolutional neural networks to classify hate-speech, In Proceedings of the 1st Workshop on Abusive Language Online, Vancouver, BC, Canada, 2017, pp. 85–90, https://doi.org/10.18653/v1/W17-3013.

[70] H. Yenala, A. Jhanwar, M.K. Chinnakotla, J. Goyal, Deep learning for detecting inappropriate content in text, International Journal of Data Science and Analytics 6 (4) (2018) 273–286, https://doi.org/10.1007/s41060-017-0088-4.

[71] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for abusive language detection in English, In Proceedings of the 5th Workshop on Online Abuse and Harms, 2021, pp.17–25, doi:10.18653/v1/2021.woah-1.3.

[72] S. Sai, A.W. Jacob, S. Kalra, Y. Sharma, Stacked embeddings and multiple fine-tuned XLM-RoBERTa models for enhanced hostility identification. Combating Online Hostile Posts in Regional Languages during Emergency Situation, Springer International Publishing, Cham, 2021, pp. 224–235.

[73] E.W. Pamungkas, V. Basile, V. Patti, Towards multidomain and multilingual abusive language detection: a survey, Personal and Ubiquitous Computing 27 (1) (2023) 17–43, https://doi.org/10.1007/s00779-021-01609-1.

[74] M. Mozafari, R. Farahbakhsh, N. Crespi, Cross-lingual few-shot hate speech and offensive language detection using meta learning, IEEE Access 10 (2022) 14880–14896, https://doi.org/10.1109/ACCESS.2022.3147588.

[75] G. Vadakkekara Suresh, B.R. Chakravarthi, J.P. McCrae, Meta-learning for offensive language detection in code-mixed texts, Fire 21 (2022) 58–66, https://doi.org/10.1145/3503162.3503167.

[76] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, D. Woodard, Surveying stylometry techniques and applications, ACM Computing Surveys 50 (6) (2017) 86, https://doi.org/10.1145/3132039. Article.

[77] E. Stamatatos, Intrinsic plagiarism detection using character n-gram profiles, Threshold (2009) 38-46, Available: http://137.226.34.227/Publications/CEUR-WS/Vol-502/paper8.pdf.

[78] D.I. Holmes, J. Kardos, Who was the author? An introduction to stylometry, CHANCE 16 (2) (2003) 5–8, https://doi.org/10.1080/09332480.2003.10554842.

[79] M. Khonji, Y. Iraqi, A. Jones, An evaluation of authorship attribution using random forests, in: International Conference on Information and Communication Technology Research (ICTRC), 2015, pp. 68–71, https://doi.org/10.1109/ICTRC.2015.7156423.

[80] R. Manna, A. Pascucci, J. Monti, Profiling Fake News spreaders through stylometry and lexical features. UniOR NLP @PAN2020, in: Conference and Labs of the Evaluation Forum (CLEF), 2020.

[81] I. Markov, N. Ljubešić, D. Fišer, W. Daelemans, Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection,, In Proceedings of the 11th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2021, pp. 149–159.

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

[83] J.H. Clark, C.J. Hannon, A classifier system for author recognition using synonym-based features. MICAI 2007: Advances in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 839–849.

[84] R. Sarwar, Q. Li, T. Rakthanmanon, S. Nutanong, A scalable framework for cross-lingual authorship identification, Information Sciences 465 (2018) 323–339, https://doi.org/10.1016/j.ins.2018.07.009.

[85] G.J. Kootstra, J.G. Van Hell, T.O.N. Dijkstra, Priming of code-switches in sentences: The role of lexical repetition, cognates, and language proficiency, Bilingualism: Language and Cognition 15 (4) (2012) 797–819, https://doi.org/10.1017/S136672891100068X.

[86] J. Calvillo, L. Fang, J. Cole, D. Reitter, Surprisal predicts code-switching in Chinese-English bilingual text, in: In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, 2020, pp. 4029–4039.

[87] L. Zhou, J.K. Burgoon, J.F. Nunamaker, D. Twitchell, Automated linguistics based cues for detecting deception in text-based asynchronous computer-mediated communication: An empirical investigation, Group Decision & Negotiation 13 (1) (2004) 81–106, https://doi.org/10.1023/B:GRUP.0000011944.62889.6f.

[88] M.D. Capua, E.D. Nardo, A. Petrosino, Unsupervised cyber bullying detection in social networks, in: the 23rd International Conference on Pattern Recognition (ICPR), 2016, pp. 432–437, https://doi.org/10.1109/ICPR.2016.7899672.

[89] H. Watanabe, M. Bouazizi, T. Ohtsuki, Hate speech on twitter: A pragmatic approach to collect hateful and offensive expressions and perform hate speech detection, IEEE Access 6 (2018) 13825–13835.

[90] N.S. Halim, M. Maros, The functions of code-switching in Facebook interactions, Procedia - Social and Behavioral Sciences 118 (2014) 126–133, https://doi.org/10.1016/j.sbspro.2014.02.017.

[91] Y. Kang, L. Zhou, Helpfulness assessment of online reviews: The role of semantic hierarchy of product features, ACM Transactions on Management Information Systems 10 (2019) 1–18, https://doi.org/10.1145/3365538.

[92] J.P. Kincaid, R.P. Fishburne, R.L. Rogers, B.S. Chissom, Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel. Naval Air Station Memphis: Chief of Naval Technical Training, Research Branch Report (1975) 8–75. Available: https://apps.dtic.mil/sti/pdfs/ADA006655.pdf.

[93] W. Xu, Z. Yao, D. Chen, Readability of Chinese annual reports: measurement and testing, Accounting Research 3 (2021) 28–44.

[94] D.M. Howcroft, V. Demberg, Psycholinguistic models of sentence processing improve sentence readability ranking, in: In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics 1, 2017, pp. 958–968.

[95] J.W. Pennebaker, C.K. Chung, M. Ireland, A. Gonzales, R.J. Booth, The development and psychometric properties of LIWC2007, LIWC.net, Austin, TX, 2007.

[96] T. Kumar, V. Nukapangu, A. Hassan, Effectiveness of code-switching in language classroom in India at primary level: A case of L2 teachers' perspectives, Pegem Eğitim ve Öğretim Dergisi 11 (2021) 379–385, https://doi.org/10.47750/pegegog.11.04.37.

[97] S. Foster, A. Welsh, A 'new normal'of code-switching: Covid-19, the Indonesian media and language change, Journal Contribution (2021). https://doi.org/10.17509/ijal.v11i1.34621.

[98] B. He, C. Ziems, S. Soni, N. Ramakrishnan, D. Yang, S. Kumar, Racism is a virus: anti-asian hate and counterspeech in social media during the COVID-19 crisis, in: In Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, Virtual Event, Netherlands, 2022, https://doi.org/10.1145/3487351.3488324.

[99] J.T. Huang, M. Krupenkin, D. Rothschild, J.Lee Cunningham, The cost of anti-Asian racism during the COVID-19 pandemic, Nature Human Behaviour 7 (5) (2023) 682–695, https://doi.org/10.1038/s41562-022-01493-6.

[100] J. Huang, D. Tang, W. Zhong, S. Lu, L. Shou, M. Gong, D. Jiang, N. Duan. WhiteningBERT: An easy unsupervised sentence embedding approach, In Findings of the Association for Computational Linguistics: EMNLP, Punta Cana, Dominican Republic, 2021, pp. 238–244, https://doi.org/10.18653/v1/2021.findings-emnlp.23.

[101] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Minneapolis, Minnesota, 2019, pp. 4171–4186.

[102] Y. Liu, et al., RoBERTa: A robustly optimized BERT pretraining approach, ArXiv (2019) abs/1907.11692.

[103] H. Touvron, et al., LLaMA: Open and efficient foundation language models, ArXiv (2023) abs/2302.13971.

[104] J. Pennington, R. Socher, and C. D. Manning, Glove: Global vectors for word representation, In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, 2014. Available: https://www.aclweb.org/anthology/D14-1162/, https://doi.org/10.3115/v1/d14-1162">https://www.aclweb.org/anthology/D14-1162/, https://doi.org/10.3115/v1/d14-1162.

[105] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, Distributed representations of words and phrases and their compositionality, Advances in Neural Information Processing Systems, Stateline, NV, 2013. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.

[106] M. Corazza, S. Menini, E. Cabrio, S. Tonelli, S. Villata, A multilingual evaluation for online hate speech detection, ACM Transactions on Internet Technology, 20 (2) (2020) 10, https://doi.org/10.1145/3377323. Article.

[107] F.-z. El-Alami, S.Ouatik El Alaoui, N. En Nahnahi, A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model, Journal of King Saud University - Computer and Information Sciences 34 (8) (2022) 6048–6056, https://doi.org/10.1016/j.jksuci.2021.07.013. Part B.

[108] Y.R. Tausczik, J.W. Pennebaker, The psychological meaning of words: LIWC and computerized text analysis methods, Journal of Language and Social Psychology 29 (1) (2010) 24–54, https://doi.org/10.1177/0261927x09351676.

[109] N. Goyal, I.D. Kivlichan, R. Rosen, L. Vasserman, Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity annotation, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–28, https://doi.org/10.1145/3555088.

[110] M. Avalle, et al., Persistent interaction patterns across social media platforms and over time, Nature 628 (8008) (2024) 582–589, https://doi.org/10.1038/s41586-024-07229-y.

[111] M. Suarez Estrada, Y. Juarez, C.A. Piña-García, Toxic Social Media: Affective Polarization After Feminist Protests, Social Media + Society 8 (2) (2022), https://doi.org/10.1177/20563051221098343.

[112] M. Subramanian, et al., Offensive language detection in Tamil YouTube comments by adapters and cross-domain knowledge transfer, Computer Speech & Language 76 (2022) 101404, https://doi.org/10.1016/j.csl.2022.101404.

[113] S. Liu, T. Forss, New classification models for detecting hate and violence web content, in: the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K) 01, 2015, pp. 487–495.

[114] P. Burnap, M.L. Williams, Us and them: identifying cyber hate on Twitter across multiple protected characteristics, EPJ Data Sci 5 (1) (2016) 11, https://doi.org/10.1140/epjds/s13688-016-0072-6.

[115] G. Menardi, N. Torelli, Training and assessing classification rules with imbalanced data, Data Mining and Knowledge Discovery 28 (1) (2014) 92–122, https://doi.org/10.1007/s10618-012-0295-5.

[116] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, P.J. Kennedy, Training deep neural networks on imbalanced data sets, in: International Joint Conference on Neural Networks (IJCNN), 2016, pp. 4368–4374, https://doi.org/10.1109/IJCNN.2016.7727770.

[117] D. Miller, C. Solis-Barroso, R. Delgado, The foreign language effect in bilingualism: Examining prosocial sentiment after offense taking, Applied Psycholinguistics 42 (2) (2021) 395–416, https://doi.org/10.1017/S0142716420000806.

Lina Zhou is currently a Professor in the Department of Business Information Systems and Operations Management and the School of Data Science at the University of North Carolina (UNC) at Charlotte, USA. Prior to UNC Charlotte, she was a Professor in the Department of Information Systems at the University of Maryland Baltimore County. Her research interests include online misinformation, text mining, social media analytics, and medical informatics.

Zhe Fu is currently a PhD student at the Department of Software and Information Systems, University of North Carolina at Charlotte. His research interests include big data analysis, deep learning, and personalized recommender systems.