A Scalable Framework for Cross-lingual Authorship Identification

Raheem Sarwar[a], Qing Li[a], Thanawin Rakthanmanon[b,c], Sarana Nutanong[a,*]

[a]*Department of Computer Science, City University of Hong Kong, Tat Chee Avenue, Kowloon, HKSAR, China*
[b]*School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Thailand*
[c]*Department of Computer Engineering, Kasetsart University, Thailand*

Abstract

Cross-lingual authorship identification aims at finding the author of an anonymous document written in one language by using labeled documents written in other languages. The main challenge of cross-lingual authorship identification is that the stylistic markers (features) used in one language may not be applicable to other languages in the corpus. Existing methods overcome this challenge by using external resources such as machine translation and part-of-speech tagging. However, such solutions are not applicable to languages with poor external resources (known as low resource languages). They also fail to scale as the number of candidate authors and/or the number of languages in the corpus increases. In this investigation, we analyze different types of stylometric features and identify 10 high-performance language-independent features for cross-lingual stylometric analysis tasks. Based on these stylometric features, we propose a cross-lingual authorship identification solution that can accurately handle a large number of authors. Specifically, we partition the documents into fragments where each fragment is further decomposed into fixed size chunks. Using a multilingual corpus of 400 authors with 825 documents written in 6 different languages, we show that our method can achieve an accuracy level of

*Corresponding author

Email addresses: rsarwar2-c@my.cityu.edu.hk (Raheem Sarwar), itqli@cityu.edu.hk (Qing Li), thanawin.r@ku.ac.th (Thanawin Rakthanmanon), snutanon@cityu.edu.hk (Sarana Nutanong)

96.66%. Our solution also outperforms the best existing solution that does not rely on external resources.

Keywords: Stylometric Features, Similarity Search, Cross-lingual, Authorship Identification, Cyber forensic, Writeprint

1. Introduction

Authorship identification aims to identify the most likely author of a disputed document from a set of candidate authors [32, 5]. Recently, the practical applications of authorship identification have grown in several areas such as *criminal law*, e.g., identifying the writers of harassing letters or ransom notes [22]; *intelligence agencies work*, e.g., linking intercepted messages to known terrorists or enemies [1]; *civil law*, e.g., solving estate disputes or copyright issues [12]; *plagiarism detection*, e.g., determining whether work submitted by a student was written by someone else [7]. Authorship identification has also become a major part of other identification technologies including intrusion detection systems, cryptography and signatures [37].

The science of authorship identification is based on the observation that each individual author has a distinctive writing style [32]. There exists a long history of stylistic investigations focusing on authorship identification since the 19th century [32]. Most of the existing research in this area has used monolingual corpora and English is the most studied language [9, 29, 36, 43, 45]. However, nowadays, users may participate in several platforms regardless of the language [46]. For example, an Italian user may have a blog in Italian, primarily post in English on Facebook, and publish articles in both languages. Similarly, many novelists write in different languages, for example, Vladimir Nabokov wrote in both Russian and English and an Irish novelist Samuel Beckett wrote in both French and English [4]. Moreover, nowadays, around 45% content on the web is written in non-English languages [35]. Another aspect is that people are becoming increasingly proficient in more than one language. It has been shown that more than half of world population is bilingual [35]. The European Union

report shows that on average 94.5% pupils in secondary education learn two or more languages [1]. Consequently, there is a substantial need for cross-lingual authorship identification solutions.

Note that authorship identification in a multilingual corpus can be done by applying a monolingual authorship identification technique to each language independently. However, the ability to cross-compare stylistic variations of documents written in multiple languages allows for more data to be used to construct an authorship prediction model. Furthermore, many authors may have written a large number of documents in their respective native languages and a much smaller number of documents in some foreign language. Consequently, foreign language documents can be difficult to be identified if the analysis is limited to only one language at a time.

Bogdanova and Lazaridou [4] formally define the cross-lingual authorship identification problem as follows.

Definition 1.1. *[Cross-lingual Authorship Identification]* Cross-lingual authorship identification *aims to identify the author of a query document $\mathcal{Q}$ written in one language X from set of candidate authors using*

(i) writing samples from candidate authors written in a set $\mathcal{Y}$ of languages;

(ii) writing samples from the original author written in a set $\mathcal{Z}$ of languages, where $X \notin \mathcal{Z}$ and $\mathcal{Z} \subseteq \mathcal{Y}$.

Note that the restriction $X \notin \mathcal{Z}$ is imposed to ensure that while assessing a cross-lingual solution, the solution actually pertains the cross-lingual capability.

Recently developed solutions in this area have reported a high accuracy, e.g., 97% with 6 candidate authors and 2 languages. However, these studies have at least one of the following limitations.

Limitation 1: *Language Knowledge Dependency.* A class of existing methods rely on a certain form of internal language knowledge, e.g., a machine transla-

[1] `http://ec.europa.eu/eurostat/statistics-explained/index.php/Foreign_language_learning_statistics`

tor [4, 28, 46, 47]. These methods have reported a high accuracy for resourceful languages using a high dimensional feature set (i.e., word-based n-grams) [4, 28, 47]. However, they have the following limitations. First, the performance drastically drops when used with languages that we do not have enough knowledge to derive an extensive feature set or to construct a reliable machine translator. We call such languages *low-resource languages (LRLs)* [28]. Second, a translator might have brought its own stylistic elements masking the original literary style of the author [4]. Our objective in this investigation is to develop a solution that does not rely on any internal language knowledge.

Limitation 2: *A Variety of Languages in the Corpus.* Increasing the number of languages in the corpus negatively affects the classification accuracy [4, 46, 28, 46]. For example, Llorens-Salvador and Delany [28] show that increasing the number of languages from 1 to 2 decreases the accuracy from 95% to 88%. Our objective is to handle many languages effectively.

Limitation 3: *Size of the Candidate Author Set.* Existing studies report a drastic accuracy drop as the number of candidate authors increases in the set of candidate authors [29, 4, 28]. For example, Luyckx et al. [4] show that increasing the number of authors from 2 to 8 decreases the accuracy from 55% to 19%. Moreover, in existing studies based on cross-lingual authorship identificaton [4, 28, 47, 46], the number of candidate authors does not exceed 8. Our objective is to design a solution that can support hundreds of candidate authors for the given document.

Limitation 4: *Small Number of Document Samples per Class.* Existing solutions for cross-lingual authorship identification have reported a drastic drop as the number of training samples per candidate author (class) is reduced. For example, Llorens-Salvador and Delany [28] show that as the number of writing samples per class is dropped from 15 to 5 documents, the accuracy gets dropped from 81% to 47%. However, there exist a few studies which have achieved good accuracy using few text samples per author [39, 41, 38, 40, 42]. For example, Qian et al. [39] used on average 10 samples per author in their experiments. However, these studies were limited to the English language only. In this inves-

tigation, we design a solution that can handle an extremely data-poor condition, in which the average number of documents per candidate author is between 2 and 3.

In order to address the first two limitations, we identify a set of features that can be used across a large number of languages. Specifically, our feature space relies on a minimal set of linguistic assumptions: (i) the ability to tokenize a writing sample into words; (ii) the ability to identify sentence boundaries; and (iii) the use of punctuations. Following these three assumptions, we formulate a feature space with 16 language-independent features. A further dimensionality reduction analysis [14] is performed which in turn increases the accuracy and reduces the computational cost.

In order to address the third and fourth limitations, we adopt an instance-based learning method called the *probabilistic k nearest neighbor (PkNN)* classifier [20]. However, the main problem of the PkNN method is that the classifier is sensitive to outliers. To mitigate this problem, we use a stylometric data representation, which makes use of set similarity search [36] such that the stylistic variations between documents can be measured as a set distance [21].

We conducted an extensive set of cross-lingual performance studies using a large multilingual corpus. Specifically, our corpus contained documents in 6 different languages written by 400 authors, which were significantly larger than those for any existing study on cross-lingual authorship identification in terms of the number of languages and the number of authors [4, 28]. We also compared our solution against four different classifiers and the best language-independent competitor [28] (more details in Table 1). Experimental results show that our solution significantly outperforms the competitors.

Summary of contributions. The core contributions of this paper are as follows.

- A formulation of a language-independent feature space, which relies on a minimal set of linguistic assumptions.
- A cross-lingual authorship identification method that *does not* rely on machine translation or any internal knowledge of the languages in the

corpus.

- A multilingual corpus with the number of candidate authors and the number of languages being significantly greater than those in existing cross-lingual studies.
- An extensive set of experimental studies evaluating the performance of the proposed method against four classifiers and the best language-independent competitor in different cross-lingual settings.

The rest of the paper is organized as follows. Section 2 provides a literature review. Section 3 presents a solution overview. Our proposed solution is discussed in Section 4. In Section 5, we report results from our extensive experimental studies. Section 6 presents our concluding remarks and suggestions for future work.

2. Literature Review

The objective of authorship identification is to determine the author of a given anonymous document using a set of writing samples obtained from known candidate authors. Generally, an authorship identification task is performed in two main steps: feature engineering and analysis.

2.1. Stylometric Feature Engineering

Stylometry is the statistical analysis of variations in the authorship literary styles [32]. Stylometric features can be categorized into four types: idiosyncratic, structural, syntactic, and lexical.

- Idiosyncratic features include grammatical mistakes, misspellings and other usage anomalies [6].
- Structural features include style markers relating to the structure of the text sample and its layout [45]. For instance, the average number of words per paragraph and the average number of words per line.
- Syntactic features include the part-of-speech [45] and the use of function words [34].

- Lexical features include word-based and character-based statistical measures of lexical variations, e.g., the average word length and vocabulary richness [13, 8]. Other lexical features used in monolingual authorship identification include the frequencies of word n-grams and the frequency of stop words [2]. Several studies have reported a superior performance when using n-gram-based lexical features to distinguish between the authorial styles [10, 23, 19]. In these studies, the feature space contained the frequencies of stop-words and n-grams of the text samples.

While solving cross-lingual authorship identification, one cannot simply apply the earlier discussed features to multilingual corpora. It has been shown that while applying the features such as most frequent words or n-grams to a multilingual corpus, these feature sets of different languages are often orthogonal to each other. This renders documents written in different languages incomparable [2, 48, 26]. Similarly, the idiosyncratic features [6, 5] are not applicable in the cross-lingual authorship identification task, since grammatical mistakes and spelling errors are also language-dependent.

Cross-lingual authorship identification requires a set of *language-independent* features in order to conduct a meaningful analysis [28]. One approach is to use part-of-speech information as features, e.g., the number of nouns and the number of verbs in a writing sample [4]. However, the main drawback of this approach is that it assumes prior knowledge of the language and relies on the accuracy of part of speech taggers, which can be a challenge for low-resource languages [28]. Another approach is to translate all documents into a common language and perform a monolingual analysis [4]. Again, this approach relies on prior knowledge of each language in corpus and the quality of the machine translator [28, 15, 4, 17, 16, 18]. Due to the reliance on prior language knowledge, they are not considered completely language-independent. In addition, relying on a translator leads to issues related to the construction or availability of a reliable machine translator for low resource languages [28, 4]. Moreover, using a machine translator may bring its own stylistic elements masking the original literary style of the author [4].

One approach to achieve language independence is to use vocabulary richness features [28], e.g., the entropy of the word frequency distribution and the frequency of words that appear only once in the text sample. In addition, one may also use structural features [45] such as the average number of words per sentence and the number of sentences per paragraph.

Llorens-Salvador and Delany [28] proposed a language-independent feature extraction method. Specifically, they formulated two sampling techniques to obtain writing samples from each document. In the first technique, 500 and 1000-token chunks were randomly sampled from the document. In the second technique, 500 and 1000-token bags were randomly sampled from the document. The difference between a chunk and a bag is that words in a chunk are contiguous while those in a bag are randomly sampled from anywhere in the document. For each writing sample (chunk/bag), 8 language-independent features, such as vocabulary richness and language-independent token counts are extracted to create an 8D feature vector. We call this method RF-VRFS for short. The main disadvantage of this approach is that all structural features, which can be considered language-independent, are ignored.

Summary. Recall that the main objective of this investigation is to provide a solution which does *not* rely on any prior knowledge of the languages in the corpus. This objective implies the following two conditions.

- First, all features used in this investigation must be *language-independent* [28]. This condition prevents us from using the following types of features. (i) **Idiosyncratic features** are not applicable in the cross-lingual authorship identification task since grammatical mistakes and spelling errors are language-dependent [5]. (ii) **N-grams frequency features** or most frequent words features are applicable to only mono-lingual authorship identification. This is because, while applying these set of features to a multilingual corpus, most of these features from different languages are often orthogonal to each other. This makes documents written in different languages incomparable [48, 26, 2].

- Second, we cannot rely on any machine translation aid or a part-of-speech tagger. (i) The main drawback of part-of-speech features is that they assumes prior knowledge of the language and relies on the accuracy of part of speech taggers, which can be a challenge for low-resource languages [28]. (ii) One approach to dealing with cross-lingual corpora is to translate all documents into a common language and perform a monolingual analysis [4]. This approach relies on prior knowledge of each language in corpus and the quality of the machine translator [28, 15, 4, 17, 16, 18]. In addition, relying on a translator leads to issues related to the construction or availability of a reliable machine translator for low resource languages. [28, 4]. Moreover, using a machine translator may bring its own stylistic elements masking the original literary style of the author [4].

By incorporating these conditions into our solution design, our proposed method can operate in cross-lingual settings and become applicable to low-resource languages.

2.2. Stylometric Analysis Techniques

Stylometric analysis is concerned with obtaining authorship identification results from feature vectors extracted from text samples. Traditionally, this can be done by applying machine learning models directly to the feature vectors. Machine learning models used for cross-lingual authorship identification include support vector machines, naive bayes, logistic regression, nearest neighbors and random forest [4, 28]. The random forest classifier has reported a reasonable accuracy using vocabulary richness features [28]. On the other hand, in a study which employed machine translation [4], the logistic regression classification has led to superior accuracy using a larger set of features (i.e., word n-grams) in comparison to other classifiers. The nearest neighbor classifier has reported a reasonable accuracy (71%) when used in a corpus with a large number of candidate authors [36].

Recently, character-level convolutional neural networks (CNNs) have shown promising results in a monolingual setting [24, 49]. Kim et al. [24] proposed a

CNN architecture for a sentence classification problem. They have shown that a neural network with 5 convolutional layers and 3 fully connected layers can obtain the accuracy of up to 89.6% in a sentence classification problem with 6 classes.

Summary. Note that, cross-lingual authorship identification recently received attention by researchers. One earlier attempt towards cross-lingual authorship identification was made by Bogdanova and Lazaridou in [4]. One of the major limitations of their technique is that it relies on machine translation. They report that, using a translator brings its own stylistic elements masking the original literary style of the author and negatively affects the classification accuracy. In addition, relying on a translator arises the issue of constructing or availability of a reliable machine translator for low resource languages [4]. Later, Llorens-Salvador and Delany tried to address this limitation in [28] using language-independent stylometric features without the help of machine translation. The main disadvantage of this approach is that all structural features, which can be considered language-independent, are ignored. Since this technique uses language-independent features, we consider it as our competitor and call this technique RF-VRFS (description of RF-VRFS is given in Section 2.1).

2.3. Summary of Literature Review

Table 1 provides a summary of related authorship identification methods. As can be seen, in terms of the corpus size, our study has the largest corpus in terms of the number of languages, the number of authors, and the number of documents in comparison to all other cross-lingual studies.

Let us now consider the language independence of each method. The first two methods [4] rely on prior knowledge of the studied languages (part of speech taggers and machine translation). As a result, they are not completely language-independent. As for the kNN-based method [36], the feature sets include language-specific feature types: *lexical* and *syntactic*. Hence, it cannot be directly applied to cross-lingual settings.

Although the study of the CNN-based method [24] is confined to monolingual

corpora, the proposed character-level features can be considered to be language-independent as long as the studied languages share approximately the same character set. In other words, the method assumes no prior knowledge of the studied languages. However, when we conducted a preliminary experiment, we found that their method requires a large number of training samples per class. This means that the CNN-based method is *not* suitable for us in our data-poor setting.

The method (RF-VRFS) proposed by Llorens-Salvador and Delany [28] is language-independent, since it makes use of vocabulary richness features. We consider this method as our direct competitor in the experimental studies (Section 5).

3. Solution Overview

Figure 1 provides an overview of our *Cross-Lingual Set Similarity (CLSS)* solution which consists of four components: *feature extraction*, *features analysis*, *set similarity search* and *prediction aggregation*. In order to perform cross-lingual authorship identification, we partition the documents into fragments, where each fragment consists of 30,000 tokens[2]. We then further decompose each fragment into chunks of 1,500 tokens. We extract 16 stylometric features from each chunk and represent it as a 16-dimensional vector. As a result, each document is represented as a collection of fragments, where each fragment is in turn presented as a point set in a 16D vector space.

[2] A white space-separated sequence of characters

Table 1: Comparison of related authorship identification methods

Ref.	Classification Method	Prior Knowledge / Restrictions	#Authors	#Docs	Avg. #Docs. per Author	#Langs	Features
[4]	SVM	Part of Speech	6	34	5.2	2	Part of Speech
[4]	Logistic Regression	Machine Translation	6	34	5.2	2	Word 1,2,3-gram
[36]	kNN	English Only	136	2,386	17	1 (Monolingual)	Lexical, Syntactic, Structural
[24]	Conv. Neural Net.	Same Character Set	6	NA	NA	1 (Monolingual)	Characters
[28]	Random Forest	-	8	120	15	4	Vocabulary Richness
Proposed Method	kNN	-	400	825	2.06	6	Vocabulary Richness, Structural, Punctuations

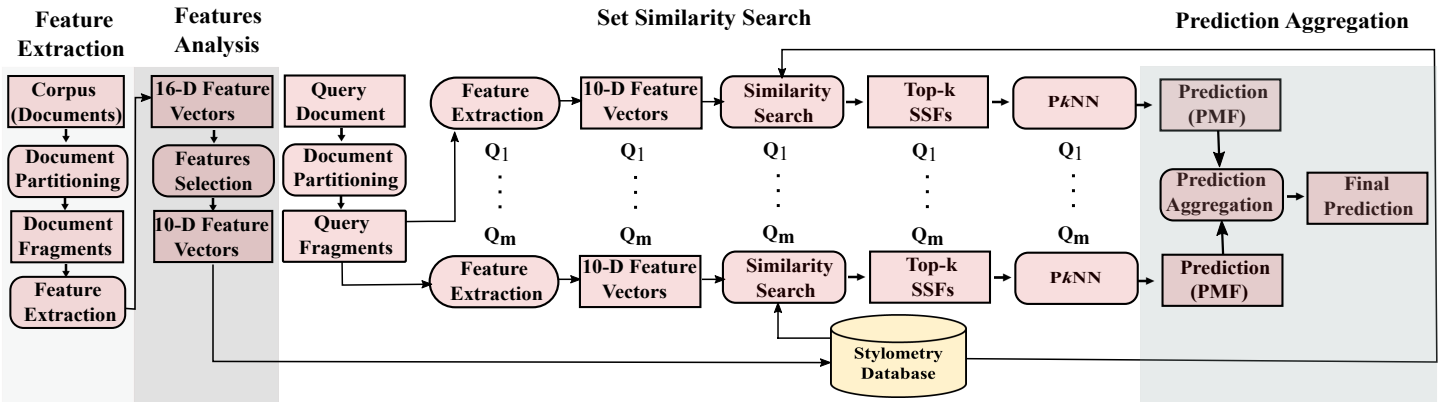

Figure 1: System overview: Cross-Lingual Set Similarity (CLSS)

Based on our document representation model, we are able to formulate the authorship identification problem as the following *set similarity search problem*.

- We partition the query document $\mathcal{Q}$ into m query fragments, where each query fragment Q contains a fixed number of points in a vector space.
- We identify stylistically similar document fragments (SSFs) in the corpus for each query fragment Q.
- In order to identify the top-k SSFs with the minimum distances, we compare the document fragments with respect to the query fragment Q.

We use 3 different set distance measures, which includes well known *standard Hausdorff distance (SHD)*, *partial Hausdorff distance (PHD)* [21], and *modified Hausdorff distance (MHD)* [27]. .

Note that each query fragment corresponds to an independent set similarity query. For example, if a query document $\mathcal{Q}$ contains four query fragments, four independent set similarity queries will be executed. This results in four different predictions being produced by the PkNN classifier. Results from these PkNN predictions are then aggregated to make a final prediction for the query document $\mathcal{Q}$.

In order to understand the prediction aggregation process, we need to first redefine the cross-lingual authorship attribution problem discussed in Section 1 as a probabilistic classification problem. Specifically, instead of just providing one single authorship prediction for each query fragment Q, we make a probabilistic prediction over a set of candidate authors. The updated problem definition is as follows.

Definition 3.1. *[Probabilistic Cross-lingual Authorship Identification] Cross-lingual authorship identification aims to assess the authorship likelihood by providing the probability mass function (PMF) over a set of likely authors of a query fragment Q written in one language X from set of candidate authors using*

(i) writing samples from the candidate authors written in a set $\mathcal{Y}$ of languages;

(ii) writing samples from the original author written in set $\mathcal{Z}$ of languages, where $X \notin \mathcal{Z}$ and $\mathcal{Z} \subset \mathcal{Y}$.

In this way, we can reliably combine predictions from different query fragments Q by computing the average PMF over all query fragments Q corresponding to the same document $\mathcal{Q}$. Note that based on this probabilistic definition of cross-lingual authorship identification, one can easily convert a probabilistic prediction into a non-probabilistic one by using the mostly likely outcome.

4. Proposed Solution

In this section, we describe the four components of our solution, namely *feature extraction*, *feature analysis*, *set similarity search*, and *prediction aggregation*.

4.1. Feature Extraction

As described in Section 3, writing samples are decomposed into 1,500-token chunks. For each chunk, we extract language-independent stylometric features. In order for our method to be applicable to a broad range of languages, our feature space relies on the following minimal set of linguistic assumptions: (i) the ability to tokenize a writing sample into words; (ii) the ability to identify the start and end of sentences; and (iii) the use of punctuations.

From the stated assumptions, we could identify 16 features, which could be categorized into three different classes: vocabulary richness [13], structural [45], and punctuation-based [29, 45, 25]. Table 4 provides a summary of these 16 language independent stylometric features.

Let us now discuss these feature types in details using the writing sample given in Table 2.

- The first type of features are concerned with the **vocabulary richness**. There are 10 of them. In order to compute the 10 vocabulary richness features (Features 1 to 10 in Table 4), we first need to determine the frequency F_j of each distinct word j as shown in Table 3. From the frequency table, we can obtain the total number N of words as the summation of the frequency column and the number V of distinct words as the number

of rows. In this case, we have $N = 18$ and $V = 13$. Next we build a type-token ratio table to obtain the frequency V_i of each frequency value ($i = F_j$) in Table 3. In this case, we have $V_1 = 9$, $V_2 = 3$, and $V_3 = 1$. That is, there are 9 words that appear once, 3 words that appear twice, and one word that appears thrice, respectively. Using these values, we can obtain the results for Features 1 to 10 through substitution as shown in Table 4.

- The second type of features are **structural** features, i.e., the average number of words per sentence and the number of sentences in the chunk. These features and their corresponding values (as obtained from the text sample) are also shown in Table 4 as Features 11 and 12.
- The third type of stylometric features are **punctuation** frequencies, i.e., (i) the frequency of quotations; (ii) frequency of punctuations; (iii) frequency of commas; and (iv) frequency of special characters. These features and their corresponding values (obtained from the text sample) are also shown in Table 4 as Features 13 to 16.

We call this feature space *Vocabulary_richness-Structural-Punctuation* or *VSP* for short.

Table 2: The example text for feature extraction

Chunk
Truth is stranger than fiction, but it is because Fiction is obliged to stick to possibilities; Truth isn't.

4.2. Feature Analysis.

The feature analysis component of our solution finds the high variance feature subspace. Specifically the feature analysis consists of two tasks including subspace selection and subspace evaluation. The subspace selection task of the feature analysis component is completely unsupervised. This implies that, only training data points were used to identify a feature subspace with a high variance. Specifically, we use the *recursive feature elimination (RFE)* technique

Table 3: The term frequencies for the given text sample in table 2

Word j	**Word**	**Frequency** F_j
1.	is	3
2.	truth	2
3.	fiction	2
4.	to	2
5.	stranger	1
6.	than	1
7.	but	1
8.	it	1
9.	because	1
10.	obliged	1
11.	stick	1
12.	possibilities	1
13.	isn't	1

proposed by Guyon et al. [14] to identify a high variance features subspace. We used RFE technique to construct stylometric feature subspaces with 15, 14, 13, 12, 11, 10, 9 and 8 numbers of dimensions. In the evaluation task of the feature evaluation process, we assessed the performance/accuracy of these feature subspaces using a 10-fold nested cross-validation method using *only* the training data points from the corpus and their labels. Our results from feature analysis component showed that the feature subspace with 10 dimensions yielded the best performance. The retrieved set of high variance features in turn reduces the computational and storage costs and improves the classification accuracy. Using this method, the Features 1,3,7, and 9-15 from Table 4 were selected. Finally, we stored these features in the database where each document is represented as a collection of point sets (fragments).

Table 4: The language-independent VSP stylometric features (* Features selected after dimensionality reduction analysis; Types: Vocabulary Richness (V), Structural (S), and Punctuation (P))

		Stylometric Features	**Values**	**Type**
1.	*	V	13	V
2.		$VR = \frac{V}{N}$	0.72	V
3.	*	$VR_S = \frac{V_2}{V}$	0.23	V
4.		$VR_R = \frac{V}{\sqrt{(N)}}$	3.06	V
5.		$VR_C = \frac{\log V}{\log N}$	0.88	V
6.		$VR_K = \frac{\log V}{\log(\log N)}$	12.33	V
7.	*	$VR_N = \frac{(1-V^2)}{(V^2 log N)}$	-0.79	V
8.		$VR_H = \frac{(100 \log N)}{(1-V_1/V)}$	403.22	V
9.	*	$VR_K = \frac{10^4(\sum i^2 V_i - N)}{N^2}$	740.74	V
10.	*	Entropy $= -\sum_{j=1}^{V} \frac{F_j}{N} \log \frac{F_j}{N}$	1.01	V
11.	*	Average number of words per sentence	18	S
12.	*	Number of sentences	1	S
13.	*	Frequency of punctuations	4	P
14.	*	Frequency of quotations	0	P
15.	*	Frequency of commas	1	P
16.		Frequency of special characters	0	P

4.3. Set Similarity Search

When a query document $\mathcal{Q}$ is submitted to our system, we repeat the same feature extraction process. That is, the query document $\mathcal{Q}$ is represented as a collection of points sets, where each set contains twenty points in a 10 dimensional vector space.

After the feature extraction process, we use each query fragment Q to iden-

tify the top-k *stylistically similar fragments (SSFs)*. Specifically, we use three set distance measures, namely *standard Hausdorff distance (SHD)*, *partial Hausdorff distance (PHD)*, and *modified Hausdorff distance (MHD)*. .

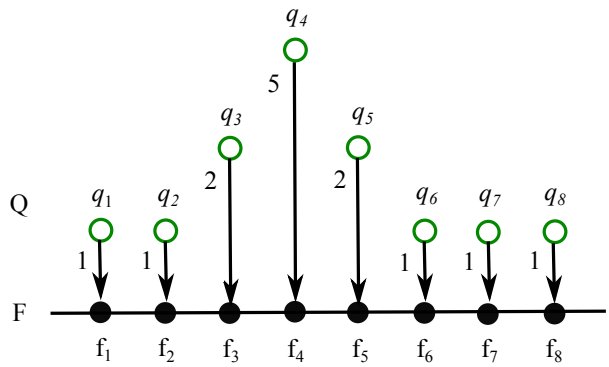

Figure 2: Hausdorff Distance Calculations

Table 5: Hausdorff Distance Calculations

Rank	Percentile	Min. Dist.	Dist.	SHD [100%]	MHD (50%,100%]	PHD (50%,75%]
1.	100	d(q_4, f_4)	5	✓	✓	
2.	87.5	d(q_3, f_3)	2		✓	
3.	75.0	d(q_5, f_5)	2		✓	✓
4.	62.5	d(q_1, f_1)	1		✓	✓
5.	50.0	d(q_2, f_2)	1			
6.	37.5	d(q_6, f_6)	1			
7.	25.0	d(q_7, f_7)	1			
8.	12.5	d(q_8, f_8)	1			
				5	2.5	1.5

Let us now briefly describe the set distance measures with the help of Figure 2 and Table 5. In Figure 2, we show the minimum distance between each data point in the query fragment Q and a document fragment F where each edge value shows the distance value. According to the definition of the *standard Hausdorff*

distance (SHD), the distance between two points sets is the maximum of all the minimum distances. As a result, the SHD from Q to F is $d(q_4, f_4)$, which is 5 units.

A drawback of using SHD is that the distance value is highly affected by outliers. In this case, most of the entries in Q have a minimum distance less than 2 units and the distance $d(q_4, f_4)$ of 5 units can be considered as an outlier. To mitigate this problem, one may use the *modified Hausdorff distance (MHD)* which is calculated by (i) ranking all data points in Q according to the minimum distance to F; and (ii) computing the average of the minimum distances within a given percentile range as shown in Table 5. In this example, we assume a percentile range of (50%, 100%]. (The second parameter is always 100% for MHD.) As a result, the entries marked with ✓, i.e., $d(q_4, f_4)$, $d(q_3, f_3)$, $d(q_5, f_5)$, and $d(q_1, f_1)$ are used in the calculation and the average distance is 2.5 units.

The partial Hausdorff distance (PHD) handles outliers in a more aggressive fashion. That is, a range of top-ranked distances are ignored completely. In this example, we assume a percentile range of (50%, 75%], i.e., the top 25% are ignored from the calculation. This range corresponds to $d(q_5, f_5)$ and $d(q_1, f_1)$ and the final distance value of 1.5 units.

Using one of the described set distances, the set-similarity component provides a set of *stylistically similar fragments (SSFs)* for each query fragment Q. Finally, we apply the PkNN classifier to the retrieved top-k SSFs in order to produce a prediction for each query fragment Q. Specifically, we adopt the PkNN variant [20] which uses the distances of k nearest neighbors (SSFs in this case) to weight the probability contributions. An exponential function is used to smoothen the distance-probability mapping. The final product is a probability mass function (PMF) over all classes (candidate authors) corresponding to the retrieved SSFs.

Finally, we aggregate multiple fragment-wise predictions in order to produce a final prediction for the entire query document $\mathcal{Q}$. Consider the example in Table 6. The query document $\mathcal{Q}$ has 4 fragments $\{Q_1, Q_2, Q_3, Q_4\}$ as shown in the first column. The prediction (PMF) corresponding to each fragment is

given in the second column. In this example, there are three candidate authors, namely A, B, and C. The final prediction is computed as the average PMFs of all four PMFs.

Note that, in order to handle a large dataset, we apply the probabilistic k-nearest neighbor (PkNN) classification technique to compute the probabilistic distribution over the candidate authors [20]. The motivation for using probabilistic k-nearest neighbor (PkNN) classification technique [20] is that it is an instance-based learning method. That is, the classification is performed through a comparison with instances stored in memory instead of building a generalized model. In addition, the advantages of using PkNN include (i) little or no training is required to perform classification task [33]; (ii) the learning model can make use of a complex target function [33]; (iii) it can incrementally add new information at runtime [3]; (iv) there is no information loss through generalization [33]; (v) it can learn from a limited set of examples [3]; (vi) it is a non-parametric method and does not require a priori knowledge relating to probability distributions for the classification problem [30]; and (vii) by using our set representation with the PkNN method, we effectively transform the cross-lingual authorship identification problem into a set similarity problem. This enables us to make use of a large array of set distance measures associated with outlier handling techniques. Consequently, it enable us to handle a large number of languages in the corpus and a greater number of candidate authors than any existing cross-lingual authorship identification technique.

Table 6: Predictions Aggregation

Query Fragment	Prediction (PMF)
Q_1	$[A : 0.40, B : 0.30, C : 0.30]$
Q_2	$[A : 0.50, B : 0.25, C : 0.25]$
Q_3	$[A : 0.30, B : 0.40, C : 0.30]$
Q_4	$[A : 0.65, B : 0.15, C : 0.20]$
Average PMF	$[A : 0.46, B : 0.28, C : 0.26]$

5. Performance Evaluation

In this section, we report results from our extensive experimental studies. Our experimental studies were organized into two sets of studies. First, we used a large corpus to show that our *Cross-Lingual Set Similarity (CLSS)* method could handle a large number of candidate authors (Section 5.2). Second, we reduced the corpus size in order to fairly compare our proposed CLSS method with our competitors (Section 5.3 and Section Appendix C). The experimental setup details are as follows.

5.1. Experimental Setup

Dataset. We extracted our dataset from an online book archive, Project Gutenberg[3], whose statistics[4] shows that the top six languages in terms of number of documents are English, French, German, Finnish, Dutch and Portuguese. However, in terms of second language of authors in Project Gutenberg archive, the top six languages are English, French, German, Finnish, Dutch and Spanish. We chose the documents written in these languages for experiments. Our corpus contained 825 novels from 400 different authors. The author distribution with respect to the number of languages in which they write is given in Table 7. Note that there are 196 monolingual authors and 204 multilingual authors, while 25 authors write in 3 languages or more.

Table 7: Number of authors by the number of languages they use.

Number of Languages	Number of Authors
1 (Monolingual)	196
2 (Bilingual)	179
> 3	25
Total	400

[3]https://www.gutenberg.org
[4]http://www.gutenbergnews.org/statistics/

Table 8 shows the language distribution of our dataset. As can be seen, the number of documents written in different languages are approximately the same. Clearly, there is no bias towards any particular language.

Table 8: Dataset description: Data sizes per language in terms of the number of documents, number of fragments, number of chunks, and number of tokens.

Language	#Documents	#Fragments	#Chunks	#Tokens
Dutch	133	3,676	73,535	110,302,500
English	143	4,092	81,845	122,767,500
French	141	3,917	78,341	117,511,500
Finnish	133	3,886	77,732	116,598,000
German	135	3,737	74,757	112,135,500
Spanish	140	3,868	77,368	116,052,000
Total	825	23,176	4,63,578	695,367,000

Note that, in terms of the number of authors and the number of languages, our corpus is significantly larger than any of those in the existing studies on cross-lingual authorship identification [4, 28]. For example, the studies of Bogdanova and Lazaridou [4] and the studies of Llorens-Salvador and Delany [28] involve fewer than 9 authors and 120 documents, and the number of languages does not exceed 4. This test condition was designed to verify the claim that our method is designed to overcome the following two limitations: (i) the language variety and (ii) the number of candidate authors, as stated in the introduction (Section 1).

In terms of the number of documents per candidate author (class), as can be seen in Table 7, our average is 2.06 (825 documents per 400 candidate authors), which is much lower than any existing studies [4, 28]. We use a much lower number of documents per candidate author than any existing study in order to evaluate our proposed method in an extreme data-poor condition. As stated in the introduction (Section 1), this is also one of the limitations of existing techniques we aim to overcome.

Evaluation Measures. As exemplified by Table 12, predictions are made at two different levels: *fragment* and *document*. Hence, we evaluate the accuracy as follows.

(i) *Fragment accuracy (FA)*: The method makes the correct prediction for a particular query fragment Q, i.e., the correct author is identified as the most likely author of Q.

(ii) *Document accuracy (DA)*: The aggregate prediction obtained from different query fragments corresponding to the same query document results with the correct author being the most likely author.

Parameters. We compared the accuracy of our method by varying (i) the number Ω of authors; and (ii) the number $\mathcal{L}$ of languages. As for the value k of PkNN, ideally we want k to be just large enough to obtain stable statistics, while keeping the retrieval cost low. We tested different k values and found that the k value 10 provides the best trade-off. As for the Hausdorff distance variants, following an experimental analysis, we chose the MHD percentage range of (50%,100%] and the PHD percentage range of (50%, 75%]. As for the fragment size $|Q|$, we found that the size of 20 chunks per fragment provides the best result. A summary of these parameter settings is given in Table 9.

Table 9: Parameters: The description of parameters and their values.

Parameter	Value	Description
k	10	The top-k closest fragments with respect to query fragment to consider for PkNN
Q	20 points	The size of a fragment, i.e. 20 chunks
MHD	(50%, 100%]	Average of ranked distances that fall in the specified range
PHD	(50%, 75%]	Average of ranked distances that fall in the specified range

Evaluation Strategy. As can be seen from Table 7, most of the multilingual authors are bilingual. In order to make sure that all 6 languages can be used

Table 10: List of abbreviations and their description

Abbreviation	Description
VSP	Our feature space consists of vocabulary richness (V), structural (S) and punctuation (P) based features and we call it VSP
VRFS	Feature space used by language-independent competitive method [28]
CLSS	Our proposed Cross-Lingual Set Similarity (CLSS) method
CLSS-VSP	The VSP feature space applied to our CLSS method (proposed solution)
CLSS-VRFS	The VRFS feature space applied to our CLSS method
RF-VRFS	The VRFS feature space applied to *random forest (RF)* method (Languguage-independent competitive method [28])
RF-VSP	The VSP feature space applied to *random forest (RF)* method
SVM-VRFS	The VRFS feature space applied to the *support vector machines (SVM)* method
SVM-VSP	The VSP feature space applied to the *support vector machines (SVM)* method
LR-VRFS	The VRFS feature space applied to the *logistic regression (LR)* method
LR-VSP	Our proposed VSP feature vectors applied to the *Logistic Regression (LR)* method
NB-VRFS	The VRFS feature space applied to the *naive bayes (NB)* method
NB-VSP	The VSP feature space applied to the *naive bayes (NB)* method

Table 11: Query-identifier pairs of documents organized according to identification check types. (Author#,Query Doc#, Identifying Doc#) Reference: https://www.gutenberg.org/wiki/Gutenberg:Offline_Catalogs

		Query					
		English	Dutch	Spanish	German	Finnish	French
Ident.	English	-	(1714,17528,5157)	(1708,10821,21700)	(26292,24174,26339)	(115,16944,203)	(112,15554,1686)
		-	(3026,17523,14031)	(913,7109,18569)	(1995,24746,7014)	(708,52852,27523)	(907,9800,7409)
	Dutch	(1714,5157,17528)	-	(2183,14795,21848)	(3624,31527,10664)	(528,14433,21945)	(136,17949,1399)
		(3026,14031,17523)	-	(492,24988,40169)	(4959,14340,17637)	(5355,17628,27489)	(239,24007,16102)
	Spanish	(1708,21700,10821)	(2183,21848,14795)	-	(481,50887,46196)	(2769,12382,29511)	(25891,23520,27121)
		(913,18569,7109)	(492,40169,24988)	-	(35,49424,45438)	(593,26724,25671)	(1336,20950,14236)
	German	(26292,26339,24174)	(3624,10664,31527)	(481,46196,50887)	-	(1772,19260,10507)	(60,5097,15559)
		(1995,7014,24746)	(4959,17637,14340)	(35,45438,49424)	-	(586,13328,10425)	(251,41211,5097)
	Finnish	(115,,20316944)	(528,21945,14433)	(2769,29511,12382)	(1772,10507,,19260)	-	(35,42,18123)
		(708,27523,52852)	(5355,27489,17628)	(593,25671,26724)	(586,10425,13328)	-	(314,14918,45263)
	French	(112,1686,15554)	(136,1399,17949)	(25891,23520,27121)	(251,5097,41211)	(35,18123,42)	-
		(907,7409,9800)	(239,16102,24007)	(1336,20950,14236)	(60,15559,5097)	(314,45263,14918)	-

to identify each other, there are 15 language pairs to check and each pair has two identification checks as shown in Figure 3. For example, the language pair (English, French), corresponds to the following identification checks.

- French → English: Checking whether French documents can be used to correctly identify the authorship of an English document written by the same author.
- English → French: Checking whether English documents can be used to correctly identify the authorship of a French document written by the same author.

For each identification check $X \rightarrow Y$, two query documents written in Y are used. Details of the query documents from Project Gutenberg used in our experiments is given in Table 11.

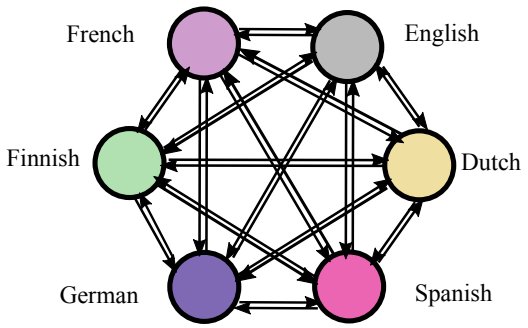

Figure 3: Language Evaluation Strategy (Edge represents the identifier language and arrow represents the query language)

According to Definitions 1.1 and 3.1, for each identification check, we needed to ensure that there is no self-language contamination. For example, assume that Author 44 is bilingual and writes Doc 112 in French and Doc 116 in English. The French → English identification check can be done by

(i) using the English document (Doc 116) as the query document $\mathcal{Q}$;

(ii) temporally removing all English documents written by Author 44 from the corpus;

(iii) leaving the French document (Doc 112) in the corpus; and

(iv) building a model and observing the authorship identification results.

The process was reversed when checking whether the English document can be used to correctly identify the authorship of the French document.

5.2. Large Corpus: Proposed Method Only

Hausdorff Distance Variants. In the first study, we assessed the performance of our method when used with different Hausdorff distance variants described in Section 4.3. Table 12 shows that MHD significantly outperforms all other variants in terms of the fragment accuracy and document accuracy. Recall that the standard Hausdorff distance (SHD) had no outlier handling mechanism. The fact that MHD had outperformed SHD showed that our dataset did in fact has noise (or outliers) to be handled. Further, the fact that MHD had outperformed PHD showed that the former had a better outlier handling mechanism than the latter. Due to the obvious performance gaps, MHD was adopted as the only set distance function for the rest of the studies.

Table 12: Effect of Hausdorff distance variants on accuracy (%)

Method	Fragment Accuracy	Document Accuracy
SHD	75.00	72.78
MHD	94.65	96.66
PHD	52.97	58.00

Language-Pair Identification Checks. Table 13 presents results from each of the 30 identification checks specified in Table 11. As can be seen, all fragment accuracies are greater than 90%, while the document accuracies are mostly 100% except for the identification checks of Finnish → Dutch and German → Spanish. In each of those two cases, the accuracy has dropped to 50%, i.e., one of the two query documents resulted with a misprediction. When we conducted a further investigation, we found that the two query documents were much shorter than other documents in the corpus, i.e., 180,000 tokens and 180,019 tokens, while the average document length is 842,869 tokens. As a result, we could not obtain substantial statistical information to make accurate predictions for those queries.

Table 13: Fragment accuracy (%) & Document accuracy (%) for each cross-lingual identification check type

		Query					
		English	Dutch	Spanish	German	Finnish	French
Ident.	English	-	90.18 & 100	94.27 & 100	96.37 & 100	92.66 & 100	99.74 & 100
	Dutch	93.40 & 100	-	92.40 & 100	95.52 & 100	91.48 & 100	98.10 & 100
	Spanish	94.87 & 100	90.12 & 100	-	96.20 & 100	92.47 & 100	99.69 & 100
	German	95.37 & 100	90.04 & 100	93.83 & 50	-	92.66 & 100	99.74 & 100
	Finnish	94.12 & 100	90.02 & 50	93.61 & 100	95.63 & 100	-	99.31 & 100
	French	96.65 & 100	90.54 & 100	95.10 & 100	98.03 & 100	93.70 & 100	-

We can also see that in terms of the fragment accuracy, we obtain a near-perfect accuracy when the query document is written in French, while Dutch query documents result in a poorer than average fragment accuracy. Moreover, the language of the identifier document marginally affects the fragment accuracy.

5.3. Small Corpus: Comparison

For comparison purposes, we reduced the corpus size from 400 to 24 candidate authors (or less). This was because our competitors [24, 28] were not designed to handle a large number of candidate authors. As for the number of documents per candidate author, we set it to 2, which was approximately the same as that of the large corpus (i.e., 825 documents per 400 authors). Next, we experimented with two and four languages to show the effect of the number $\mathcal{L}$ of languages on the accuracy of each method.

As stated in Section 2, we can decompose the authorship identification into 2 steps: *feature extraction* and *analysis*. Consider our *CLSS* method as an example. We first formulate a 10D feature space called *VSP* and then apply the set similarity PkNN method [36] to the analysis part. Similarly, for the competitor [28], the authors proposed a set of *vocabulary richness (VRFS)* features and applied an array of machine learning algorithms to the extracted VRFS feature vectors.

In this subsection, in addition to directly comparing our proposed method with the competitor [28], we *cross-compare* the feature extraction part and the analysis part of CLSS and VRFS, by formulating the following methods.

- **RF-VRFS**: The VRFS feature space applied to the *Random Forest* method.
- **RF-VSP**: The VSP feature vectors applied to the *Random Forest* method.
- **CLSS-VRFS**: The VRFS feature space applied to our CLSS method.
- **CLSS-VSP** (*proposed method*): The VSP feature space applied to the CLSS method.

Note that the *Random Forest* is used as our comparative classification method due to its superior performance when used with the VRFS feature space reported by Llorens-Salvador and Delany [28]. Our experimental results shown

in Table C.19 also conforms the superior performance of Random Forest when used with VRFS feature space. While training a RF classifier, we capped the number of training samples per class at that of the class with the least samples to avoid the class-imbalance problem. This is because different authors may have a different number of documents, and different documents may have different lengths.

Comparison across different languages of $\mathcal{Q}$. In this phase, we compare the performance of the 4 methods. Fragment accuracy is used for the two CLSS methods, CLSS-VRFS and CLSS-VSP, and the sample accuracy is used for the two RF methods, RF-VRFS and RF-VSP. We omit the document accuracy for conciseness.

As for queries, we used 24 documents from twelve authors written in four different languages. We chose English, Spanish, German, and French for conformity with the existing evaluation [28].

As can be seen from Table 14, the proposed method (CLSS-VSP) has outperformed the competitors significantly. Moreover, regardless of the features set (VSP and VRFS), CLSS outperformed RF as the classifier used in the analysis part of the authorship identification pipeline. As for the feature extraction part, we can also see that VSP has significantly outperformed VRFS. The experimental results show that our proposed solution CLSS-VSP has been the best performer. We can also see that due to the reduced corpus size, the fragment accuracy of CLSS-VSP has increased from 94.65% (as reported in Table 12) to over 99% across 4 different languages.

Varying the number Ω of authors and the number $\mathcal{L}$ of languages. In this study, we show the effect of varying both Ω and $\mathcal{L}$. Let us first consider the effect of the number Ω of authors. As can be seen from Table 15, for any number $\mathcal{L}$ of languages, the accuracy decreases as Ω increases. Specifically, for RF-VRFS and RV-VSP, the accuracy drastically drops by at least 3 folds as Ω increases from 6 to 24. As for CLSS-VSP, on the other hand, we see a slight accuracy drop of 2 percentage points as Ω increases from 6 to 24.

For the effect of the number $\mathcal{L}$ of languages, we can see that the accuracy

Table 14: Accuracy comparison organized by the language of the query document with the number Ω of candidate authors set to 12

	Language of Query Doc. $\mathcal{Q}$			
Method	English	Spanish	German	French
RF-VRFS	12.13	10.80	07.19	12.03
RF-VSP	19.02	19.61	20.49	21.95
CLSS-VRFS	79.04	73.17	78.84	82.99
CLSS-VSP	99.59	99.40	99.75	99.97

drops as $\mathcal{L}$ increases. This effect is consistent across all methods and Ω values. Once again our method, CLSS-VSP, is least affected by the changing $\mathcal{L}$, and is the best performer in all cases.

Table 15: Comparison of fragment/sample accuracy (%) of proposed method against Comparative techniques

$\mathcal{L}$	Method	The number Ω of Authors			
		6	12	18	24
2	RF-VRFS	26.64	13.02	09.09	06.98
2	RF-VSP	30.84	21.11	15.58	7.79
2	CLSS-VRFS	85.03	80.19	77.36	75.21
2	CLSS-VSP	99.84	99.76	98.92	97.89
4	RF-VRFS	24.22	10.53	7.94	5.41
4	RF-VSP	29.01	20.26	13.27	06.89
4	CLSS-VRFS	82.13	78.51	77.30	72.63
4	CLSS-VSP	99.79	99.67	98.83	97.88
6	RF-VRFS	19.18	07.38	04.67	02.83
6	RF-VSP	26.96	18.19	11.74	05.81
6	CLSS-VRFS	80.58	77.97	76.60	69.38
6	CLSS-VSP	99.46	99.20	98.09	97.60

6. Conclusion and Future Work

In this paper, we have presented a scalable method for cross-lingual stylometric analysis. Specifically, we have identified a high-performance language-independent feature set that can be used to accurately identify the authorship of a document in a cross-lingual setting. We have shown that the language-independent features used in this paper have led to the proposed solution outperforming existing state-of-the-art methods. Furthermore, our proposed solution does not rely on any prior knowledge of the languages in the corpus, any machine translation aid, or a part-of-speech tagger. Experimental results have shown that our proposed solution is scalable in terms of the number of languages and the number of candidate authors. We have also demonstrated that our proposed solution can handle a small number of document samples per candidate author. As future work, we plan to apply the proposed solution to other cross-lingual stylometric analysis tasks such as authorship profiling. In addition, provided the relevant dataset, we plan to investigate how the accuracy of cross-lingual author identification will be affected when two languages are very different, for example, in a language pair, one is an Asian language and the other is an European language.

Acknowledgment

This research was partially supported by two CityU research grants (CityU Project No.7200387 and No.6000511).

[1] Abbasi, A., & Chen, H. (2005). Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems*, *20*, 67–75. doi:10.1109/MIS.2005.81.

[2] Arun, R., Saradha, R., Suresh, V., Murty, M., & Madhavan, C. (2009). Stopwords and stylometry: a latent dirichlet allocation approach. In *NIPS workshop on Applications for Topic Models*.

[3] Bay, S. D. (1999). Nearest neighbor classification from multiple feature subsets. *Intelligent data analysis*, *3*, 191–209.

[4] Bogdanova, D., & Lazaridou, A. (2014). Cross-language authorship attribution. In *LREC* (pp. 2015–2020).

[5] Burns, K. (2006). Bayesian inference in disputed authorship: A case study of cognitive errors and a new system for decision support. *Information Sciences*, *176*, 1570–1589. doi:`10.1016/j.ins.2005.04.011`.

[6] Chaski, C. E. (2001). Empirical evaluations of language-based author identification techniques. *Forensic Linguistics*, *8*, 1–65.

[7] Clarke, R., & Lancaster, T. (2006). Eliminating the successor to plagiarism? identifying the usage of contract cheating sites. In *PICAI*.

[8] Drozdz, S., Oswiecimka, P., Kulig, A., Kwapien, J., Bazarnik, K., Grabska-Gradzinska, I., Rybicki, J., & Stanuszek, M. (2016). Quantifying origin and character of long-range correlations in narrative texts. *Information Sciences*, *331*, 32–44. doi:`10.1016/j.ins.2015.10.023`.

[9] Eder, M. (2010). Does size matter? authorship attribution, small samples, big problem. *PDH*, (pp. 132–135). doi:`10.1093/llc/fqt066`.

[10] Escalante, H. J., Solorio, T., & Montes-y Gómez, M. (2011). Local histograms of character n-grams for authorship attribution. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1* (pp. 288–298). Association for Computational Linguistics.

[11] Fujii, R., Domoto, R., & Mochihashi, D. (2017). Nonparametric bayesian semi-supervised word segmentation. *TACL*, *5*, 179–189.

[12] Grant, T. (2007). Quantifying evidence in forensic authorship analysis. *International Journal of Speech, Language & the Law*, *14*.

[13] Grieve, J. (2007). Quantitative authorship attribution: An evaluation of techniques. *LLC*, *22*, 251–270. doi:10.1093/llc/fqm020.

[14] Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, *46*, 389–422.

[15] Hajmohammadi, M. S., Ibrahim, R., & Selamat, A. (2014). Bi-view semi-supervised active learning for cross-lingual sentiment classification. *Inf. Process. Manage.*, *50*, 718–732. doi:10.1016/j.ipm.2014.03.005.

[16] Hajmohammadi, M. S., Ibrahim, R., & Selamat, A. (2014). Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning. *Eng. Appl. of AI*, *36*, 195–203. doi:10.1016/j.engappai.2014.07.020.

[17] Hajmohammadi, M. S., Ibrahim, R., Selamat, A., & Fujita, H. (2015). Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples. *Information Sciences*, *317*, 67–77.

[18] Hajmohammadi, M. S., Ibrahim, R., Selamat, A., & Yousefpour, A. (2014). Combination of multi-view multi-source language classifiers for cross-lingual sentiment classification. In *Intelligent Information and Database Systems - 6th Asian Conference, ACIIDS 2014, Bangkok, Thailand, April 7-9, 2014, Proceedings, Part I* (pp. 21–30). doi:10.1007/978-3-319-05476-6_3.

[19] Hirst, G., & Feiguina, O. (2007). Bigrams of syntactic labels for authorship discrimination of short texts. *Literary and Linguistic Computing*, *22*, 405–417. doi:10.1093/llc/fqm023.

[20] Holmes, C., & Adams, N. (2002). A probabilistic nearest neighbour method for statistical pattern recognition. *J R Stat Soc Series B Stat Methodol*, *64*, 295–306.

[21] Huttenlocher, D. P., Klanderman, G. A., & Rucklidge, W. (1993). Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.*, *15*, 850–863. doi:10.1109/34.232073.

[22] Iqbal, F., Binsalleeh, H., Fung, B. C. M., & Debbabi, M. (2013). A unified data mining solution for authorship analysis in anonymous textual communications. *Information Sciences*, *231*, 98–112. doi:10.1016/j.ins.2011.03.006.

[23] Kešelj, V., Peng, F., Cercone, N., & Thomas, C. (2003). N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING* (pp. 255–264). volume 3.

[24] Kim, Y. (2014). Convolutional neural networks for sentence classification. *CoRR*, *abs/1408.5882*.

[25] Kulig, A., Kwapien, J., Stanisz, T., & Drozdz, S. (2017). In narrative texts punctuation marks obey the same statistics as words. *Information Sciences*, *375*, 98–113. doi:10.1016/j.ins.2016.09.051.

[26] Lertnattee, V., & Theeramunkong, T. (2004). Effect of term distributions on centroid-based text categorization. *Information Sciences*, *158*, 89–115. doi:10.1016/j.ins.2003.07.007.

[27] Lipikorn, R., Shimizu, A., & Kobatake, H. (1994). A modified hausdorff distance for object matching. In *Pattern Recognition* (pp. 566–568). volume 1.

[28] Llorens-Salvador, M., & Delany, S. J. (2016). Deep level lexical features for cross-lingual authorship attribution. In *ECIR* (pp. 16–25).

[29] Luyckx, K., & Daelemans, W. (2010). The effect of author set size and data size in authorship attribution. *LLC*, (pp. 35–55). doi:10.1093/llc/fqq013.

[30] Mao, C., Hu, B., Moore, P., Su, Y., & Wang, M. (2015). Nearest neighbor method based on local distribution for classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 239–250).

[31] Meknavin, S., Charoenpornsawat, P., & Kijsirikul, B. (1997). Feature-based thai word segmentation. In *Proceedings of Natural Language Processing Pacific Rim Symposium* (pp. 41–46). volume 97.

[32] Mendenhall, T. C. (1887). The characteristic curves of composition. *Science*, (pp. 237–249).

[33] Mitchell, T. M. (1997). *Machine Learning*. (1st ed.). New York, NY, USA: McGraw-Hill, Inc.

[34] Mosteller, F., & Wallace, D. (1964). Inference and disputed authorship: The federalist, .

[35] Narducci, F., Basile, P., Musto, C., Lops, P., Caputo, A., de Gemmis, M., Iaquinta, L., & Semeraro, G. (2016). Concept-based item representations for a cross-lingual content-based recommendation process. *Information Sciences*, *374*, 15–31. doi:10.1016/j.ins.2016.09.022.

[36] Nutanong, S., Yu, C., Sarwar, R., Xu, P., & Chow, D. (2016). A scalable framework for stylometric analysis query processing. In *ICDM*.

[37] Orebaugh, A. (2006). An instant messaging intrusion detection system framework: Using character frequency analysis for authorship identification and validation. In *Carnahan Conferences Security Technology, Proceedings 2006 40th Annual IEEE International* (pp. 160–172).

[38] Qian, T., Liu, B., Chen, L., & Peng, Z. (2014). Tri-training for authorship attribution with limited training data. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, Volume 2: Short Papers* (pp. 345–351).

[39] Qian, T., Liu, B., Chen, L., Peng, Z., Zhong, M., He, G., Li, X., & Xu, G. (2016). Tri-training for authorship attribution with limited training data: a comprehensive study. *Neurocomputing*, *171*, 798–806. doi:10.1016/j.neucom.2015.07.064.

[40] Qian, T., Liu, B., Li, Q., & Si, J. (2015). Review authorship attribution in a similarity space. *J. Comput. Sci. Technol.*, *30*, 200–213. doi:10.1007/s11390-015-1513-6.

[41] Qian, T., Liu, B., Zhong, M., & He, G. (2014). Co-training on authorship attribution with very fewlabeled examples: methods vs. views. In *The 37th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '14, Gold Coast , QLD, Australia - July 06 - 11, 2014* (pp. 903–906). doi:10.1145/2600428.2609470.

[42] Qian, T., Xiong, H., Wang, Y., & Chen, E. (2007). On the strength of hyperclique patterns for text categorization. *Information Sciences*, *177*, 4040–4058. doi:10.1016/j.ins.2007.04.005.

[43] Ramnial, H., Panchoo, S., & Pudaruth, S. (2016). Authorship attribution using stylometry and machine learning techniques. In *IJISTA* (pp. 113–125).

[44] Shu, X., Wang, J., Shen, X., & Qu, A. (2017). Word segmentation in chinese language processing. *Statistics and Its Interface*, *10*, 165–173.

[45] Stamatatos, E. (2009). A survey of modern authorship attribution methods. *JASIST*, *60*, 538–556. doi:10.1002/asi.21001.

[46] Steinberger, R., Pouliquen, B., & Hagman, J. (2002). Cross-lingual document similarity calculation using the multilingual thesaurus EUROVOC. In *Computational Linguistics and Intelligent Text Processing* (pp. 415–424). doi:10.1007/3-540-45715-1_44.

[47] Stuart, L. M., Tazhibayeva, S., Wagoner, A. R., & Taylor, J. M. (2013). Style features for authors in two languages. In *IEEE/WIC/ACM* (pp. 459–464). doi:10.1109/WI-IAT.2013.65.

[48] Wu, M. (2015). Modeling query-document dependencies with topic language models for information retrieval. *Information Sciences*, *312*, 1–12. doi:10.1016/j.ins.2015.03.056.

[49] Zhang, X., & LeCun, Y. (2015). Text understanding from scratch. *CoRR*, *abs/1502.01710*.

Appendix A. Performance Comparison Among Stylometric Feature Spaces

In this phase of our study, we investigated the importance of vocabulary richness features for cross-lingual authorship identification. We conducted an experimental study using only the *vocabulary richness features (V)* shown in Table 4 (Feature 1 to Feature 10 of type V). This feature space V did not contain any structural or punctuations-based features. We compared the performance of the feature space V against *our feature space (VSP)* which contained the vocabulary-richness (V), structural (S) and the punctuation (P) based features. For this experimental study, we used our main corpus containing 825 documents written in 6 languages from 400 authors (detailed description of the corpus is given in Tables 7 and 8).

The experimental results are shown in Table A.16. As can been seen, our features space *VSP* outperformed the feature space *V* containing only the *vocabulary richness* features. It can be seen that the *structural* and *punctuation* features can help improve the accuracy of cross-lingual authorship identification. However, using only *vocabulary richness* features has also led to an outperformance over current techniques. Since, the features space VSP outperforms the features space V, for the rest of the experimental studies in this investigation,

we focus on our feature space VSP against the feature space *VRFS* proposed by the *competitive method* [28].

Recall that our feature space VSP relies on only a small set of linguistic assumptions: (i) the ability to tokenize a writing sample into words; (ii) the ability to identify sentence boundaries; and (iii) the use of punctuations. However, the feature space V relies on the first linguistics assumption only. Consequently, this feature space can be used instead of VSP when the other two assumptions do not apply.

We note that, 6 of the features in feature space VSP and all of the features in feature space V rely on the ability to identify word boundaries. This does not pose a challenge for the 6 languages we use in our corpus since words in these languages are separated by whitespace characters. However, this is not the case with Asian languages such as Chinese, Thai and Japanese. In order to apply our proposed method to these languages a more sophisticated method has to be used to identify word boundaries. Several recent developments have achieved state of the art accuracy on word segmentation for Asian languages such as Chinese [44], Thai [31] and Japanese [11]. Due to the cross-lingual nature of this task, our corpus has to contain a substantial number of authors who write in two different languages. Unfortunately, such a corpus is *not* publicly available for any European-Asian language pair. Due to the unavailability of such a corpus for this recently developed area, we consider it for the future work, provided a relevant dataset is available.

Table A.16: Comparison of feature spaces

Method	Accuracy
CLSS-VSP	94.65%
CLSS-V	90.01%
CLSS-VRFS	60.11%

Appendix B. Largescale Experiments

In this section, we present experimental results obtained from large-scale experiments to validate the effectiveness of proposed solution. Due to the cross-lingual nature of this task, it is essential to have documents written by an author in two languages in different language pairs (L_1, L_2). Specifically, for all language pairs (L_1, L_2) we wish to test, there has to be a substantial number of bilingual authors writing in both languages. Unlike in existing studies, we tested all possible language pairs in our original experimental setting. The major obstacle to testing a large number of language pairs is the unavailability of corpora containing a sufficient number of bilingual authors who write in different L_1-L_2 combinations [4]. In order to expand our corpus we reduced the number of tested language pairs by setting English to be one of the languages in all tested language pairs. This testing strategy was also used by the previous works [4, 28]. Specifically, we formulated an additional dataset of 3,000 documents from 1450 authors written in 6 languages. As for the test dataset, we performed testing on 856 documents from 196 authors written in 6 languages. A description of the test dataset is given in Table B.17. In comparison to our main test dataset shown in Table 11, we increased the number of authors from 30 to 196 authors, i.e., a 553% increase and the number of test documents from 60 to 856 documents, i.e., a 1327% increase. In this test dataset, for all 196 authors, English is one of the languages in all language pairs. The experimental results are shown in Table B.18. As can be seen, there is no significant change in the experimental results which in turn validates the effectiveness of our algorithm.

Appendix C. Accuracy: Competitors

In this section, we compare the accuracy of proposed solution CLSS-VSP against several classifiers, namely logistic regression (LR), naive bayes (NB), support vector machines (SVM) and convolution neural networks (CNN). We also compare the proposed method against the main language-independent competitor (RF-VRFS). In addition to applying these classifiers including our language-

Table B.17: Large Scale Experiments (Test dataset description): Data sizes per language in terms of the number of documents, number of fragments, number of chunks, and number of tokens.

Language	#Authors	#Documents	#Fragments	#Chunks	#Tokens
English	196	400	6,729	134,585	201,877,500
Dutch	28	77	1,837	36,756	55,134,000
French	86	180	4,551	91,033	136,549,500
Finnish	29	49	652	13,041	19,561,500
German	33	100	1,642	25,630	38,445,000
Spanish	20	50	739	14,781	22,171,500
Total	196	856	16,150	315,826	473,739,000

Table B.18: Large Scale Experiments (LS): Accuracy comparison organized by the language of the query document.

	Language of Query Doc. $\mathcal{Q}$					
Method	French	Spanish	German	Finnish	English	Dutch
CLSS-VSP (LS)	98.86	96.11	95.78	94.86	93.32	93.11
CLSS-VSP	97.74	94.27	96.37	92.66	94.88	90.18

independent competitor, we *cross-compare* the feature extraction part and the analysis part of VRFS, by formulating the following methods.

- **RF-VRFS**: The VRFS feature space applied to the *Random Forest* method.
- **RF-VSP**: The VSP feature space applied to the *Random Forest* method.
- **SVM-VRFS**: The VRFS feature space applied to the *SVM* method.
- **SVM-VSP**: The VSP feature space applied to the *SVM* method.
- **LR-VRFS**: The VRFS feature space applied to the *LR* method.
- **LR-VSP**: The VSP feature space applied to the *LR* method.
- **NB-VRFS**: The VRFS feature space applied to the *NB* method.
- **NB-VSP**: The VSP feature space applied to the *NB* method.

While training these classifiers, we capped the number of training samples per class at that of the class with the least samples to avoid the class-imbalance

problem. This was because different authors may have different numbers of documents, and different documents may have different lengths.

Similar to other experimental studies in our investigation, for this study we also evaluated the performance of each method by varying the number Ω of authors, the number $\mathcal{L}$ of languages and the feature spaces. Specifically, we reduced the corpus size from 400 to 24 candidate authors and vary the number Ω of authors between 6 and 24. This was because, unlike with our proposed method, the competitive methods are not designed to handle a large number of candidate authors. In terms of the number of documents per candidate author, we set it to 2, which was approximately the same as that of the large corpus (i.e., 825 documents per 400 authors). For the value of $\mathcal{L}$, we varied it between 2 and 4. As shown in Table C.19, the proposed method significantly outperformed all other classifiers. We also note that, the RF classifier had outperformed the CNN, SVM, NB and LR. As a result, we have compared the proposed method against RF in rest of the experimental studies conducted in this investigation.

Table C.19: Comparison of fragment/sample accuracy (%) of Comparative techniques

$\mathcal{L}$	Method	The number Ω of Authors			
		6	12	18	24
2	CNN	27.00	12.50	05.88	02.50
2	SVM-VRFS	21.93	12.76	05.30	05.66
2	SVM-VSP	29.45	18.23	10.93	06.86
2	LR-VRFS	19.43	10.01	06.03	05.89
2	LR-VSP	26.35	17.98	10.46	06.22
2	NB-VRFS	17.52	08.79	04.54	03.57
2	NB-VSP	18.09	11.17	08.13	04.21
2	RF-VRFS	26.64	13.02	09.09	06.98
2	RF-VSP	30.84	21.11	15.58	7.79
2	CLSS-VRFS	85.03	80.19	77.36	75.21
2	CLSS-VSP	99.84	99.76	98.92	97.89
4	CNN	25.00	08.30	05.50	02.27
4	SVM-VRFS	19.66	11.31	05.12	02.41
4	SVM-VSP	27.18	15.79	07.03	06.88
4	LR-VRFS	18.71	07.13	05.61	03.69
4	LR-VSP	26.32	13.81	04.79	04.16
4	NB-VRFS	14.85	05.03	04.96	02.69
4	NB-VSP	17.22	09.61	05.33	03.11
4	RF-VRFS	24.22	10.53	7.94	5.41
4	RF-VSP	29.01	20.26	13.27	06.89
4	CLSS-VRFS	82.13	78.51	77.30	72.63
4	CLSS-VSP	99.79	99.67	98.83	97.88

Vitae

Raheem Sarwar received the MS degree in computer science from Information Technology University, Lahore, Pakistan in 2015. He is currently a PhD student in the Department of Computer Science, City University of Hong Kong. His research interests include stylometry, query optimization, and largescale machine learning.

Qing Li is a Professor in department of Computer Science, City University of Hong Kong. His research interests include object modeling, multimedia databases, social media and recommender systems. He is a Fellow of IET, a senior member of IEEE, a member of ACM SIGMOD and IEEE Technical Committee on Data Engineering. He is the chairperson of the Hong Kong Web Society, and is a steering committee member of DASFAA, ICWL, and WISE Society.

Thanawin Rakthanmanon received the PhD degree in computer science from the University of California, USA. He is an assistant professor in the Department of Computer Engineering, Kasetsart University, Thailand. His research interests include machine learning, data mining, scientific data management, large-scale machine learning and time series analysis.

Sarana Nutanong received the PhD degree in computer science from the University of Melbourne, Australia, in 2010. He is an assistant professor in the Department of Computer Science, City University of Hong Kong. His research interests include scientific data management, data intensive computing, spatial-temporal query processing, and large-scale machine learning.