Native Language Identification of Fluent and Advanced Non-Native Writers

RAHEEM SARWAR, School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Thailand
ATTAPOL T. RUTHERFORD, Department of Linguistics at Faculty of Arts Chulalongkorn University, Thailand
SAEED-UL HASSAN, Department of Computer Science, Information Technology University, Pakistan
THANAWIN RAKTHANMANON, Department of Computer Engineering, Kasetsart University, Thailand and School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Thailand
SARANA NUTANONG, School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Thailand

Native Language Identification (NLI) aims at identifying the *native* languages of authors by analyzing their text samples written in a *non-native* language. Most existing studies investigate this task for educational applications such as *second language acquisition* and require the learner corpora. This article performs NLI in a challenging context of the *user-generated-content* (UGC) where authors are fluent and advanced non-native speakers of a second language. Existing NLI studies with UGC (i) rely on the content-specific/social-network features and may not be generalizable to other domains and datasets, (ii) are unable to capture the variations of the language-usage-patterns within a text sample, and (iii) are not associated with any outlier handling mechanism. Moreover, since there is a sizable number of people who have acquired non-English second languages due to the economic and immigration policies, there is a need to gauge the applicability of NLI with UGC to other languages. Unlike existing solutions, we define a topic-independent feature space, which makes our solution generalizable to other domains and datasets. Based on our feature space, we present a solution that mitigates the effect of outliers in the data and helps capture the variations of the language-usage-patterns within a text sample. Specifically, we represent each text sample as a *point set* and identify the top-*k* stylistically similar text samples (SSTs) from the corpus. We then apply the *probabilistic k nearest neighbors'* classifier on the identified top-*k* SSTs to predict the native languages of the authors. To conduct experiments,

The research was partially supported by the Digital Economy Promotion Agency (project# MP-62-0003), the Thailand Research Funds (MRG6180266, MRG6280175), Office of the Higher Education Commission, and Grants for Development of New Faculty Staff at Chulalongkorn University.

Authors' addresses: R. Sarwar and S. Nutanong, School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Wangchan Valley 555 Moo 1 Payupnai, Wangchan, Rayong, Thailand, 21210; email: {raheem.s, snutanon}@vistec.ac.th; A. T. Rutherford, Department of Linguistics at Faculty of Arts Chulalongkorn University, Phayathai Road, Pathumwan, Bangkok, Thailand; email: attapolrutherford@gmail.com; S.-U. Hassan, Department of Computer Science, Information Technology University, 346-B, Ferozepur Road, Lahore, Punjab, Pakistan; email: saeed-ul-hassan@itu.edu.pk; T. Rakthanmanon (corresponding author), Department of Computer Engineering, Kasetsart University, 50 Thanon Ngamwongwan, Lat Yao, Chatuchak, Bangkok, Thailand, 10900 and School of Information Science and Technology, Vidyasirimedhi Institute of Science and Technology, Wangchan Valley 555 Moo 1 Payupnai, Wangchan, Rayong, Thailand, 21210; email: thanawin.r@ku.ac.th.

2375-4699/2020/04-ART55 $15.00
https://doi.org/10.1145/3383202

we create three new corpora where each corpus is written in a different language, namely, *English*, *French*, and *German*. Our experimental studies show that our solution outperforms competitive methods and reports more than 80% accuracy across languages.

CCS Concepts: • **Computing methodologies** → **Language resources**; **Supervised learning by classification**; *Classification and regression trees*; • **Information systems** → *Content analysis and feature selection*; *Information extraction*; • **Applied computing** → *Investigation techniques*; *Evidence collection, storage and analysis*;

Additional Key Words and Phrases: Author profiling, stylometry, text classification, forensic investigation, native language identification

ACM Reference format:
Raheem Sarwar, Attapol T. Rutherford, Saeed-Ul Hassan, Thanawin Rakthanmanon, and Sarana Nutanong. 2020. Native Language Identification of Fluent and Advanced Non-Native Writers. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 19, 4, Article 55 (April 2020), 19 pages.
https://doi.org/10.1145/3383202

1 INTRODUCTION

Native Language Identification (NLI) aims at identifying the *native language* (L1) of authors by analyzing their text samples written in a *non-native* language (L2). This process relies on the observation that authors sharing the same linguistic background exhibit specific language production and error patterns, in a subsequently learnt L2 language [7, 23]. Given a corpus of the text samples written in a *non-native* (L2) language of the authors labeled with their *native* (L1) languages, the NLI task is generally performed by (i) computing features from the text samples; and (ii) applying a classifier to them to build a model that can identify the native language (L1) of the author of an anonymous text sample.

Most existing studies investigate NLI for educational applications such as second language acquisition [6, 10, 16, 19, 23, 42, 43] and require the learner corpora such as *test of English as a foreign language* [6, 46], *international corpus of learner English* [10], *ASK corpus of learners of Norwegian* [42], and *Jinan Chinese learner* corpus [46]. The second language (L2) learners with different native languages (L1s) make different types of errors[1] [10, 19]. In the context of educational applications, NLI was used to identify L1-specific error patterns in their subsequently learnt languages that can help develop teaching instructions and methods that are specific to their native languages.

The NLI task is not limited to the language of learners. It is relevant also, perhaps even more so, in more challenging context of the fluent and advanced non-native speakers that are likely to produce less errors in comparison to the beginners [9]. While the world has witnessed the ubiquity of English as the *lingua franca*, the native English speakers are outnumbered by the non-native English speakers. Specifically, more than one billion people use English today as their second language and English is the first language of over 400 million people [21]. Consequently, a huge amount of *user-generated-content* (USG), such as blog posts, product reviews, articles, and novels, is continuously being generated by the non-native writers [9, 30]. Therefore, performing NLI with UGC can be useful in several areas such as *forensic linguistics*, author profiling, and authorship identification [9, 18, 29, 30, 34, 37, 38]. For example, in the context of the *forensic linguistics*, a juncture where the linguistic stylistics and the legal system intersect [23], NLI can be considered as a useful tool to provide evidence regarding the linguistic background of an author. That is, there are several situations where a text (e.g., an anonymous letter) is the central evidence in an investigation [23, 34]. In such a situation, the ability to retrieve additional useful information from

[1] Spelling and grammatical mistakes.

an anonymous text can help intelligence agencies and authorities to learn more about the threats and people responsible for them. Specifically, NLI techniques can be used to provide evidence regarding the linguistic background of the author [1, 23] that can help law enforcement agencies identify the source of anonymous text [1].

Nowadays people can be proficient in more than one language [26, 33, 39]. Specifically, it has been estimated that more than half of the people in the world are proficient in two or more languages. While English is the most widely spoken language in the world, there is a sizeable number of people who have acquired non-English second languages due to economic and work-related immigration [23, 26]. Moreover, around 45% of the web-content is written in non-English languages and the number of non-English webpages is rapidly growing [26]. Consequently, there is a substantial need to gauge the applicability of NLI with UGC to other languages. We note that the NLI task with UGC recently received attention by researchers [9, 18], but existing studies are limited to English as L2.[2] To the best of our knowledge, there is no benchmark corpora available to conduct such an investigation. We create three corpora where each corpus is written in a different language, namely, English, French, and German, with same number of L1 classes across the languages and will be made publicly available (see Section 3 for details). These languages are reported among the top five widely spoken languages in the world [26].

Novelty of Our Solution and the Limitations of Existing NLI Studies with UGC.

- ***Feature Space.*** Previous NLI studies with UGC [9, 18, 45] rely on (i) the content-specific features (linguistic features) such as word- and character-based n-grams, which are not likely to generalize to other domains [9] (see Section 2.2 for more details), and (ii) the social network features that are specific to the Reddit dataset[3] only [30] and are hard to generalize to other datasets. These features include karma, average score, average number of submissions, average number of comments, and most popular subreddits [9]. Unlike previous NLI studies with UGC [9, 18, 45], we define a *topic-independent* feature space without relying on the social network features and content-specific features, which makes our solution generalizable to other domains and datasets. Specifically, our feature space is based on (i) *part-of-speech* (POS) n-grams, (ii) function words, (iii) context-free grammar production rules (CFG rules), and (iv) the structure of the text sample such as average sentence length (see Section 4.1 for more details).
- ***Number of Samples per L1 Class.*** Existing NLI studies require a large number of samples for each L1 class. For example, the existing most recent and state-of-the-art NLI study with UGC [9] used a dataset where the average number of writing samples per L1 class is 5,797. However, such a large amount of text sample per L1 class may not be available in many real-world cases. To address this issue, we identify the top-k stylistically similar text samples (SSTs) from the corpus with respect to the given test sample. We then apply the *probabilistic k nearest neighbors* classifier (PkNN) [14] on the identified top-k SSTs to predict the native languages of the authors. The motivation behind adopting PkNN is that it can learn from the limited set of training samples [5]. Moreover, it is an instance-based classifier: It predicts the class of the test sample by comparing it with instances stored in the memory rather than a generalized model [5]. Consequently, there is no information loss through generalization [5].
- ***Language-Usage-Patterns Variations and Outlier Handling.*** The NLI process requires us to capture the variations of the L1-specific *language-usage-patterns* within and across the text samples. One of the main issues associated with existing NLI solutions is that they

[2] Malmasi and Draz applied NLI to other languages [23]. However, their study is limited to learner corpora.
[3] http://cl.haifa.ac.il/projects/L2/.

are unable to capture the variations of language-usage-patterns within a text sample. This is because existing solutions represent each text sample as a data point (vector) in multidimensional space. To capture the language-usage-patterns variations within a text sample, instead of representing each text sample as a point (feature vector), we represent it as a set of points (set of feature vectors) in a multidimensional space (see Section 4.1 for details). As a result, each NLI prediction relies on multiple data points instead of one single data point. We note that representing each text sample as a point set requires a set distance measure such as *standard Hausdorff distance* (SHD) to compute the similarity between two text samples (see Section 4.2 for more details).

Another issue associated with existing NLI solutions is that they are not associated with any outlier handling mechanism [9, 18, 45]. However, the accuracy of some important features (e.g., CFG Rules) predicate upon the availability of the accurate NLP tools, and unfortunately it is not the case for *all the languages* [23], which produces noise in the data and negatively affects the NLI accuracy. As mentioned earlier we adopt P*k*NN classifier to predict the native language of an author. We note that the P*k*NN classifier is also sensitive to outliers in the data. This issue can be addressed by using *partial Hausdorff distance* (PHD) [15] as a set similarity measure between two text samples, that is, associated with an outlier handling mechanism [8, 15, 29, 38] (see Section 4.2 for details). One of the main motivations behind adopting the P*k*NN classifier is that allows us to apply set distance measures associated with outlier handling mechanism to mitigate the effect of outliers in the data, which help to increase the performance of the NLI task. Our extensive experimental studies show that our solution can obtain a high accuracy across languages.

Research Questions. In addition to addressing the aforementioned limitations of existing studies, we aim to answer the following questions in this article.

- ***Research Question 1:*** How important it is to capture the variations of the language-usage-patterns within a text sample in the *native language identification* process.
- ***Research Question 2:*** How much accuracy improvement can be obtained in the *native language identification* process using the set similarity measures associated with outlier handling mechanisms such as PHD, in comparison to the one without outlier handling mechanism such as the standard Hausdorff distance (SHD)?
- ***Research Question 3:*** What is the contribution of each feature type of our feature space in the native language identification process with UGC? Specifically, our feature space contains four types of features that are based on (i) *part-of-speech* (POS) n-grams, (ii) function words, (iii) context-free grammar production rules (CFG rules), and (iv) the structure of the text sample such as average sentence length (see Section 4.1 for more details).
- ***Research Question 4:*** Recall that to make our solution robust and generalizable to other domains and datasets, our feature space relies on the *topic-independent* features only. How robust is our feature space when applied to cross-domain settings and different dataset?

Summary of Contributions.

- We propose a state-of-the-art NLI system that (i) is applicable to different domains/datasets, (ii) is capable of capturing the variations of L1-specific language-usage-patterns within the text sample, (iii) is capable of mitigating the effect of outliers in the data, (iv) can handle large number of L1 classes, and (v) outperforms the previous works with 84.51% accuracy.
- We conduct the first NLI study with UGC to gauge its applicability to other languages including French and German. We create three new NLI corpora where each corpus was

extracted from Project Gutenberg[4] and written in a different language, namely, English, French, and German.
- We perform extensive experimental studies to compare the accuracy of our solution against the existing state-of-the-art NLI solutions with UGC [9] across languages.

The rest of the article is organized as follows. Section 2 reviews previous NLI studies. Section 3 illustrates our new corpora. Section 4 illustrates our solution. Section 5 reports experimental results. Section 6 contains our concluding remarks and recommended future work directions.

2 LITERATURE REVIEW

Natural language processing (NLP) is a subfield of computer science, linguistics, and information engineering, concerned with the interactions between computers and human languages [3, 4, 12, 13, 27, 28, 31, 32, 35, 36, 41, 44, 47]. In context of the objectives of this investigation given in the Introduction, we organize the literature review in following two sections. In Section 2.1, we provide a brief discussion on existing NLI studies with the learners' corpora. In Section 2.2, we provide a detailed discussion on the existing NLI studies with UGC corpora.

2.1 Native Language Identification for Learners

The pioneer NLI study was conducted by Koppel et al. [16] on the *International Corpus of Learning English (ICLE)* [11], which contains text samples written by five groups of English learners who speak Czech, Bulgarian, French, Spanish, and Russian. They applied a *support vector machines* (SVM) classifier for native language identification and reported around 80% accuracy. Several other studies [48, 49] adopted the same experimental setup for NLI. This task became more popular when *Educational Testing Service* (ETS) released the non-native *Test of English as a Foreign Language (TOEFLE 11)* corpus. This corpus is also used by the first NLI shared task [43] and 2017 NLI shared task [24]. The NLI shared task is an NLI competition where different teams of NLI researchers participate to improve the accuracy of the NLI task. The NLI task is also performed on *non-English learner corpora* including ASK corpus of learners of Norwegian [42] for *Norwegian* and Jinan Chinese Learner corpus [46] for *Chinese*, with an accuracy level of 81.8% and 76.5% for Norwegian and Chinese corpora, respectively. All these studies identify the native language of the learners, which is an easier task in comparison to identifying the native language of the advanced non-native speakers of a second language [9]. For brevity, we limit our discussion on NLI for learners, since this investigation focuses on NLI for fluent and advanced writers. The following section reviews the existing relevant NLI studies.

2.2 Native Language Identification for Advanced Authors

There are three recent studies that investigated NLI with UGC, and all of them are limited to English as L2.

- **SVM-WC.** The first study [18] summarizes the shared NLI competition performed on an English corpus based on Facebook comments, written by native speaker of six different Indian languages. The best performance was achieved by the team called *DalTeam* with an accuracy level of 48.80% [17]. They used a linear SVM classification method trained using *Stochastic Gradient Descent* method as available in *scikit-learn*[5] library. They used different types of content-specific features including word n-grams (of order 1–2) and character n-grams (of order 2–5). They adopted a *term frequency–inverse document frequency* (TF-IDF)

[4] https://www.gutenberg.org/.
[5] https://scikit-learn.org/stable/modules/sgd.html.

Table 1. A Comparison between the Competitive Study and Our Investigation where OHM represents Outlier Handling Mechanism and LUP represents a Mechanism to Capture the Variations of the Language-usage-patterns within a Text Sample

Ref.	OHM	LUP	Generalizable to other domains/datasets	Classification Method	Test on L2 Languages	# L1s	Avg. # Authors per L1	Types of Features	# words per text sample	Avg. # texts/L1
[9] LR-UGC	No	No	No	Logistic Regression	English	23	1,500	(i) Content specific (ii) Content independent (iii) Social network	1500	5,797
Our Investigation	Yes	Yes	Yes	P*k*NN	English, French, German.	5	100	Content independent	1000	20

weighting scheme and selected 50,000 features using $\mathcal{X}^2$ feature selection. However, the feature space used by this study is *content specific* and highly domain dependent and may not generalize across domains [9]. We call this solution SVM-WC for short, where SVM denotes the classification method and WC denotes the feature space.

- **LR-W.** Later, Volkova et al. [45] tried to investigate the effect of stylistic, syntactic, and lexical features to identify the foreign languages of the non-native English speakers, not necessarily the native languages, using a corpus of Twitter posts. They used a *logistic regression* (LR) model for classification and report that word uni-grams or tri-grams yield the best results. Moreover, their feature space also contains *content-specific* features such as word n-grams that are not likely to generalize to other domains. We call this solution LR-W for short, where LR denotes the classification method and W denotes the feature space (consists of word tri-grams only).
- **LR-UGC.** Finally, the most recent and state-of-the-art NLI study was conducted by Gili et al. [9], and we extensively compare the accuracy of our solution against LR-UGC only. This is because, similarly to our datasets, the authors of only this study report that the non-native writers in their dataset are highly advanced, almost at the level of native speakers. Moreover, this study is performed on a very large dataset written by 34,511 individuals from 23 different langauge backgrounds (native-langauges) and reports a good accuracy level of 69% [9]. We provide a summary of the comparison between the LR-UGC [9] and our investigation in Table 1. The LR-UGC used three types of features including (i) content-specific, (ii) content-independent, and (iii) social media features, as shown in Table 1. The content-specific features include (i) character tri-grams, (ii) token uni-grams, and (iii) spelling and grammar mistakes. The content-independent features include (i) function word, (ii) POS tri-grams, and (iii) sentence length. The social network features include karma, average score, average # submissions, average # comments, and the most popular reddits. As for the classification model, LR-UGC adopts *logistic regression* (LR) as available in the *scikit-learn library*[6] to identify the native languages of the authors. We note that the *social network features* used by LR-UGC are specific to the *Reddit*[3] dataset only [30]. To test the robustness of their feature types for the domain noise, they evaluated their methods in two different scenarios: For the in-domain scenario, they perform the training and testing on the subreddits related to Europe only, and for the out-of-domain scenario, they train the model on the subreddits related to Europe and test it on the chunks from non-European subreddits, while ensuring that they belong to the different authors.

[6] https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html.

Table 2. A Breakdown of the Three Languages and Their L1 Classes: *Texts* Represents Number of Text Samples in Each L1

English		French		German	
L1	Texts	L1	Texts	L1	Texts
French	32	English	38	English	36
German	31	Dutch	23	French	24
Spanish	23	Finnish	19	Dutch	20
Swedish	7	Portuguese	14	Finnish	12
Norwegian	7	Russian	6	Portuguese	8

Comparison to Our Work. Recall that LR-UGC used three types of features: (i) content-specific, (ii) content-independent, and (iii) social media features, as shown in Table 1. However, our solution relies on the content-independent feature space only, which broadens the scope our solution since it can be generalized to other domains/datasets. The main difference between the content-independent features used by LR-UGC and our work is that, in addition to all the features used by LR-UGC, we use the context-free grammar (CFG) production rules-based features as well. We perform an experimental study to show that using CFG rules-based features improves the accuracy of NLI (see Section 5.2.1).

Unlike LR-UGC, our solution represents each text sample as a *point set*, which help us obtain a high accuracy by (a) mitigating the effect of outliers in the data, and (b) capturing the language-usage-patterns variation within the text sample. Moreover, we test our solution on three different languages including English, French, and German. Recall that this investigation aims at performing NLI with fluent and advanced non-native speakers of a second language. This is a more challenging task to perform in comparison to identifying the native language of the learners where they are likely to make more errors while writing in a second language. Moreover, in our corpus, the average number of samples per class is significantly lower than LR-UGC, and the length of a text sample is lower than LR-UGC, which makes the NLI task more challenging to perform.

3 CORPORA

To the best of our knowledge, there is no benchmark corpora available to conduct such an investigation. The NLI task requires the text samples written in a non-native (L2) language of the authors labeled with their native (L1) languages. We retrieved the corpora from Project Gutenberg.[4] We used *Scrapy* and *Spider* for web crawling, open source frameworks that allows us to locate and extract the data from a web page and process it with external Python scripts. We retrieved three corpora where each corpus is written in a different language, namely, English, French, and German, with the same number of L1 classes across languages as shown in Table 2. Each text sample is written in the second language (L2) of the authors and manually labeled with the native language of these authors. The breakdown of these three languages and their relevant L1 classes are shown in Table 2. Since one of the main objectives of this investigation is to gauge the applicability of the native language identification process to other languages, it requires the balanced corpora where each corpus (written in a different second language) is similar across languages in terms of (i) the number of L1 classes, (ii) the topic of the text samples, (iii) the average number of text samples per L1 class, (iv) the total number of users (writers), and (iv) the total number of text samples. Our corpora have all the aforementioned characteristics. The rest of the characteristics of our corpora are as follows: (i) To avoid the topic-bias in the native language identification process, all the text samples in a corpus are restricted to the same topic across the L1's. (ii) To avoid the influence of same authorship of the text samples in the native language identification process,

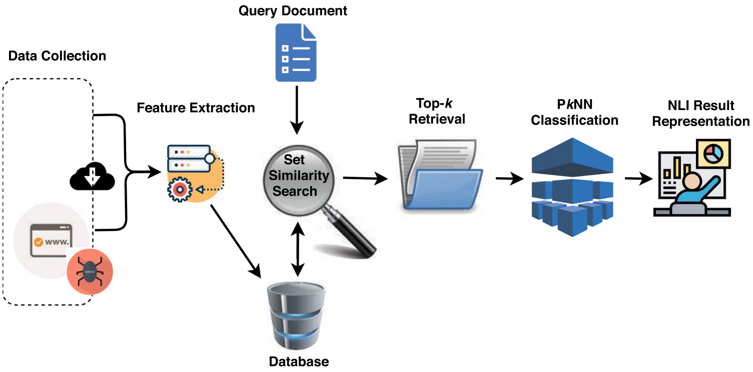

Fig. 1. System overview.

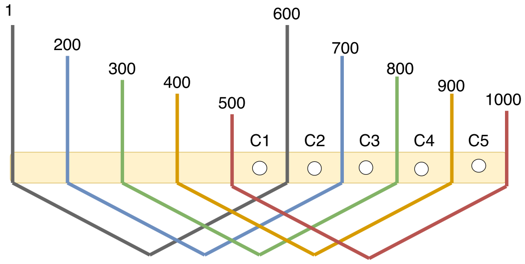

Fig. 2. An example of feature extraction process from a document with sliding window: For features extraction, each text sample is partitioned into *chunks* where the size of each chunk is fixed to 600 tokens, and the size of sliding window is fixed to 100 tokens. Consequently, a 1,000-token text sample results in 5 chunks of 600 tokens each.

each text sample is written by a different author. Note that, unlike the existing method (UGC-NLI) where each text sample contains 1,500 tokens, we limit the length of each text sample to 1,000 tokens only, which makes this task more realistic and challenging to perform.

4 METHOLDOGY

Our solution consists of three main steps, namely, (i) *feature extraction*, (ii) *set similarity search*, and (iii) *probabilistic k nearest neighbor (PkNN) classification*. The overview of our solution is provided in Figure 1.

4.1 Feature Extraction

After collecting the required corpora, we extract features from each text sample. Figure 2 shows that for features extraction, each text sample is partitioned into *chunks*,[7] where the size of each chunk is fixed to 600 tokens. While the text partitioning process, the concept of sliding window is used to generate more chunks from each text, where the size of sliding window is fixed to 100 tokens. Consequently, a 1,000-token text sample results in five chunks. Following the text partitioning process, we extract the features from each *chunk*. As a result, each chunk is represented as a *point*, and each text sample is represented as a *point set* in a multidimensional space. We note that each *text sample* is represented as a *point set* and is the unit for classification. The main motivation for representing each sample as a *point set* is twofold:

[7] Collection of tokens separated by white space without considering the punctuations.

Table 3. Summary of Language Processing Tools and Characteristics

Languages	# Function Words	# POS Tags
English	400[10]	36 [25]
French	463[9]	30 [2]
German	603[9]	55 [40]

(1) To mitigate the effect of outliers in our data. That is, by representing each text sample as a point set, the similarity between two text samples can be computed as a set distance using *partial Hausdorff distance* [15], that is associated with outlier handling mechanism.
(2) To capture the variations of the language-usage-patterns within a text sample.

We extract the following types of features from each chunk using Stanford CoreNLP[8]: (i) *part-of-speech* (POS) n-grams, (ii) function words, (iii) context-free grammar production rules, and (iv) average sentence length. A detailed description of each feature type is as follows.

POS n-grams. An n-gram is a contiguous sequence of *n* items from a text. The *part-of-speech* (POS) are *linguistic categories* associated with words that signify their syntactic roles in the text sample, e.g., nouns, verbs, and adjectives [22]. The POS n-grams are considered as content-independent features and identify the preferences of word categories and the *local syntactic patterns* of the *language usage* from text samples that help differentiate among the groups of authors with respect to their native languages [9, 16, 22]. We calculate the POS n-grams of order 1–5 from each chunk. Specifically, the value of a certain POS n-gram is calculated as the ratio of its occurrences in the chunk and the total number of n-grams in the chunk. For English, we use *Penn Treebank*, which classifies the words into 36 linguistic categories [25]. For the French and German, we use *French Treebank* [2] and *Stuttgart/Tübinger* [40], which classify words into 30 and 55 linguistic categories, respectively (see Table 3 for more details).

Function Words. The *function words*–based features are considered as highly topic independent [9, 16, 22]. The examples of the function words include conjunctions, determiners, articles, and auxiliary verbs. Unlike content words, the function words do not have meaning themselves and actually indicate the grammatical relations between other words. We obtain the function words lists for all languages from the multilingual IR resources[9] except English, which is obtained from *Onix Text Retrieval Toolkit* [10] (see Table 3 for more details). We use the normalized frequency of a function word as the feature value.

Context-free Grammar Production (CFG) Rules. The CFG rules (phrase structure rules) are used to generate the constituent parts of the sentences from each chunk such as noun phrases. The CFG rules help to *capture* the global syntactic patterns and the structure of grammatical constructions that help identify the native langauges of the authors [49]. To extract the CFG rules, at first, we extract the constituent parses for all sentences in a chunk and then extract the production rules without lexicalizations. Each rule is considered as a classification feature.

Structural Feature. The non-native speakers are likely to produce simple and shorter sentences compare to the native speakers [9]. We compute the average length of sentences from each chunk

[8] http://nlp.stanford.edu/software/corenlp.shtml.
[9] http://members.unine.ch/jacques.savoy/clef/index.html.
[10] http://www.lextek.com/manuals/onix/stopwords1.html.

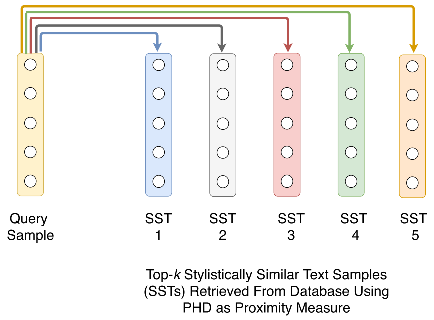
Query
Sample
SST
1
SST
2
SST
3
SST
4
SST
5
Top-*k* Stylistically Similar Text Samples
(SSTs) Retrieved From Database Using
PHD as Proximity Measure

Fig. 3. Set similarity search to retrieve top-*k* stylistically similar text samples (SSTs) from database: The rectangle represents a text sample and circles in a rectangle represent data points (chunks).

as a feature for native language identification. After completing the feature extraction process, we store the feature vectors into the database.

4.2 Set Similarity Search

Given a query text sample $\mathcal{T}$, we extract the features from $\mathcal{T}$ using the same procedure as described in Section 4.1. As a result, $\mathcal{T}$ is transformed into a *point set*. We then execute a *set similarity query* to retrieve top-*k* *stylistically similar text samples* (SSTs) from the database. While retrieving the top-*k* SSTs, we tried two set similarity measures, (i) *standard Hausdorff Distance (SHD)* and (ii) *partial Hausdorff Distance (PHD)* [15], as a proximity measure between two point sets. The SHD between two points sets $\mathcal{T}$ and U can be calculated by:

$$SHD(\mathcal{T}, U) = \max_{t_i \in \mathcal{T}} \min_{u_j \in U} d(t_i - u_j).$$

That is, the SHD between $\mathcal{T}$ and U can be calculate by (i) ranking all data points in a test sample $\mathcal{T}$ in accordance with the minimum distance to the text sample U and (ii) selecting the maximum of the of the minimum distances. Researchers have argued that SHD is sensitive to outliers. To mitigate the outlier sensitivity issue associated with SHD, researchers formulated a variant of SHD known as *partial Hausdorff distance (PHD)* [15]. Assume that there are two point sets $\mathcal{T}$ and U. The PHD between $\mathcal{T}$ and U can be calculated by (i) ranking all data points in a test sample $\mathcal{T}$ in accordance with the minimum distance to the text sample U and (ii) computing the average of minimum distances within a given range, i.e., $(50\%, 75\%]$ [20] (cf. Algorithm 1). This step of our solution retrieves a set of top-*k* SSTs from the database with respect to the query sample $\mathcal{T}$. We note that this step retrieves a set of top-*k* SSTs (instead of chunks) (see Figure 3).

4.3 Probabilistic *k* Nearest Neighbor Classification

A simple *k*NN classifier can be used to predict the native language of an author. Instead of using *k*NN classifier, we use probabilistic *k*NN classifier (P*k*NN) [14], which aims to identify the *native language likelihood* by determining the PMF[11] over a *set of likely native languages* of a query sample. This step of our solution produces a probabilistic NLI prediction for a query sample. As a result, we can determine the uncertainty of the probabilistic prediction, using entropy as an uncertainty measure. Consequently, based on the entropy value, we can analyze whether to use the prediction.

[11] Probability Mass Function.

ALGORITHM 1: PHD calculations

```
procedure PHD($\mathcal{T}$ , $U$)
    $MinDists \leftarrow []$
    $PHDDist \leftarrow 0$
    for $t$ in $\mathcal{T}$ do
        $dmin \leftarrow \infty$
        for $u$ in $U$ do
            $dist \leftarrow dist(t, u)$
            if $dist < dmin$ then
                $dmin \leftarrow$ dist
            end if
        end for
        $MinDists.Append(dmin)$
    end for
    $MinDists \leftarrow Sort(MinDists)$
    $PHDDist \leftarrow ComputeAvg[MinDists, (50\%, 75\%]]$
    return $PHDDist$
end procedure
```

The experimental results given in Section 5.2.1 show that excluding the uncertain predictions help to increase the accuracy of the NLI task.

The motivation to adopt P*k*NN is that a little or no training is required since it performs classification through a comparison with text samples stored in the memory rather than a generalized model. Consequently, there is no information loss through generalization. Furthermore, by using our document representing model, P*k*NN allows us to use variety of set similarity measures including those having outlier handling techniques associated with them such as PHD, which in turn help increase the performance of NLI task.

5 PERFORMANCE EVALUATION

In this section, we describe the experimental setup and report the results from our extensive experimental studies.

5.1 Experimental Setup

In this subsection, we illustrate the evaluation measures, parameter settings, and evaluation strategy.

Evaluation Measure. We use accuracy and F1 score as the evaluation measures, which can be defined as follows.

- **Accuracy.** A query sample is correctly predicted if the correct *native-language* is identified as the *most likely native-language* of the query sample.
- **F1 Score.** The F score is defined as the weighted harmonic mean of the *precision* and *recall.*

Parameter Setting. As for the parameters, although not shown here, we have tested different values for each parameter. The parameter values given in Table 4 resulted in the best accuracy. The *k* value denotes the number of closest text samples relating to query sample to use for P*k*NN. The value (50%, 75%], denote the PHD range. That is, to measure the PHD value between two point sets, we average the ranked distances falling in this specified range. The *L* value denotes the size

Table 4. Default Parameters Setting

k	PHD	L	l
5	(50%, 75%]	1,000 tokens	100 tokens

Table 5. Our Solution: Effect of Capturing the Variations of the Language-usage-patterns within a Text Sample

Method	Accuracy		
	English	French	German
XNLI-PFCS (SHD)	**69.91%**	**67.76%**	**69.18%**
Baseline	62.28%	61.58%	63.81%

of the chunk, i.e., 600 tokens. The value of parameter l represents the size of the sliding window, i.e., 100 tokens.

Evaluation Strategy: In each experimental study, the results are reported as the average accuracy obtained using *fivefold* cross-validation.

5.2 Experimental Results

In context of the main objectives and research questions of this investigation listed in the Introduction, we divide this subsection into three parts. The first part presents the experimental studies associated with our solution. In the second part, we provide the performance comparison between our solution and the existing state-of-the-art solution (LR-UGC) reported in References [9]. In third part, we provide a detailed comparison between our solution and the existing NLI solutions with UGC explained in Section 2.2.

5.2.1 *Experimental Results: Our Solution Only.*

This part of the article presents the experimental studies associated with our solution only.

Variations of the Language-usage-patterns. The main objective of this study is to show the effect of capturing the variations of the language-usage-patterns within a text sample. Recall that to capture the variations of the language-usage-patterns within a text sample, instead of representing each text sample as a point (feature vector), we represent it as a set of points (set of feature vectors) (see Section 4.1 for details). As a result, each NLI prediction relies on multiple data points instead of one single data point. We note that representing each text sample as a point set requires a set distance measure to compute the similarity between two text samples. In this study, we use *standard Hausdorff distance* (SHD) as a proximity measure between two point sets. To show the effect of capturing the variations of the language-usage-patterns within a text sample, we formulate a baseline method and compare its performance against our solution (XNLI-PFCS). The difference between the baseline method and our method (XNLI-PFCS) is that the former represents each text sample as a one single data point and the latter represents each text sample as a set of points. The experimental results given in Table 5 show that our method (XNLI-PFCS) is capable of capturing the language-usage-patterns within a text sample and thus outperforms the baseline method across the languages.

Outlier Handling Mechanism. The objective of this study is to show the effect of outliers in the data on the accuracy of the *native language identification* (NLI) process. Specifically, we provide the accuracy comparison between two set distance measures: (i) *standard Hausdorff distance (SHD)*

Table 6. Our Solution: The Effect of Outlier Handling Mechanism Associated with Partial Hausdorff Distance (PHD)

Set Distance Measure	Accuracy		
	English	French	German
XNLI-PFCS (PHD)	**83.29%**	**82.44%**	**84.51%**
XNLI-PFCS (SHD)	69.91%	67.76%	69.18%

Table 7. Our Solution: Effect of Feature Types

	Feature Types				Accuracy		
	POS n-grams (1–5)	Function Words	CFG Rules	Avg. Sentence Length	English	French	German
	–	✓	✓	✓	59.18%	57.22%	55.47%
	✓	–	✓	✓	62.79%	60.07%	59.89%
	✓	✓	–	✓	71.83%	73.68%	75.83%
	✓	✓	✓	–	78.11%	79.02%	81.79%
Combined	✓	✓	✓	✓	**83.29%**	**82.44%**	**84.51%**

and (ii) *partial Hausdorff distance (PHD)*, where the former is not associated with outlier handling mechanism, and the latter is associated with outlier handling mechanism (i.e., PHD). The experimental results given in Table 6 show that PHD outperforms the SHD. This is due to the fact that our dataset has noise to be handled, and using PHD, which is associated with outlier handling mechanism, improves the accuracy of the NLI process. Since the PHD measure provides a better performance in comparison to the SHD measure across the three languages, we will confine the rest of the studies to PHD only.

Feature Evaluation. In this study, we evaluated the contribution of each feature type based on (i) *part-of-speech* (POS) n-grams, (ii) function words, (iii) context-free grammar rules, and (iv) average sentence length in the *native language identification* (NLI) process across the languages. We hypothesize that combining all features into one feature vector will outperform other feature combinations. For other combinations, we remove one feature at a time from combined feature set and report the NLI accuracy. It enables us to observe the contribution of each feature to the NLI process. Table 7 shows that our most important type of feature is based on *part-of-speech* (POS) n-grams. That is, excluding the POS-based features from our feature space (combined feature set) significantly drops the NLI accuracy compared to other types of features. The function words-based features take the second rank in terms of their contribution in the NLI process. Furthermore, the features based on the CFG rules and average sentence length takes third and fourth ranks, respectively. It can also be seen that the combined feature types (i.e., using all types of features) outperforms the other feature types combinations across three languages. It indicates that the stylistic information captured by some of our features is complementary and orthogonal. Consequently, combining these feature sets improved the performance of the *NLI* task. Since the *combined feature types* provides a better performance across the three languages, all other experimental studies are based on the *combined feature types* only.

Comparison among L2's in Terms of Their L1's. In this study, we assess the performance of our method across languages in terms of their L1 classes. Table 8 shows that the F1 scores of our method is higher than 80% across three languages in terms of their L1 classes.

Table 8. The Performance of Our Method among L2's in Terms of Their L1

English		French		German	
L1	F1 Score	L1	F1 Score	L1	F1 Score
French	0.84	English	0.84	English	0.85
German	0.84	Dutch	0.83	French	0.83
Spanish	0.83	Finnish	0.81	Dutch	0.85
Swedish	0.86	Portuguese	0.83	Finnish	0.87
Norwegian	0.86	Russian	0.83	Portuguese	0.88

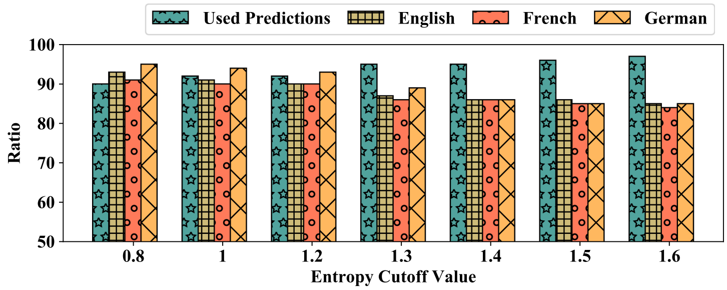

Fig. 4. Used prediction ratio and NLI accuracy.

Entropy Analysis of the *k*NN Predictions. Recall that instead of using simple *k*NN classifier, we adopt the P*k*NN so that we can measure the prediction uncertainty using entropy as an uncertainty measure. As a result, based on the entropy value, we can decide whether to use the prediction. We show the used prediction ratio and the NLI accuracy for different entropy cutoff values in Figure 4. We use the result of a prediction if the entropy of this prediction is lower than the cutoff value. We used the same entropy cutoff value for all the languages. We can see from Figure 4 that increasing the cutoff value decreases the number of used predictions while the NLI accuracy increases, across the languages. For English as an example, when the cutoff value is 0.8, the used prediction ratio is 90% while the NLI accuracy is 93.89%. However, using all the predictions results with 83.29% accuracy. We can find the similar trend across the languages in our corpora. Thus, ruling out the uncertain predictions based on entropy help to increase the accuracy of our method.

5.2.2 Comparison between Our Method and LR-UGC. In this part of the article we compare the accuracy of our solution against the existing state-of-the-art solution (LR-UGC) [9] with the help of three studies. The first study is performed on our corpora where each corpus is written in different language: English, German, and French. The second study is performed on the *Reddit dataset* used by LR-UGC. However, this corpus is limited to English only. The third study evaluates both of our methods in cross-corpus settings where the model training is performed on one dataset and the testing is performed using other dataset.

Comparison on Our Dataset. Recall that our feature space contains *content-independent* features only and the length of each text sample is fixed to 1,000 tokens. However, the LR-UGC feature space contains three types of features: (i) content-specific, (ii) content-independent, and (iii) social network features. Moreover, for LR-UGC, the size of each text sample is 1,500 tokens. Despite the fact that the LR-UGC takes the advantage of *longer* text samples and the *content-specific* features (which are not likely to generalize to other domains), our method outperforms the competitive technique, as can be seen from Table 9. We note that the experimental results associated with LR-UGC are

Table 9. Our Dataset: Performance Comparison between Our Solution and LR-UGC

Methods	Accuracy		
	English	French	German
XNLI-PFCS (Our Method)	**83.29%**	**82.44%**	**84.51%**
LR-UGC	53.57%	50.04%	51.97%

Table 10. Reddit Dataset: Performance Comparison between Our Solution and LR-UGC

Methods	Accuracy	
	In-domain	Out-of-domain
XNLI-PFCS (Our Method)	**84.97%**	**83.91%**
LR-UGC	77.69%	62.77%

obtained without the *social network features*, since they are specific to the *Reddit* dataset and cannot be computed from our corpus, which limits the scope of LR-UGC to the Reddit dataset only.

Table 9 shows that our method can achieve high accuracy (i.e., more than 80% across languages) and outperforms LR-UGC. This is due to the fact that, unlike LR-UGC, our method is capable of (i) capturing the variations of language-usage-patterns of an author within a text sample, since each prediction is based on multiple data points, and (ii) mitigating the effect of outliers in the data with the help of set similarity measures associated with outlier handling mechanism. Moreover, unlike LR-UGC, our feature space contains CFG Rules-based features, which play an important role in improving the NLI accuracy (see the experimental results given in Table 7).

Comparison on Reddit Dataset. Recall that the *Reddit dataset* contains more L1 classes (i.e., 23) than our corpora as shown in Table 1. Consequently, by applying our technique on Reddit dataset we can evaluate whether our method can handle a large number of L1 classes. Moreover, the text samples in Reddit dataset are categorized in different domains (i.e, related to European and non-European). Consequently, by applying our method on the Reddit dataset in cross-domain settings, we can evaluate whether our solution, which relies on the topic-independent features only, is robust and can achieve a high accuracy level, i.e., similarly to accuracy reported on our corpora, which is topic controlled. Specifically, two evaluation scenarios are used in this study: (i) in-domain and (ii) out-of-domain, which are explained in Section 2.2. The experimental results are given in Table 10. The findings of this study are twofold: (1) Our method outperforms the competitive technique for both of the scenarios, (i) in-domain and (ii) out-of-domain, which indicates that our solution is more robust in comparison to the competitive study, and (2) our method can handle the dataset with large number of L1 (i.e., 23 different L1's). Note that, in this study, the LR-UGC method takes the advantage of the social network features, which are specific to Reddit dataset and cannot be computed from other datasets. We note that, to perform fair comparison between our method and the LR-UGC, no data down sampling is performed in any experiment of this article. It enables us to evaluate all the methods in more challenging scenarios where (i) the number of users (writers) are not evenly distributed among the L1 classes and (ii) some of the users are over-represented in the corpus in terms of their text samples.

Cross-corpus Comparison between Our Method and LR-UGC. In this study, we further evaluate the robustness of our method by conducting two experiments. In the first experiment, we train the model using our corpus (source corpus) and test it on the Reddit corpus (target corpus). In the second experiment, we swap the source and the target corpora. The experimental results

Table 11. Cross-corpus: Performance Comparison between Our Solution and LR-UGC

Methods	Accuracy	
	Source = Our Dataset, Target = Reddit Dataset	Source = Reddit Dataset, Target = Our Dataset
XNLI-PFCS (Our Method)	**83.07%**	**84.91%**
LR-UGC	40.18%	49.82%

Table 12. The Performance Comparison between Our Solution and Competitive Methods

Methods	Feature Spaces			
	PFCS	**WC**	**W**	**UGC**
XNLI	83.29%	71.26%	68.29%	75.98%
SVM	59.79%	52.09%	51.37%	52.44%
LR	58.06%	50.79%	52.17%	53.57%

are given in Table 11. Note that the set of L1's in our English corpus is the subset of L1's in the Reddit dataset. For this study, we reduce the Reddit dataset such that both corpora have the same L1's. Table 11 shows that our method outperforms the LR-UGC method. We note that in this study the social network features cannot be computed, which is one of the main reasons behind the large accuracy gap between the two methods, along with the other advantages associated with our method, such as (i) the ability to mitigate the effect of outliers in the data, (ii) the ability to capture the variations of the L1 language-usage-patterns within a text sample, and (iii) additional features in our feature space.

5.2.3 Detailed Comparison between Our Method and Competitive Techniques. In this section, we compare the performance of our solution (XNLI-PFCS) against three competitive studies explained in Section 2.2: (i) SVM-WC [17], (ii) LR-W [45], and (iii) LR-UGC [9]. Specifically, we cross-compare the feature extraction part and the classification part of all competitive studies against our method as shown in Table 12. The findings of this study are twofold:

(1) Our method (XNLI) outperforms other methods. This is due to the fact that, unlike competitive methods (i.e., SVM and LR) that represent each text sample as a point, our method (XNLI) represents each text sample as a set of points. As a result, our method (XNLI) is capable of (i) capturing the variations of L1-specific language-usage-patterns within a text sample, since each prediction is based on multiple data points, and (ii) mitigating the effect of outliers in the data with the help of set similarity measures associated with outlier handling mechanism.

(2) Our feature space (PFCS) reports higher accuracy than all other feature spaces. This is because, unlike WC and W feature spaces, our feature space contains CFG Rules-based features, POS n-grams-based features, and structural features, which play an important role in improving NLI accuracy (see the experimental results given in Table 7). Moreover, as for the UGC feature space, it does not contain the CFG-rules-based features, and some of the features in UGC, such as social network features, are specific to the Reddit dataset and cannot be computed from our dataset.

6 CONCLUSIONS

This article performs *native language identification* in a challenging context of the fluent and advanced non-native speakers of English and gauges its applicability to other languages such as

French and German. To conduct such an investigation, we create three new corpora where each corpus is written in a different language, namely, English, French, and German. Unlike existing solutions, we define a topic-independent feature space without relying on the social network features and content-specific features, which makes our solution generalizable to other domains and datasets. Based on our feature space, we present a solution that transforms each text sample into a *point set* and adopts the *probabilistic k nearest neighbors* classifier to predict the native languages of the authors. Our experimental studies show that our solution (i) can mitigate the effect of noise in the data, (ii) can capture the variations of the language-usage-patterns within the text sample, (iii) can handle small number of samples per L1, and (iv) significantly outperforms the existing NLI studies across languages. A straightforward future research direction is the extension of our experimental studies to additional languages, provided the relevant corpora. It can help us to gain better insights into the differences among different L1–L2 language pairs.

REFERENCES

[1] Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intell. Syst.* 20, 5 (2005), 67–75.

[2] Anne Abeillé, Lionel Clément, and François Toussenel. 2003. Building a treebank for french. In *Treebanks*. 165–187.

[3] Sophia Ananiadou, Paul Thompson, and Raheel Nawaz. 2013. Enhancing search: events and their discourse context. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'13)*. 318–334.

[4] Riza Theresa Batista-Navarro, Georgios Kontonatsios, Claudiu Mihaila, Paul Thompson, Rafal Rak, Raheel Nawaz, Ioannis Korkontzelos, and Sophia Ananiadou. 2013. Facilitating the analysis of discourse phenomena in an interoperable NLP platform. In *Proceedings of the 14th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing'13)*. 559–571.

[5] Stephen D. Bay. 1999. Nearest neighbor classification from multiple feature subsets. *Intell. Data Anal.* 3, 3 (1999), 191–209.

[6] Daniel Blanchard, Joel Tetreault, Derrick Higgins, Aoife Cahill, and Martin Chodorow. 2013. TOEFL11: A corpus of non-native english. *ETS Res. Rep. Ser.* 2013, 2 (2013), i–15.

[7] Lingzhen Chen, Carlo Strapparava, and Vivi Nastase. 2017. Improving native language identification by using spelling errors. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL'17)*. 542–546.

[8] Marie-Pierre Dubuisson and Anil K. Jain. 1994. A modified hausdorff distance for object matching. In *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Conference A: Computer Vision & Image Processing (ICPR'94)*. 566–568.

[9] Gili Goldin, Ella Rabinovich, and Shuly Wintner. 2018. Native language identification with user generated content. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 3591–3601.

[10] Sylviane Granger. 2003. The international corpus of learner english: a new resource for foreign language learning and teaching and second language acquisition research. *Tesol Quart.* 37, 3 (2003), 538–546.

[11] Sylviane Granger. 2009. The contribution of learner corpora to second language acquisition and foreign language teaching. *Corpora and language teaching* 33 (2009), 13–32.

[12] Saeed-Ul Hassan, Naif R. Aljohani, Nimra Idrees, Raheem Sarwar, Raheel Nawaz, Eugenio Martínez-Cámara, Sebastián Ventura, and Francisco Herrera. 2019. Predicting literature's early impact with sentiment analysis in twitter. *Knowledge-Based Systems* 192 (2019), 105383.

[13] Saeed-Ul Hassan, Raheem Sarwar, and Amina Muazzam. 2016. Tapping into intra-and international collaborations of the organization of islamic cooperation states across science and technology disciplines. *Sci. Publ. Policy* 43, 5 (2016), 690–701.

[14] C. C. Holmes and N. M. Adams. 2002. A probabilistic nearest neighbour method for statistical pattern recognition. *J. Roy. Stat. Soc. Ser. B* 64, 2 (2002), 295–306.

[15] Daniel P. Huttenlocher, Gregory A. Klanderman, and William Rucklidge. 1993. Comparing images using the hausdorff distance. *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 9 (1993), 850–863.

[16] Moshe Koppel, Jonathan Schler, and Kfir Zigdon. 2005. Determining an author's native language by mining a text for errors. In *Proceedings of the Eleventh ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*. 624–628.

[17] Dijana Kosmajac and Vlado Keselj. 2017. DalTeam@ INLI-FIRE-2017: Native language identification using SVM with SGD training. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'17)*. 118–122.

[18] M. Anand Kumar, Barathi Ganesh H. B., Shivkaran Singh, Soman K. P, and Paolo Rosso. 2017. Overview of the INLI PAN at FIRE-2017 track on indian native language identification. In *Proceedings of the Forum for Information Retrieval Evaluation (FIRE'17)*. 99–105.

[19] Claudia Leacock, Martin Chodorow, Michael Gamon, and Joel R. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners*. Morgan & Claypool Publishers.

[20] Rajalida Lipikorn, Akinobu Shimizu, and Hidefumi Kobatake. 1994. A modified hausdorff distance for object matching. In *Pattern Recognition*, Vol. 1. 566–568.

[21] Dylan Lyons. 2018. How many people speak english, and where is it spoken. Retrieved November 11, 2018 from https://www.babbel.com/en/magazine/how-many-people-speak-english-and-where-is-it-spoken.

[22] Shervin Malmasi and Mark Dras. 2015. Large-scale native language identification with cross-corpus evaluation. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 1403–1409.

[23] Shervin Malmasi and Mark Dras. 2017. Multilingual native language identification. *Nat. Lang. Eng.* 23, 2 (2017), 163–215.

[24] Shervin Malmasi, Keelan Evanini, Aoife Cahill, Joel R. Tetreault, Robert A. Pugh, Christopher Hamill, Diane Napolitano, and Yao Qian. 2017. A report on the 2017 native language identification shared task. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*. 62–75.

[25] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: the penn treebank. *Comput. Ling.* 19, 2 (1993), 313–330.

[26] Fedelucio Narducci, Pierpaolo Basile, Cataldo Musto, Pasquale Lops, Annalina Caputo, Marco de Gemmis, Leo Iaquinta, and Giovanni Semeraro. 2016. Concept-based item representations for a cross-lingual content-based recommendation process. *Inf. Sci.* 374 (2016), 15–31.

[27] Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2012. Identification of manner in bio-events. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*. 3505–3510.

[28] Raheel Nawaz, Paul Thompson, and Sophia Ananiadou. 2013. Negated bio-events: Analysis and identification. *BMC Bioinf.* 14, 1 (2013), 14.

[29] Sarana Nutanong, Chenyun Yu, Raheem Sarwar, Peter Xu, and Dickson Chow. 2016. A scalable framework for stylometric analysis query processing. In *Proceedings of the IEEE 16th International Conference on Data Mining (ICDM'16)*. IEEE, 1125–1130.

[30] Ella Rabinovich, Yulia Tsvetkov, and Shuly Wintner. 2018. Native language cognate effects on second language lexical choice. *Trans. Assoc. Comput. Linguist.* 6 (2018), 329–342.

[31] Fahad Sabah, Saeed-Ul Hassan, Amina Muazzam, Sehrish Iqbal, Saira Hanif Soroya, and Raheem Sarwar. 2019. Scientific collaboration networks in pakistan and their impact on institutional research performance: A case study based on scopus publications. *Libr. Hi Tech* 37, 1 (2019), 19–29.

[32] Raheem Sarwar and Saeed-UI Hassan. 2015. A bibliometric assessment of scientific productivity and international collaboration of the islamic world in science and technology (S&T) areas. *Scientometrics* 105, 2 (2015), 1059–1077.

[33] Raheem Sarwar, Qing Li, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. A scalable framework for cross-lingual authorship identification. *Inf. Sci.* 465 (2018), 323–339.

[34] Raheem Sarwar and Sarana Nutanong. 2016. The key factors and their influence in authorship attribution. *Res. Comput. Sci.* 110 (2016), 139–150.

[35] Raheem Sarwar, Thanasarn Porthaveepong, Attapol Rutherford, Thanawin Rakthanmanon, and Sarana Nutanong. 2019. StyloThai: A scalable framework for stylometric authorship identification of thai documents. *ACM Trans. Asian Low-Res. Lang. Inf. Process.* 19, 3 (2019), 36:1–36:15.

[36] Raheem Sarwar, Saira Hanif Soroya, Amina Muazzam, Fahad Sabah, Sehrish Iqbal, and Saeed-Ul Hassan. 2019. A bibliometric perspective on technology-driven innovation in the gulf cooperation council (GCC) countries in relation to its transformative impact on international business. In *Technology-Driven Innovation in Gulf Cooperation Council (GCC) Countries: Emerging Research and Opportunities*. IGI Global, 49–66.

[37] Raheem Sarwar, Norawit Urailertprasert, Nattapol Vannaboot, Chenyun Yu, Thanawin Rakthanmanon, Ekapol Chuangsuwanich, and Sarana Nutanong. 2020. *CAG*: Stylometric authorship attribution of multi-author documents using a co-authorship graph. *IEEE Access* 8 (2020), 18374–18393.

[38] Raheem Sarwar, Chenyun Yu, Sarana Nutanong, Norawit Urailertprasert, Nattapol Vannaboot, and Thanawin Rakthanmanon. 2018. A scalable framework for stylometric analysis of multi-author documents. In *Proceedings of the 23rd International Conference on Database Systems for Advanced Applications (DASFAA'18)*. 813–829.

[39] Raheem Sarwar, Chenyun Yu, Ninad Tungare, Kanatip Chitavisutthivong, Sukrit Sriratanawilai, Yaohai Xu, Dickson Chow, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. An effective and scalable framework for authorship attribution query processing. *IEEE Access* 6 (2018), 50030–50048.

[40] Anne Schiller, Simone Teufel, and Christine Thielen. 1995. Guidelines fur das tagging deutscher textcorpora mit STTS. Technical Report. Universität Stuttgart and Universität Tübingen, Germany.

[41] Matthew Shardlow, Riza Batista-Navarro, Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2018. Identification of research hypotheses and new knowledge from scientific literature. *BMC Med. Inf. Decis. Making* 18, 1 (2018), 1–13.

[42] Kari Tenfjord, Paul Meurer, and Knut Hofland. 2006. The ASK corpus - a language learner corpus of norwegian as a second language. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC'06).* 1821–1824.

[43] Joel R. Tetreault, Daniel Blanchard, and Aoife Cahill. 2013. A report on the first native language identification shared task. In *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications*. 48–57.

[44] Paul Thompson, Raheel Nawaz, John McNaught, and Sophia Ananiadou. 2017. Enriching news events with meta-knowledge information. *Lang. Resourc. Eval.* 51, 2 (2017), 409–438.

[45] Svitlana Volkova, Stephen Ranshous, and Lawrence Phillips. 2018. Predicting foreign language usage from english-only social media posts. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 608–614.

[46] Maolin Wang, Shervin Malmasi, and Mingxuan Huang. 2015. The jinan chinese learner corpus. In *Proceedings of the 10th Workshop on Innovative Use of NLP for Building Educational Applications*. 118–123.

[47] Xinglong Wang, Rafal Rak, Angelo C. Restificar, Chikashi Nobata, C. J. Rupp, Riza Theresa Batista-Navarro, Raheel Nawaz, and Sophia Ananiadou. 2011. Detecting experimental techniques and selecting relevant documents for protein-protein interactions from biomedical literature. *BMC Bioinf.* 12, 8 (2011), S11.

[48] Sze-Meng Jojo Wong and Mark Dras. 2009. Contrastive analysis and native language identification. In *Proceedings of the Australasian Language Technology Association Workshop*. 53–61.

[49] Sze-Meng Jojo Wong and Mar Dras. 2011. Exploiting parse structures for native language identification. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*. 1600–1610.

Received October 2019; revised January 2020; accepted February 2020