UrduAI: Writeprints for Urdu Authorship Identification

RAHEEM SARWAR*, Research Group in Computational Linguistics, Research Institute of Information and Language Processing, University of Wolverhampton, United Kingdom
SAEED-UL HASSAN, Department of Computer Science, Information Technology University, Pakistan

The authorship identification task aims at identifying the original author of an anonymous text sample from a set of candidate authors. It has several application domains such as digital text forensics and information retrieval. These application domains are not limited to a specific language. However, most of the authorship identification studies are focused on English and limited attention has been paid to Urdu. On the other hand, existing Urdu authorship identification solutions drop accuracy as the number of training samples per candidate author reduces, and when the number of candidate authors increases. Consequently, these solutions are inapplicable to real-world cases. Moreover, due to the unavailability of reliable POS taggers or sentence segmenters, all existing authorship identification studies on Urdu text are limited to the word n-grams features only. To overcome these limitations, we formulate a stylometric feature space, which is not limited to the word n-grams feature only. Based on this feature space we use an authorship identification solution that transforms each text sample into a point set, retrieves candidate text samples, and relies on the nearest neighbors classifier to predict the original author of the anonymous text sample. To evaluate our solution, we create a significantly larger corpus than existing studies and conduct several experimental studies which show that our solution can overcome the limitations of existing studies and report an accuracy level of 94.03%, which is higher than all previous authorship identification works.

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: Authorship identification, stylometry, text classification, forensic investigation

1 Introduction

The authorship identification task aims at identifying the *original* author of an anonymous text from a set of candidate authors. It has applications in several domains such as *criminal law*, identifying the original author of harassing letters or ransom notes [36]; *intelligence agencies work*, linking intercepted messages to know enemies [1, 26, 36]; and *plagiarism detection*, identifying the original authors of student submissions [27]. Nowadays, large text repositories have become available on the internet, while managing them has become a major challenge and an increasingly important task. Consequently, authorship identification has received significant attention from researchers in the areas of information retrieval [44], web information management [30], and *natural language processing* (NLP) [38].

Authorship identification solutions consist of two steps: (i) the stylometric features (also known as writing style markers) are computed from the text samples of the candidate authors, and (ii) a classification model is learned on features to predict the original author of the disputed text (see Section 2 for more details). Specifically, stylometric features can be organized into lexical, structural and syntactic features. The *lexical features* are the statistical measures of the character or word-based variations in a text samples such as word-length distributions [27], and vocabulary

*Corresponding Author

Authors' addresses: Raheem Sarwar, Research Group in Computational Linguistics, Research Institute of Information and Language Processing, University of Wolverhampton, Wulfruna St, Wolverhampton, Midlands, United Kingdom, WV1 1LY, R.Sarwar4@wlv.ac.uk; Saeed-Ul Hassan, Department of Computer Science, Information Technology University, 346-B, Ferozepur Road, Lahore, Punjab, Pakistan, 21210, saeed-ul-hassan@itu.edu.pk.

richness [36]. The *structural features* are related to the organization of the text such as the average sentence length [27]. Finally, the *function words*, and the *part-of-speech (POS)* tags-based features are examples of the *syntactic features* [25].

Most existing authorship identification studies are focused on English. However, the authorship identification applications are not limited to a specific language, community, ideology, culture or ethnicity. Consequently, this paper investigates the authorship identification problem of Urdu texts, which is the eight most commonly spoken language in the world with more than 163 million speakers[1]. Unfortunately, Urdu authorship identification has received little research attention. To the best of our knowledge, there are two notable publications that investigate the authorship identification of Urdu texts [3, 33]. The limitations of these studies are discussed later in this section.

Urdu is a member of the Indo-Aryan language family. Urdu is the national language of Pakistan and is widely spoken in the Indian subcontinent [2]. It uses Arabic script in cursive format (Nastaliq style) with the segmental writing system (see Fig. 1). Specifically, the Urdu language is based on an "*abjad*" system where the long vowels and consonants are necessarily written while the short vowels (diacritics) are optional. It is a bidirectional language where the numerals are written from left-to-right, while the characters are written from right-to-left (see Fig. 1). When characters are joined to make the words, they develop different shapes based on the context. Specifically, a character can have a maximum four shape variants known as initial, medial, final and isolated. The characters that can develop all four shapes are known as joiners, while the characters that can only have two shapes (final and isolated) are known as non-joiners (see Table 1 for joiners and non-joiner characters) [4, 6, 8, 17, 23, 24, 35].

Table 1. A Urdu sentence sample, joiner and non-joiner characters.

Sentence Sample	اردو کو ۱۹۵۴ میں پاکستان کی قومی زبان قرار دیا گیا تھا۔ Urdu was declared the national language of Pakistan in 1954.
Joiners	ی ہ ن م ل گ ک ق ف غ ع ظ ط ض ص ش س خ ح چ ج ث ٹ ت پ ب
Non-Joiners	ے ء و ژ ز ڑ ر ذ ڈ ا

Unlike English, a white space character is not considered as a reliable word boundary indicator in several languages [21]. Similarly, Urdu does not have consistent word boundary markings. For example, a writer may insert a space within a word قابل.احترام (respectable) in order to make it visually correct, where the character . represents the ASCII space character. If the writer omits the space it may lead to an incorrect visual form قابلاحترام of the same word. Contrarily, the writer may omit space between two words اردوزبان (Urdu language) because the shape of characters with or without space remains the same. That is, the Urdu words ending with non-joiner characters exhibit correct shape even without space. Consequently, a writer may omit space between words ending with non-joiner characters. These characteristics of Urdu makes the lexical stylometric features extraction process noisier in comparison to English and requires a solution associated with outlier handling techniques (see Section 3 for more details).

In addition to this, it is challenging to extract the syntactic or structural features from the Urdu text samples. For example, in English, if a period is followed by space and a word starting with the capital letter, it has the maximum probability to become a sentence marker. Unlike English, there is no capitalization in Urdu and the punctuations like '-', '!', '.' can be used as a sentence terminators and can also be used inside the sentence. For example, '-' is used in dates, part of the abbreviation, to describe range between two values and as a line breaker as well (see Table. 2). The unavailability of reliable NLP tools for Urdu, such as POS tagger or sentence segmenter, make us unable to extract

[1] https://www.ethnologue.com/language/urd

important syntactic and structural stylometric features from text samples. Consequently, there is a need to formulate new features that can capture the stylistic information about authors on structural, syntactic and lexical levels, without relying on POS taggers or sentence segmenters.

Table 2. The overview of the Urdu System.

احمد سات ۔آٹھ سال سے یو۔ ایس۔ اے۔ میں مقیم ھے اور ۳۱۔۱۲۔۲۰۲۰ کو واپس لوٹے گا۔
Ahmed has been living in U.S.A for the last six to seven years and will return on 31-12-2020.

As mentioned earlier that there are only 2 studies which investigate the authorship identification problem on Urdu text [3, 33] (see Section 2 for more details). In this investigation, we aim at addressing the limitations of these studies, and other challenges related to feature extraction from Urdu texts which can be illustrated as follows.

The Limitations of Existing Studies and Our Solution.

(1) Due to the unavailability of the reliable POS taggers or sentence segmenters, all existing authorship identification studies on Urdu text are limited to the word n-grams features only [3, 33]. In this investigation, we identify a *stylometric feature space* (WC) to perform authorship identification on Urdu texts. Unlike existing works which rely on word n-grams features only, our feature space is based on vocabulary richness, word n-grams, character richness, and character n-grams features. The extraction of these features do not require POS tagger, or sentence segmenter, and the performance of these features have never been evaluated on Urdu texts. There are four main advantages of including the character *n*-grams features in our feature space [18, 42]: (i) they can capture *complicated stylistic* information about authors on the structural, syntactic and lexical levels [18]; (ii) they are capable of tolerating *noise* in text samples (i.e., پاکستان and پاکسٹان have many common character 2-grams) [42]; (iii) they require high-dimensional representation, which is difficult for humans to understand and thus the deception attempts are likely to fail [42]; and (iv) they do not require tokenizers, parsers, taggers, or any non-trivial NLP tools, which makes them feasible for authorship attribution for low-resource languages.

(2) Existing solutions are unable to achieve good performance when the average number of text samples per candidate author is lower than 15. For example, the best existing authorship identification study on Urdu texts [3] uses 400 text samples for each candidate author. However, in real-world cases, such a huge amount of samples per candidate authors might not be available. For example, in the online review domain, most reviewers (authors) only write a few reviews (text samples). Specifically, it has been reported that on average each reviewer (author) only wrote 2.72 English reviews (text samples) in amazon.com, and only 8% of the authors wrote at least 5 text samples [16] . In addition to this, existing solutions are limited to 15 candidate authors only and increasing the number of candidate authors drastically drops the accuracy. However, a real-world scenario such as the number of reviewers on amazon.com or plagiarism detection may involve hundreds of candidate authors.
To handle large number of authors where each author has limited number of text samples, we adopt the *probabilistic k nearest neighbors* classifier (P*k*NN) [14]. This is because P*k*NN can learn from the limited set of training text 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].

(3) Authorship identifications process requires to capture the writing style variations within a text sample and across the text samples. However, existing Urdu authorship identification solutions are unable to capture the

writing style variations within a text sample. This is because they represent each text sample as one single vector in multidimensional space. Instead of representing each text sample as one single vector, we represent it as a set of vectors to capture the writing style variations within a text sample (see Section 3 for more details). Consequently, each authorship prediction relies on multiple vectors instead of one single vector. We note that representing each text sample as a set of vectors (data points) requires a set distance measure to compute the proximity between two point sets, such as *standard Hausdorff distance* (SHD) [27, 37].
Since Urdu does not have consistent word boundary markings, the lexical stylometric feature extraction process is noisier than English. Unfortunately, existing Urdu authorship identifications solutions are not associated with outlier handling mechanism [3, 33]. As mentioned earlier we adopt P*k*NN classifier to predict the true author of the anonymous text. We note that the P*k*NN classifier is also sensitive to outliers in the data. Instead of using SHD as proximity measure between to point sets, this issue can be addressed by using *partial Hausdorff distance* (PHD) [15] as a set similarity measure between two text samples. This is because, PHD is associated with outlier handling mechanism [15, 22, 27] (see Section 3 for details). One of the main motivations behind adopting the P*k*NN classifier is that it 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 authorship identification process[27, 36, 37].

(4) Existing authorship identification studies on Urdu texts are unable to handle a more realistic variation of the authorship identification problem known as open-set authorship identification. Unlike a standard authorship identification problem, open-set authorship identification considers that the actual author of the anonymous text sample might not be in the candidate author set. In such a case, when the actual author of the anonymous text is not in the candidate author set, an accurate solution should not attribute the anonymous text sample to any candidate author. We also show that our solution is capable of handling open-set authorship identification as well.

To evaluate our solution (UrduAI-WC)[2], we generate a new corpus of 985 Urdu text samples from 90 authors, which is larger than existing Urdu authorship identification studies (i.e., an increase of 6 folds in terms of the number of candidate authors). By using our corpus, we perform experimental studies to show that our solution can (i) mitigate the effect of outliers in the dataset; (ii) capture the writing style variations within a text sample; (iii) handle large candidate author set; (iv) perform well with limited number of training samples per class (author); and (v) achieve the accuracy level of 94.03% which is higher than all previous works.

Research Questions. Based on the aforementioned discussion, we answer the following research questions in this paper.

- ***Q # 1.*** What is the importance of vocabulary richness, character richness, and character n-grams features in Urdu authorship identification process?
- ***Q # 2.*** How important it is to use all four categories of the stylometric features in the authorship identification process?
- ***Q # 3.*** How much accuracy improvement can be obtained by using set similarity measures associated with outlier handling mechanisms in comparison to the standard set similarity measure (i.e., without outlier handling mechanism), in Urdu authorship identification process? The set similarity measures are discussed in Section 3.

[2]We call our solution UrduAI-WC for short where UrduAI refers to Urdu authorship identification and WC refers to our features space which consists of words and characters based information

Summary of Our Contributions. The contribution of this paper can be summarized as follows.

(i) We identify an effective stylometric features space (WC) for Urdu authorship identification task. Based on WC, we use an authorship identification solution (UrduAI) for Urdu that can overcome the limitations of existing studies and achieve the accuracy level of 94.03%, which is higher than all previous works on Urdu authorship identification.
(ii) Our solution can handle both the open-set and closed-set authorship identification problems.
(iii) We create a new significantly larger Urdu authorship identification corpus than existing studies (i.e., an increase of 6 folds in terms of the number of candidate authors).
(iv) We summarize the findings of our studies obtained by comparing the accuracy of our solution (UrduAI-WC) against existing authorship identification solution for Urdu called FAUT-W15, and LIP-W12; overall state-of-the-art authorship attribution methods, and four extensively used classifiers in authorship identification studies in different settings.

The rest of the paper is organized as follows. Section 2 reviews existing studies on authorship identification of Urdu texts. Section 3 describes our solution. Section 4 presents the experimental results. Section 5 contains the concluding remarks.

2 Literature Review

This section reviews the existing studies on Urdu authorship identification, the classical machine learning and deep learning-based methods.

2.1 Stylometric Features

Stylometric features are writing style markers that can be used to differentiate between documents written by different authors. Studies have proposed various stylometric features, including lexical, syntactic, structural and idiosyncratic features [9, 13, 20, 25, 28].

- Lexical features can be defined as statistical measures of word-based and character-based lexical variations in the text, such as vocabulary richness [13], word length distributions and character *n*-gram-based features [20]. Character *n*-grams are a contiguous sequence of *n* characters from a text sample. For example, the character 3-grams of the beginning of this sentence would be "For," "or␣," "r␣e," "␣ex," etc. The motivations for incorporating character *n*-gram-based features into our existing feature space are five-fold [18, 29–31, 41–43]: (i) Character *n*-gram-based features have been proven to perform well in solving authorship identification problems regardless of the length of text samples. Specifically, the character *n*-gram features can effectively capture the *stylistic information* of the authors from smaller text samples (i.e., 500 tokens) compared to vocabulary based features. The character *n*-gram features provide the best results when the value of *n* is 5, 4 or 3 [18, 29–31, 41, 43]. (ii) Character *n*-gram features can capture *complicated stylistic* information about authors on the syntactic, structural and lexical levels [18]. (iii) Character *n*-grams can tolerate *noise* in text samples (i.e., "stilometric" and "stylometric" have many common character 3-grams) [42]. (iv) Character *n*-grams require high-dimensional representation, which is not easy for humans to understand. Thus, attempts at deception are likely to fail [42]. (v) Extracting character *n*-grams does not require tokenizers, taggers, parsers or any language-dependent and non-trivial NLP tools, which makes them feasible for performing authorship attribution tasks.
- Examples of syntactic features include *function words* and *part-of-speech* tags [25].

- The structural features are writing style markers based on the presentation of the text, such as the average number of words in a paragraph or in a sentence [20].
- The idiosyncratic features are associated with the errors in the text samples of an author, such as grammatical mistakes and misspellings [7].

2.2 Existing Urdu Authorship Identification Solutions.

To the best of our knowledge, there are two notable studies on authorship identification of Urdu texts, which can be summarized as follows.

- **FAUT-W15.** The first study [3] rely on word n-grams features where the value of n varies from 1 to 5. As for the classification model, it makes use of improved sqrt-cosine similarity measure with *latent Dirichlet allocation* (LDA) to predict the true author of the anonymous text sample. This investigation was performed on a corpus of 15 authors where each author has 400 text samples. We call this method FAUT-W15 for short, where FAUT refers to the classification model and W15 refers to the feature space. This method is implemented using Gensim library [34].
- **LIP-W12.** The second study [33] also relies on word unigrams and bigrams. As for the classification model, it is based on linear interpolation of word n-grams probabilities. This investigation was performed on a corpus of 3 authors. We call this method LIP-W12 for short where LIP refers to the classification method and W12 refers to the feature space. The parameter settings of each existing Urdu authorship identification method are as suggested in the corresponding literatures.

Comparison With Our Method. Note that unlike our solution that represents each text sample as a point set, the existing methods represent each text sample as *one single data point* in a multidimensional space (see Section 3 for more details). As a result, the existing methods are *unable to* (i) capture the writing style variations within the same text sample; and (ii) apply set distance measures to handle outliers in the dataset. Moreover, unlike the feature space used by existing methods, our feature space contains word richness features, character richness features and character n-grams features. In this investigation, we compare the accuracy of our solution against these existing Urdu authorship identification methods.

Table 3. Comparison of related authorship identification methods

Method	#Authors	# Text samples	Avg. # Samples per Author	Avg. # tokens	Type of samples
FAUT-W15 [3]	15	6000	400	1183.2	News Articles
LIP-W12 [33]	3	NA	NA	NA	Poems
LMSAA [11]	62	62,000	1000	306.9	Reviews
StyloMatrix [10]	62	62,000	1000	306.9	Reviews
Our Method	90	985	10.95	1466	News Articles

2.3 Authorship Identification Using Classical Machine Learning Methods

We also compare the accuracy of our solution against well-known extensively used classification methods for authorship identification which includes *decision trees* (DT) (*.trees.J48) *naïve bayes* (NB) (*.bayes.NaïveBayes), support

vector machines (SVM) (*.functions.LibSVM), and *random forests* (RF) (*.trees.RandomForests) [1, 36, 37]. We used Weka.Classifiers implementation of these methods with default parameter settings.

We also compared the accuracy of the our solution against LMSAA [C/LM + W/LM + POS/LM] [11] which is based on the similarity-based paradigm. Specifically this method includes the concatenation of all documents written by a certain author in a single profile, which is used for the extraction of the style-markers. For the evaluation process, an attribution model is implemented to estimate the differences between every profile and an unseen text and the most likely author is chosen. The parameter setting is same as suggested in the corresponding literature.

2.4 Authorship Identification Using Deep Learning Based Methods.

Several existing studies focused on English used deep learning to perform the authorship identification task. For example, Solorio [39] performed authorship identification using a three-layer CNN model based on character bi-grams. They reported an accuracy level of 76.1% using a corpus written by 50 authors where each author has 1,000 samples. Moreover, Ge [12] reported 95% accuracy for authorship identification using *feedforward neural networks*. Posada et al., [32] propose to use the distributed representation at the document level to perform authorship identification. The proposed method learns distributed vector representations at the document level and then uses the SVM classifier to perform the automatic authorship attribution. Also, the propose method uses the word n-grams (n:2, 3, 4, 5) as the input data type for learning the distributed representation model. We call this method SVM-DDR-W25 for short and compare its performance against our solution (see Table 10).

Posada et al., [32] propose to use the distributed representation at the document level to perform authorship identification. The proposed method learns distributed vector representations at the document level and then uses the SVM classifier to perform the automatic authorship attribution. Also, the proposed method uses the word n-grams (n:2,3,4,5) as the input data type for learning the distributed representation model. We call this method SVM-DDR-W25 for short and compare its performance against our solution (see Table 10). Specifically, we used the Doc2vec [19] method available in the freely downloadable GENSIM module in order to implement this model. The implementation of the Doc2vec method requires the following three parameters: (i) length of the vector, (ii) the size of the window that captures the neighborhood, and (iii) the minimum frequency of words to be considered into the model. We have tried different values of these parameters and the following values resulted in best accuracy. We set the length of the vector to 300 features, a window size to 10 and the minimum frequency to 4.

The Doc2vec module uses stochastic gradient descend and back-propagation algorithms for generating the model. We note that, these algorithms use random values that do not guarantee their reproducibility. In order to ensure the reproducibility of our experiments, the values of the following parameters are fixed as recommended in the user manual: the value of threshold for configuring which higher-frequency words are randomly down-sampled is set to 1e−3, negative sampling is set to 5, the seed of the random number generator is set to 1, and the number of threads is set to 1. The rest of parameters are set with default values. We trained the model several times over the unlabeled corpus but exchanging the order of entry of the documents as described in the original study [32].

We have also compared the performance of our solution against StyloMatrix [10] that incorporates different categories of linguistic features into distributed representation of words to learn simultaneously the writing style representations based on unlabeled texts for the authorship identification task. In particular, these models allow topical, lexical, syntactical, and character-level feature vectors of each document to be extracted as stylometrics and train a simple logistic regression model on the representations of a chosen modality and use it to classify the unknown document. There are four hyper-parameters for the aforementioned models. The d1, d2, and d3, respectively, denote the vector

size of the topical-lexical model, the character model, and the syntactic model. $\mathcal{W}$(tp) is only for the topical-lexical model, which denotes the size of the sliding window for context. We tried different parameters and then we picked d1 = 700, $\mathcal{W}$(tp) = 8, d2 = 600, and d3 = 600 as our final hyper-parameter values.

Comparison with Our Work. The deep learning methods may achieve high accuracy for authorship identification task only when a large amount of training data is available. However, in this investigation, we aim at designing an authorship identification solution for Urdu which can perform well in real-world cases where a huge amount of training data is not available and where the average number of writing samples for each candidate author is around 10.

3 Methodology

Our solution consists of four parts including (i) data collection, (ii) preprocessing, (iii) candidate retrieval, and (iv) probabilistic k nearest neighbour (PkNN) classification as shown in Figure 1. The data collection part of our solution extracts the text samples of the authors from the website and send it to the preprocessing part of our solution. The preprocessing part of our solution performs two processes: (i) text sample partitioning, and (ii) feature extraction. Once the feature extraction process is completed, we then store the feature values into the stylometry database. Given a query text sample, the candidate retrieval part of our solution executes the set similarity query to retrieve a set of top-k *stylistically similar text samples* (SSTs) from the stylometry database and send it to the PkNN classifier. The PkNN classifier identifies the original author of the query document from the set of top-k SSTs. Each part is discussed in the following subsections.

3.1 Data Collection.

We note that the number of publicly available corpora for Urdu authorship identification is limited; and the size of the publicly available corpora is small in terms of the number of candidate authors. To perform experiments, we created a new Urdu authorship identification corpus extracted from an online news website[3]. We wrote a data scrapper in Python which extracts the data by retrieving all the URLs of each author and based on the retrieved URLs, extracts the text of each author from the website. Our corpus contains 985 text samples from 90 authors where the average length of texts is 1466 tokens. Moreover, our corpus is significantly larger than existing studies (i.e., an increase of 6 folds in terms of the number of candidate authors). Furthermore, the average number of text sample for each class (author) is between 10 to 11, which is a more realistic scenario where a large number of writing samples per author may not be available.

3.2 Preprocessing.

The preprocessing part of our solution performs two processes: (i) text sample partitioning, and (ii) feature extraction. In the first process we partition each text sample into fixed-size chunks[4], where the size of each chunk is 500 tokens. Consequently, a 1000-token text sample results in 2 chunks. However, 2 chunks per text sample might not be enough to obtain reliable stylometric information for each author. To obtain more chunks from each text samples, we use the concept of the sliding window to partition the text samples into chunks, where the size of the sliding window is fixed to 100 tokens. Consequently, each 1000-tokens text sample results in 5 chunks. We tried different values of chunk size and sliding window size. We found that chunk size of 500-tokens and sliding window of size 100-tokens provide the best accuracy. There are two advantages of partitioning each text sample into a *set of chunks*. (i) We can compute the

[3] http://dunya.com.pk/
[4] A chunk is a collection of tokens

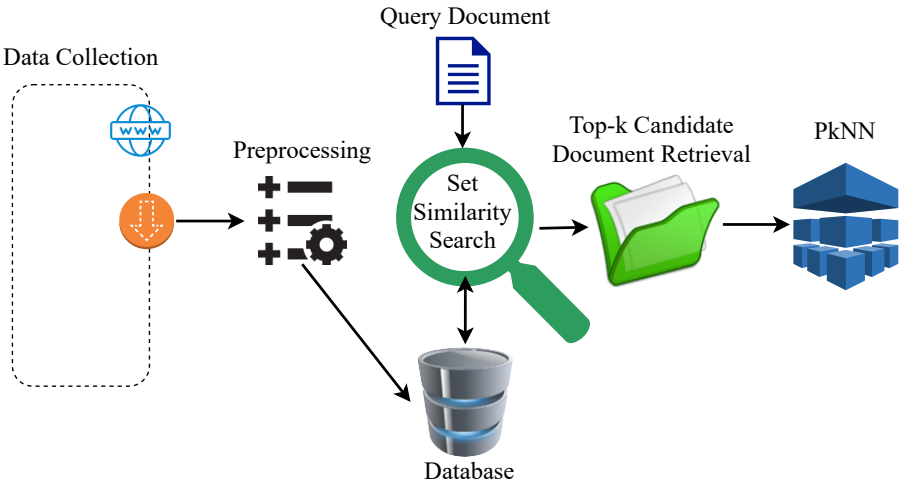

Fig. 1. System Overview.

stylistic variations between text samples as a *set distance*. Specifically, we can use those set distance measures which are capable of mitigating the effect of outliers in the data such as PHD [15]. (ii) We can capture the *stylistic variation* of an author within a text sample. This is because each authorship identification prediction is produced by multiple points rather than one single point.

In the second process, we extract the features from each chunk using_THE_NLTK toolkit. As a result, each text sample is represented as a point set in a multidimensional space. Our feature space is given in Table 4. As shown in Table 4 our feature space contains different features based on vocabulary richness (from 1 to 13), word n-grams (from 14 to 513), character richness (from 514 to 520) and character n-grams (from 521 to 1520).

In Table 4 N represents the count of words, V represents the count of distinct words, V_i represents the count of words that occur i times, and C represents the total number of characters. The word and character n-grams features are based on the most frequent n-grams with respect to the TF-IDF scores where the value of n varies from 1 to 5. Once we complete the feature extraction process, we store the feature values into the stylometry database.

3.3 Candidate Retrieval.

Given query text sample (Q), we first apply the preprocessing step of our solution on Q which transforms it into a *point set*. We then execute set similarity query to retrieve top-k *stylistically similar text samples (SSTs)* from the corpus. While retrieving the top-k SSTs, we tried two set similarity measures including SHD and PHD [15] as a proximity measure between two point sets. The SHD between two points sets Q and F can be calculated as:

$$h(Q, F) = \max_{q_i \in Q} \min_{f_j \in F} d(q_i - f_j).$$

That is, SHD can be calculated by: (i) ranking all data points in a query text Q in accordance with the *minimum distance* to the text F in the corpus; and (ii) *selecting the maximum of the minimum distances*. It has been reported that SHD is sensitive to the noise in the data [15, 22]. To mitigate the noise (outlier) sensitivity issue associated with SHD, researchers formulated another variant of SHD known as *partial Hausdorff distance (PHD)* [15]. Specifically, the PHD

measures average out the effect of the outlier over the minimum distances falling into a specified range i.e., $(50\%, 75\%]$. We have tried different values of k and PHD range. We found that the k value of 10 and the PHD range $(50\%, 75\%]$ resulted in the best accuracy. The experimental results regarding set distance measures are reported in section 4.1.

3.4 Probabilistic k Nearest Neighbors Classification (PkNN)

We apply PkNN [14] to the retrieved top-k SSTs to make a probabilistic prediction of a query sample. Unlike simple kNN classifier where the output is one single class (author), the PkNN classifier produces a probability mass function (PMF) over all classes (candidate authors) associated to the retrieved SSTs. We apply the PkNN [14] that utilizes the distance values of the k nearest neighbors (SSTs in this case) to weight the distribution of the probability. An exponential function is used to smooth the distance-probability mapping [36].

Table 4. List of Stylometric Features (N represents the count of words, V represents the count of distinct words, V_i represents the count of words that occur i times, S.D represents the standard deviation and C represents the total number of characters.)

Vocabulary Richness Features		
1. N: Total #words	2. V: Total #distinct words	3. Average word length
4. S.D. of word lengths	5. $\frac{V}{N}$	6. $\frac{10^4(\sum i^2 V_i - N)}{N^2}$
7. $\frac{V}{\sqrt{N}}$	8. $\frac{\log V}{\log N}$	9. $\frac{(100 \log N)}{(1-V_1)/V}$
10. $\frac{V_2}{V}$	11. $\frac{logV}{\log(\log N)}$	12. $\frac{(1-V^2)}{V^2(\log N)}$
13. Entropy of word freq. ditri.	14-513. Word n-grams (n:1-5)	
Character Richness Features		
514. **C:** Total # chars	515. Freq. of Urdu chars	516. Freq. of wowel and tone marks
517. $\frac{Freq.\ of\ Urdu\ chars}{C}$	518. $\frac{Freq.\ of\ White\ spaces}{C}$	519. $\frac{Freq.\ of\ Arabic\ numeric\ chars}{C}$
520. $\frac{Freq.\ of\ Arabic\ numeric\ chars}{C}$	521-1520.Char. n-grams (n:1-5)	

4 Experimental Results

This section is divided into two subsections. The first subsection reports the findings associated with our solution only. The second subsection provides a performance comparison between our solution and previous authorship identification works.

Evaluation Measure and Strategy. We used accuracy as an evaluation measure which is defined as follows: An authorship prediction is considered correct if the true author of the query text is identified as the most likely author. To evaluate the accuracy of all methods in this investigation, we use 5-fold cross-validation unless stated otherwise. Recall that, as for our method, each text is represented as a point set. To avoid test-train set contamination in the evaluation process of our method we ensure that when a text sample is used for testing it is purely used for testing.

4.1 Experimental Results of Our Solution Only.

Effect of Feature Types. This study provides answers to the first two questions mentioned in the Introduction section. As can be seen from Table 5 that the most important feature type is character n-grams, followed by word n-grams, character richness and vocabulary richness respectively. Moreover, the stylometric information captured by different features categories is complementary and orthogonal. Consequently, combining all feature categories improves the

performance of authorship identification process. Hence, we confine the rest of the experimental studies to combined categories of the stylometric features only.

Table 5. Our solution: Ablation study to show the effect of feature types by leave-one-out evaluations

Vocabulary Richness	Character Richness	Word n-grams	Character n-grams	Accuracy
-	✓	✓	✓	90.13%
✓	-	✓	✓	89.67%
✓	✓	-	✓	87.69%
✓	✓	✓	-	85.13%
✓	✓	✓	✓	**94.03%**

Writing Style Variations and Outlier Handling Mechanism. There are two objectives of this experimental study. The first objective is to show the effect of capturing the writing style variations within a text sample. Recall that, in order to capture the variations within a text sample, instead of representing each text sample as a point (feature vector), we represent it as a point set (set of feature vectors). As a result, each authorship identification 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 writing style variations within a text sample, we formulate a baseline method and compare its performance against our solution (UrduAI-WC). The difference between the baseline method and our method (UrduAI-WC) is that, the former represents each text sample as a one single data point and the latter represents each text sample as a point set. The experimental results given in Table 6 show that our method (UrduAI-WC) is capable of capturing the writing style variations within a text sample and thus outperforms the baseline method.

The second objective of this study is to show the effect of outliers in the data on the accuracy of the authorship identification process. Specifically, we provide the accuracy comparison between two set distance measures: (i) *standard Hausdorff distance (SHD*), and (ii) *partial Hausdorff distance (PHD)*, 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 the noise to be handled. and using PHD, which is associated with outlier handling mechanism, improves the accuracy of the authorship identification process. Since the PHD measure provided better performance in comparison to the SHD measure, all other results are reported based on PHD only.

Open-set Authorship Identification. To perform an open-set authorship identification study, we generate a new corpus of 500 text samples[5] from 90 authors. There are 90 test samples[6]. Out of 90 test samples, 45 of them are from authors in the candidate author set and rest of the test samples are from the non-candidate authors. All test samples are from different authors. The test sample is considered to be written by the most likely author *if and only if* the author has a probability larger than a predefined threshold value. Otherwise, the prediction is considered *uncertain* and the query sample was classified as written by a non-candidate author. According to this definition, consider the following

[5]text samples refer to the training data
[6]test sample refers to the anonymous text

Table 6. Our Solution: The effect of capturing writing style variations and outlier handling mechanism on the accuracy of authorship identification task

Set Distance Measure	Accuracy
Baseline	77.38%
UrduAI-WC (SHD)	87.96%
UrduAI-WC (PHD)	**94.03%**

two cases. (i) *Outside Author:* when the test sample is written by a non-candidate, a prediction is considered correct *if and only if* no author in the prediction has a probability greater than the probability threshold value. (ii) *Inside Author:* when the test sample is written by an author in the candidate set, a prediction is considered correct *if and only if* the most likely author is the correct author and its probability value is greater than the predefined threshold.

The experimental results given in Table 7 are two-fold. (i) Table 7 shows that as the threshold value increases from 30% to 70% the accuracy of identifying that the test sample is written by a non-candidate author increases. This is because, as the threshold value increases, it becomes more difficult for any author to clear the threshold in each prediction making it easier for each prediction to be considered uncertain. (ii) Table 7 shows that as the threshold value increases the accuracy of identifying the author inside the candidate set decreases. This is because, since it is becoming more difficult for the most likely author to clear the probability threshold, predictions with the correct author are more likely to be mistakenly treated as uncertain. We can also see that the probability threshold value of 40 provides a good trade-off between the *inside author* accuracy and *outside inside* accuracy.

Table 7. Our solution: The effect of varying the threshold from 30% to 70% on the accuracy of the open-set and closed-set authorship identification

Outside Candidate (Open-set)					Inside Candidate (Closed-set)				
30	**40**	**50**	**60**	**70**	**30**	**40**	**50**	**60**	**70**
92.13%	94.42%	95.14%	95.53%	95.78%	96.12%	94.23%	93.52%	91.61%	91.17%

The effect of Chunk Size. In this study, we vary the chunk size as 300, 400, 500, and 600 tokens. As shown in Table 8 increasing chunk size positively affects the accuracy. However, chunk size of 600 tokens shows only a marginal performance reduction over the chunk size of 500 tokens. This might be due to the fact that increasing chunk size reduces the number of result chunks from a text sample.

Table 8. Accuracy: Effect of chunk size on the accuracy of authorship identification

The effect of Chunk Size			
400	500	600	700
93.61%	93.96%	**94.03%**	94.01%

The effect of Sliding Window Size. In this study, we vary the sliding window size between 50 to 200 tokens. As shown in Table 9 increasing sliding window size negatively affects the accuracy. This might be due to the fact that increasing the sliding window size reduces the number of resulting chunks from a text sample.

Table 9. Accuracy: Effect of sliding window size on the accuracy of authorship identification

The effect of Sliding Window Size			
50	100	150	200
94.02%	**94.03%**	93.98%	93.89%

4.2 Detailed Comparison between our method and previous state-of-the-art authorship identification methods.

In this subsection, we study the effect of increasing the number of candidate author. Moreover, we cross-compare the feature extraction and the classification parts of all previous state-of-the-art authorship identification solutions against the proposed solution.

The Effect of the Candidate Author Set Size. In this study, we provide the performance comparison between our solution and the existing authorship identification studies by varying the size of the candidate author set from 30 to 90. To vary the size of the candidate author set we randomly choose the authors from our dataset. The experimental results given in Table 10 show that (i) our solution (UrduAI) can scale as the number of candidate authors increases; and (ii) our method (UrduAI) outperforms all existing authorship identification methods.

Table 10. Accuracy: Effect of candidate author set size on the accuracy of authorship identification task

Method	The effect of Number of Candidate Authors		
	30	60	90
UrduAI-WC [Our Method]	**94.12%**	**94.10%**	**94.03%**
FAUT-W15	84.39%	80.16%	74.33%
LIP-W12	81.56%	81.22%	72.19%
LMSAA [C/LM + W/LM]	80.15%	82.25%	65.85%
StyloMatrix [Lexical+Topical]	75.65%	72.87%	69.93%
SVM-DDR-W25	74.82%	71.22%	69.42%

We also cross-compare the feature space and the method part of all solutions. At this stage, we also introduce another feature space called C15 which consists of character n-grams where the value of n is between 1 to 5. The experimental results are given in Table 11. Our findings of this study are twofold: (i) Our method (UrduAI) outperforms other methods. This is because, unlike other methods that represents each text sample as a point, our method UrduAI represents each text sample as a point set. As a result, our method is capable of capturing the writing variations within a text sample since each prediction is based on multiple data points, and mitigating the effect of outliers in the data with the help of set similarity measures associated with outlier handling mechanism. (ii) Our feature space (WC) reports higher accuracy than all other feature spaces. This is because, unlike W12 and W15 feature spaces, our feature space contains vocabulary richness features, character richness features and character n-grams features which play an important role in improving the authorship identification accuracy (see the experimental results given in Table 5).

We also compare the performance of all methods on different datasets. Recall that the FAUT-W15 dataset consists 6000 text samples from 15 authors. On the other hand, our dataset consists of 985 text samples from 90 authors. The experimental results are given in Table 12. The findings of this study show that our method can outperform all methods

Table 11. The performance comparison between our solution and existing methods

Feature Space	Method							
	UrduAI	FAUT	LIP	SVM	DT	NB	RF	LMSAA
WC	**94.03%**	**85.11%**	**84.91%**	**84.40%**	**82.66%**	**83.44%**	**81.63%**	**82.34%**
W15	**92.22%**	74.33%	73.91%	73.02%	71.29%	72.14%	70.41%	76.83%
W12	**91.59%**	73.21%	72.19%	71.69%	69.73%	70.11%	68.91%	74.08%
C15	**90.78%**	71.44%	72.13%	79.11%	63.81%	70.16%	69.85%	77.34%

on every dataset. That is, our method is capable of handling large number of candidate authors where each candidate author has a small number of writing samples.

Table 12. Performance comparison between our solution and existing methods on different datasets

Method	Accuracy	
	Data used in FAUT-W15 [3]	Our Dataset
UrduAI-WC [Our Method]	**94.22%**	**94.03%**
FAUT-W15	92.89%	74.33%
LIP-W12	87.24%	72.19%
LMSAA [C/LM + W/LM]	84.20%	65.86%
StyloMatrix [Lexical+Topical]	83.65%	69.93%
SVM-DDR-W25	76.12%	64.42%

Performance of Our Solution on Urdu and English Corpora. In this study, we apply our solution on Urdu and English corpora. To perform a fair comparison, we create the subsets of CCAT50 [40] and our Urdu corpora where each subset corpus consists of 500 documents written by 50 authors. As can be seen from Table 13 that our solution outperforms the existing authorship attribution methods for both English and Urdu corpora.

Table 13. Performance comparison between our solution and existing methods on different languages

Method	Accuracy	
	CCAT50 [40] (English)	Our Dataset (Urdu)
UrduAI-WC [Our Method]	**92.97%**	**94.11%**
FAUT-W15	83.15%	82.30%
LIP-W12	82.11%	81.43%
LMSAA [C/LM + W/LM + POS/LM]	83.61%	80.12%
StyloMatrix [Lexical+Topical]	79.22%	74.29%
SV-DDR-W25	78.90%	73.04%

5 Conclusions

This paper presents an authorship identification solution for Urdu texts. By using a significantly larger corpus than existing studies, we perform extensive experimental studies to show that our solution can (i) mitigate the effect of outliers in the dataset; (ii) handle a large number of candidate authors in data-poor conditions; and (iii) achieve the accuracy level of 94.03% which is significantly higher than all previous authorship identification works. In addition to this, we have answered important research questions in this paper. This paper has laid the foundation for future work in Urdu authorship identification task and opened the door for future work on Urdu to keep up with the work in other languages.

References

[1] Malik H. Altakrori, Farkhund Iqbal, Benjamin C. M. Fung, Steven H. H. Ding, and Abdallah Tubaishat. 2019. Arabic Authorship Attribution: An Extensive Study on Twitter Posts. *ACM Trans. Asian & Low-Resource Lang. Inf. Process.* 18, 1 (2019), 5:1–5:51.

[2] Maaz Amjad, Grigori Sidorov, and Alisa Zhila. 2020. Data Augmentation using Machine Translation for Fake News Detection in the Urdu Language. In *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*. European Language Resources Association, 2537–2542.

[3] Waheed Anwar, Imran Sarwar Bajwa, M. Abbas Choudhary, and Shabana Ramzan. 2019. An Empirical Study on Forensic Analysis of Urdu Text Using LDA-Based Authorship Attribution. *IEEE Access* 7 (2019), 3224–3234.

[4] Muhammad Awais and Muhammad Shoaib. 2019. Role of Discourse Information in Urdu Sentiment Classification: A Rule-based Method and Machine-learning Technique. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 18, 4 (2019), 34:1–34:37.

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

[6] Riyaz Ahmad Bhat, Irshad Ahmad Bhat, and Dipti Misra Sharma. 2017. Improving Transition-Based Dependency Parsing of Hindi and Urdu by Modeling Syntactically Relevant Phenomena. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 16, 3 (2017), 17:1–17:35.

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

[8] Prakash Choudhary and Neeta Nain. 2016. A Four-Tier Annotated Urdu Handwritten Text Image Dataset for Multidisciplinary Research on Urdu Script. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 15, 4 (2016), 26:1–26:23.

[9] Edwin Dauber, Rebekah Overdorf, and Rachel Greenstadt. 2017. Stylometric Authorship Attribution of Collaborative Documents. In *First International Conference, CSCML 2017, Beer-Sheva, Israel, June 29-30, 2017, Proceedings*. 115–135.

[10] Steven HH Ding, Benjamin CM Fung, Farkhund Iqbal, and William K Cheung. 2017. Learning stylometric representations for authorship analysis. *IEEE transactions on cybernetics* 49, 1 (2017), 107–121.

[11] Olga Fourkioti, Symeon Symeonidis, and Avi Arampatzis. 2019. Language models and fusion for authorship attribution. *Information Processing & Management* 56, 6 (2019), 102061.

[12] Zhenhao Ge, Yufang Sun, and Mark J. T. Smith. 2016. Authorship Attribution Using a Neural Network Language Model. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA*. 4212–4213.

[13] Jack Grieve. 2007. Quantitative authorship attribution: An evaluation of techniques. *LLC* 22, 3 (2007), 251–270.

[14] C.C. Holmes and N.M. Adams. 2002. A probabilistic nearest neighbour method for statistical pattern recognition. *J R Stat Soc Series B Stat Methodol* 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] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the International Conference on Web Search and Web Data Mining, WSDM 2008, Palo Alto, California, USA, February 11-12, 2008*, Marc Najork, Andrei Z. Broder, and Soumen Chakrabarti (Eds.). ACM, 219–230.

[17] Safia Kanwal, Kamran Malik, Khurram Shahzad, Faisal Aslam, and Zubair Nawaz. 2020. Urdu Named Entity Recognition: Corpus Generation and Deep Learning Applications. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 19, 1 (2020), 8:1–8:13. https://doi.org/10.1145/3329710

[18] Vlado Kešelj, Fuchun Peng, Nick Cercone, and Calvin Thomas. 2003. N-gram-based author profiles for authorship attribution. In *Proceedings of the conference pacific association for computational linguistics, PACLING*, Vol. 3. Harifax Canada, 255–264.

[19] Quoc V. Le and Tomás Mikolov. 2014. Distributed Representations of Sentences and Documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014 (JMLR Workshop and Conference Proceedings, Vol. 32)*. JMLR.org, 1188–1196.

[20] Jiexun Li, Rong Zheng, and Hsinchun Chen. 2006. From fingerprint to writeprint. *Commun. ACM* 49, 4 (2006), 76–82.

[21] Peerat Limkonchotiwat, Wannaphong Phatthiyaphaibun, Raheem Sarwar, Ekapol Chuangsuwanich, and Sarana Nutanong. 2020. Domain adaptation of thai word segmentation models using stacked ensemble. Association for Computational Linguistics.

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

[23] Muhammad Kamran Malik. 2017. Urdu Named Entity Recognition and Classification System Using Artificial Neural Network. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 17, 1 (2017), 2:1–2:13.

[24] Khawar Mehmood, Daryl Essam, Kamran Shafi, and Muhammad Kamran Malik. 2020. Sentiment Analysis for a Resource Poor Language - Roman Urdu. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 19, 1 (2020), 10:1–10:15.

[25] Frederick Mosteller and David Wallace. 1964. Inference and disputed authorship: The Federalist. *Reading MA: Addison-Wesley* (1964).

[26] Arvind Narayanan, Hristo Paskov, Neil Zhenqiang Gong, John Bethencourt, Emil Stefanov, Eui Chul Richard Shin, and Dawn Song. 2012. On the feasibility of internet-scale author identification. In *Security and Privacy (SP), 2012 IEEE Symposium on*. IEEE, 300–314.

[27] Sarana Nutanong, Chenyun Yu, Raheem Sarwar, Peter Xu, and Dickson Chow. 2016. A Scalable Framework for Stylometric Analysis Query Processing. In *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*. 1125–1130.

[28] Mathias Payer, Ling Huang, Neil Zhenqiang Gong, Kevin Borgolte, and Mario Frank. 2015. What You Submit Is Who You Are: A Multimodal Approach for Deanonymizing Scientific Publications. *IEEE Trans. Information Forensics and Security* 10, 1 (2015), 200–212.

[29] Fuchun Peng, Dale Schuurmans, and Shaojun Wang. 2004. Augmenting naive bayes classifiers with statistical language models. *Information Retrieval* 7, 3-4 (2004), 317–345.

[30] Fuchun Peng, Dale Schuurmans, Shaojun Wang, and Vlado Keselj. 2003. Language independent authorship attribution using character level language models. In *Proceedings of the tenth conference on European chapter of the Association for Computational Linguistics-Volume 1*. Association for Computational Linguistics, 267–274.

[31] Jian Peng, Kim-Kwang Raymond Choo, and Helen Ashman. 2016. Astroturfing Detection in Social Media: Using Binary n-Gram Analysis for Authorship Attribution. In *2016 IEEE Trustcom/BigDataSE/ISPA, Tianjin, China, August 23-26, 2016*. 121–128.

[32] Juan-Pablo Posadas-Durán, Helena Gómez-Adorno, Grigori Sidorov, Ildar Batyrshin, David Pinto, and Liliana Chanona-Hernández. 2017. Application of the distributed document representation in the authorship attribution task for small corpora. *Soft Computing* 21, 3 (2017), 627–639.

[33] Agha Ali Raza, Awais Athar, and Sajid Nadeem. 2009. N-gram based authorship attribution in Urdu poetry. In *Proceedings of the Conference on Language & Technology*. 88–93.

[34] Radim Rehurek and Petr Sojka. 2010. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer.

[35] Ali Saeed, Rao Muhammad Adeel Nawab, Mark Stevenson, and Paul Rayson. 2019. A Sense Annotated Corpus for All-Words Urdu Word Sense Disambiguation. *ACM Trans. Asian Low Resour. Lang. Inf. Process.* 18, 4 (2019), 40:1–40:14.

[36] Raheem Sarwar, Qing Li, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. A Scalable Framework for Cross-lingual Authorship Identification. *Information Sciences* (2018).

[37] 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.

[38] Fabrizio Sebastiani. 2006. Classification of text, automatic. *The encyclopedia of language and linguistics* 14 (2006), 457–462.

[39] Thamar Solorio, Paolo Rosso, Manuel Montes-y-Gómez, Prasha Shrestha, Sebastián Sierra, and Fabio A. González. 2017. Convolutional Neural Networks for Authorship Attribution of Short Texts. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 2: Short Papers*. 669–674.

[40] Efstathios Stamatatos. 2008. Author identification: Using text sampling to handle the class imbalance problem. *Information Processing & Management* 44, 2 (2008), 790–799.

[41] Efstathios Stamatatos. 2009. A survey of modern authorship attribution methods. *JASIST* 60, 3 (2009), 538–556.

[42] Efstathios Stamatatos. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy* 21, 2 (2013), 421–439.

[43] Efstathios Stamatatos et al. 2006. Ensemble-based author identification using character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, Vol. 36. 41–46.

[44] Ying Zhao and Justin Zobel. 2007. Searching With Style: Authorship Attribution in Classic Literature. In *Computer Science 2007. Proceedings of the Thirtieth Australasian Computer Science Conference (ACSC2007). Ballarat, Victoria, Australia, January 30 - February 2, 2007. Proceedings*. 59–68.