






Please cite the Published Version

Sarwar, Raheem , Perera, Maneesha , Teh, Pin Shen , Nawaz, Raheel  and Hassan, Muhammad Umair  (2024) Crossing linguistic barriers: authorship attribution in Sinhala texts. ACM Transactions on Asian and Low-Resource Language Information Processing, 23 (5). pp. 1-14. ISSN 2375-4699

DOI: <https://doi.org/10.1145/3655620>

Publisher: Association for Computing Machinery (ACM)

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/634675/>

Usage rights:  In Copyright

Additional Information: © Authors 2024. This is the author's version of the work. It is posted here for your personal use. Not for redistribution. The definitive Version of Record was published in ACM Transactions on Asian and Low-Resource Language Information Processing, <http://dx.doi.org/10.1145/3655620>.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)

Crossing Linguistic Barriers: Authorship Attribution in Sinhala Texts

RAHEEM SARWAR, OTEHM, Manchester Metropolitan University, United Kingdom

MANEESHA PERERA, School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand

PIN SHEN TEH, OTEHM, Manchester Metropolitan University, United Kingdom

RAHEEL NAWAZ, Staffordshire University, United Kingdom

MUHAMMAD UMAIR HASSAN, Department of ICT and Natural Sciences, Norwegian University of Science and Technology (NTNU), Ålesund, Norway

Authorship attribution involves determining the original author of an anonymous text from a pool of potential authors. Author attribution task has applications in several domains, such as plagiarism detection, digital text forensics, and information retrieval. While these applications extend beyond any single language, existing research has predominantly centered on English, posing challenges for application in languages like Sinhala due to linguistic disparities and a lack of language processing tools. We present the first comprehensive study on cross-topic authorship attribution for Sinhala texts and propose a solution that can effectively perform the authorship attribution task even if the topics within the test and training samples differ. Our solution consists of three main parts: (i) extraction of topic-independent stylometric features, (ii) generation of the small candidate author set with the help of similarity search, and (iii) identification of the true author. Several experimental studies were carried out to demonstrate that the proposed solution can effectively handle real-world scenarios involving a large number of candidate authors and a limited number of text samples for each candidate author.

CCS Concepts: • **Computing methodologies** → **Classification and regression trees**.

Additional Key Words and Phrases: Authorship Attribution, Sinhala, Linguistic Barriers, Low-Resource

ACM Reference Format:

Raheem Sarwar, Maneesha Perera, Pin Shen Teh, Raheel Nawaz, and Muhammad Umair Hassan. 2024. Crossing Linguistic Barriers: Authorship Attribution in Sinhala Texts. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* 1, 1 (May 2024), 14 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 INTRODUCTION

The main objective of authorship attribution is to determine the genuine author of an anonymous text among a group of potential authors [39, 46, 48, 50–52, 54–56, 60, 60]. Authorship attribution has several impactful real-world applications across various domains: (1) *Plagiarism Detection*: Identifying the original authors of texts helps detect and prevent plagiarism in academic, journalistic, and online content. (2) *Digital Forensics*: In legal cases or criminal investigations, authorship

Authors' addresses: Raheem Sarwar, R.Sarwar@mmu.ac.uk, OTEHM, Manchester Metropolitan University, United Kingdom; Maneesha Perera, maneeshanick@gmail.com, School of Information, Computer and Communication Technology, Sirindhorn International Institute of Technology, Thammasat University, Thailand; Pin Shen Teh, OTEHM, Manchester Metropolitan University, United Kingdom, p.teh@mmu.ac.uk; Raheel Nawaz, Staffordshire University, United Kingdom, raheel.nawaz@staffs.ac.uk; Muhammad Umair Hassan, Department of ICT and Natural Sciences, Norwegian University of Science and Technology (NTNU), Ålesund, Norway, muhammad.u.hassan@ntnu.no.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2375-4699/2024/5-ART

<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

attribution can be crucial in determining the origin of anonymous or threatening messages, helping law enforcement or legal professionals. (3) *Security and Fraud Detection*: Authenticating the authorship of documents, emails, or online content can aid in security measures, such as verifying the legitimacy of communication to prevent fraud. (4) *Information Retrieval and Search Engines*: Enhancing search engines by attributing content to specific authors can improve content relevance and ranking, aiding users in finding more credible and relevant information. (5) *Literary Studies*: Studying and understanding authors' writing styles across different works can aid literary scholars in analyzing and contextualizing texts, contributing to literary criticism and analysis. (6) *Historical and Textual Analysis*: Authorship attribution assists in determining the authors of historical texts with uncertain origins, contributing to historical research and textual analysis. (7) *Content Curation and Recommendation Systems*: In content recommendation systems, identifying authors helps recommend similar content by the same authors or with similar writing styles. (8) *Social Media and Online Behavior Analysis*: Attribution of anonymous online posts or messages can help in understanding social media behaviour and identifying potential threats or abusive behaviour. These applications highlight the diverse and practical uses of authorship attribution across fields, showcasing its significance in various industries and disciplines [59]. Although these applications are not confined to a single language, English has been the focus of most of the existing studies of the authorship attribution task. To the best of our knowledge, this is the first comprehensive study that focuses on the cross-topic attribution of authorship of Sinhala texts.

Generally, this task is completed in two steps: In the first step, the writing style indicators (i.e., stylometric traits) are extracted as features from the text samples. In the second step, the extracted features are used to train classification models, which will identify the true author of the anonymous text sample [6, 38, 47, 49]. The authorship attribution task relies on the text's lexical, syntactic, and structural features. Some of these features can only be extracted from text samples written in resource-rich languages such as English due to the availability of advanced natural language processing (NLP) tools and technologies [42, 57]. However, such reliable NLP tools are limited or inaccurate for low-resource languages such as Sinhala.

Sinhala or Sinhalese is an Indo-Aryan language [15] spoken by about 17 million Sinhalese people in Sri Lanka as their official language. Due to linguistic variations between the two languages, existing authorship attribution procedures largely intended for the English language do not apply to Sinhala. For instance, in contrast to English: (i) Sinhala boasts 44 consonants; (ii) Sinhala encompasses 18 vowel symbols, which give rise to numerous compound vowels; (iii) there exist distinct characters known as "murdhaja," altering the sentence's meaning; (iv) each consonant inherently carries its vowel; (v) Sinhala exhibits few consonant clusters; (vi) Structurally, Sinhala follows the SOV (subject-object-verb) pattern, whereas English adheres to SVO (subject-verb-object); (vii) Pronouns in Sinhala, especially for second and third persons, are gender-specific and necessitate varied verb forms; (viii) In Sinhala, when the noun is non-living, the verb remains singular regardless of the noun's plurality; (ix) Conversely, when the noun includes the respectable suffix, the verb becomes plural, irrespective of the noun's animacy, plurality, or singularity [12]. The lack of effective NLP tools, as well as the peculiarities of the Sinhala language, make the stylometric feature extraction process more difficult [12].

As per findings in NLP tools and research, Sinhala continues to be categorized as a language lacking in resources, lacking both the economic impetus and linguistic infrastructure of its relative, English. Consequently, our objective is to pinpoint discernible features within Sinhala texts that can proficiently facilitate authorship attribution. Additionally, a significant hurdle in authorship attribution of Sinhala texts lies in the dearth of comprehensive datasets tailored to the language's nuances. Hence, our aim is to construct a benchmark corpus conducive to conducting cross-topic authorship attribution, encompassing a substantial pool of potential authors

Most previous research on authorship attribution assumes that the training and test data are drawn from the same distribution, known as intra-topic authorship attribution. Although it simplifies the authorship attribution task, this assumption is unrealistic in real scenarios [35, 45]. Another goal of this paper is to formulate a solution for cross-topic authorship attribution where the training and testing data come from two different topics. Performing cross-topic authorship attribution requires the extraction of topic-independent writing style markers (i.e., stylometric features).

This study aims to investigate using function words as stylometric features to perform cross-topic authorship attribution. Function words refer to the words that have less impact on the lexical meaning of a sentence and express the grammatical link between words within a sentence. They can be considered as a marker of an author's writing style. This is because function words are used less frequently within the author's conscious control during writing. As a result, these terms serve as a strong foundation for comparing writing styles [5, 25, 33, 45].

Function words are discussed in more detail in Section 4. Although few studies have been conducted on cross-topic authorship attribution for the English language, cross-topic authorship attribution is not limited to a particular language. This study focuses on performing cross-topic authorship attribution for the Sinhala texts. Moreover, it has been reported in previous authorship attribution studies that increasing the size of the candidate authors reduces the accuracy of the authorship attribution task significantly. Furthermore, existing solutions are not designed to work in data-poor environments where each candidate author only has a small number of writing samples (i.e., less than 10) available. To handle real-world scenarios, however, an authorship attribution solution should be able to handle many candidate authors in data-poor conditions. A summary of the objectives of this paper is as follows:

Summary of Objectives:

- Introduce the first cross-topic authorship attribution dataset for Sinhala;
- Present solution that can handle both variations of the authorship attribution task for Sinhala: (i) intra-topic authorship attribution and (ii) cross-topic authorship attribution;
- Evaluate the effectiveness of function words to perform authorship attribution for Sinhala;
- Unlike existing approaches for authorship attribution of English texts, create a solution that can handle a large number of candidate authors while also working in data-poor conditions with a limited number of writing samples for each candidate author.

In addition to achieving the aforementioned objectives, we answer the following research questions.

Research Questions.

- **Research Question 1.** Can we effectively conduct authorship attribution for Sinhala text, both within and across different topics, relying solely on function words as stylometric features?
- **Research Question 2.** How does the quantity of candidate authors impact the accuracy of authorship attribution in Sinhala texts?
- **Research Question 3.** What is the optimal number of function words needed as features to accurately perform authorship attribution in Sinhala text?

Summary of Our Contributions. The contribution of this work includes:

- (1) We present the first comprehensive study on an effective cross-topic authorship attribution solution for Sinhala, which can achieve higher accuracy levels.
- (2) We demonstrate the first comprehensive study proving that function words are strong discriminators for cross-topic authorship attribution for Sinhala.
- (3) We create a new, significantly larger Sinhala cross-topic authorship attribution corpus.

- (4) We used a corpus of 980 Sinhala documents from 140 writers in numerous experiments, much larger than many of the existing English datasets. Experiments are carried out to demonstrate that our solution can perform well in data-poor conditions with a limited number of writing samples per author; (ii) handle a large number of candidate authors; and (iii) achieve higher levels of accuracy for both intra- and cross-topic authorship attribution variations.

The rest of the paper is laid out as follows. Section 2 examines previous research on authorship identification. Our corpus is depicted in Section 3. Our solution is described in Section 4. The experimental results are presented in Section 5. The concluding observations are found in Section 6.

2 LITERATURE REVIEW

This section aims to provide additional information on practical and successful experiments on Authorship attribution and Cross-topic authorship attribution. We divide this section into four subsections: (i) Authorship attribution studies, (ii) Cross-topic authorship attribution studies, (iii) Stylometric Features, and (iv) Classification methods.

2.1 Authorship attribution studies

Mendenhall [37] was among the first to use the concept of authorship attribution on Shakespeare's plays, followed by Yule's [40, 63], and Zipf's [26] statistical studies in the first half of the twentieth century. Many detailed studies were later conducted to differentiate between candidate authors [18, 22]. These studies had several limitations, including an excessive length of textual data, an insufficient number of candidate authors, and a lack of appropriate benchmark data [41]. However, with the advancement of electronic texts, authorship attribution has become one of the fields in natural language processing, data mining, and stylometry that has received increased attention in recent years [1, 2, 16, 21]. Furthermore, efficient information retrieval techniques, powerful machine learning algorithms that handle multidimensional and sparse data, and efficient text analysis tools have all influenced the growth of the authorship attribution task. Two main paradigms have been used in authorship attribution; (i) similarity-based paradigm, where some criteria are used to calculate the distance between two documents and an anonymous document is compared to determine their similarity, or (ii) machine-learning paradigm, where known documents of candidate authors are used to build and train a classifier that can be used to classify an anonymous document [29]. Furthermore, there are several other authorship analysis tasks which can be defined as follows: (i) Authorship verification of a given text [23], (ii) Plagiarism detection between two texts [20], (iii) Author profiling by obtaining information about the author of a given text's age, education, gender, and so on [28].

2.2 Cross-topic authorship attribution studies

Cross-topic authorship attribution is a sub-topic in authorship attribution. The main difference between these two is that authorship attribution uses the same topic documents to perform training and testing, while cross-topic authorship attribution trains on one topic domain and tests on a different topic domain. The scarcity of cross-domain corpora has hampered the use of cross-topic authorship attribution. However, the studies that have been conducted on this task have yielded significant results. Corney [11] investigates the possibility of cross-domain authorship attribution using an email corpus from a small group of authors on a specific set of topics. De vel et al. [13] analyzed a corpus of 156 native English documents from three authors on three topics. Madigan et al. [32] also experimented with cross-domain social media texts. Koppel et al. successfully tested the unmasking method in cross-topic conditions for author verification of extended papers [30].

Sapkota et al. investigated the performance of several baseline cross-topic authorship attribution methods and discovered combining several topics in the training set improves the ability to identify the authors of texts on a different topic [45]. These authors were able to achieve notifiable results, however, their studies were limited to English, and their methods are not scalable in terms of the size of candidate authors with a limited number of samples.

2.3 Stylometry Features

The following four types of features are commonly used in authorship attribution studies focus on English. To the best of our knowledge, none of the following features have been investigated for the authorship attribution task of Sinhala documents.

Lexical features. These features are also known as bag-of-words because of the exclusion of the function words. Many studies have been conducted to investigate the effectiveness of lexical features in authorship attribution [14, 17, 53]. These features are simple and efficient, but they disregard the order of words, carry thematic information, and are not reliable for cross-topic authorship attribution.

Function Words. Although these features lack the semantic meaning of texts, they have been used in many authorship attribution studies and have proven to be very effective in cross-domain authorship attribution [5, 25, 33, 45].

Structural Features. Generally refers to the structure of a document (i.e. number of sentences, number of tokens per sentence). A few studies have investigated the effectiveness of stylistic features on the authorship attribution task [8, 61].

Character *n*-grams. They are one of the features that have been commonly used in authorship attribution studies [24, 62] because of their ability to capture both the theme and stylistic information of the texts. These features are simply viewed as a series of characters. However, they are known to be carrying the thematic information of the texts as well.

2.4 Classification Methods

Many classifiers have been used to perform Authorship Attribution tasks in recent years. We implemented seven different classifiers and compared their performance on this task using rapid miner¹ with their default parameter settings. This section will provide a brief overview of each classifier.

***k*-Nearest Neighbors** stores all instances in *n*-dimensional space that correspond to training data points. KNN analyzes the closest *k* saved instances and returns the most common class as the predicted class of an unknown instance. This algorithm has also been shown to be effective in the attribution of authorship [31].

Naive Bayes is a probabilistic classification algorithm that was developed based on the Bayes theorem. The Naive Bayes assumption holds that all features are conditionally independent. Although this assumption is usually unrealistic, Naive Bayes has proven to be an efficient authorship attribution classifier with a low computational cost [4, 19].

Logistic Regression is a statistical algorithm fitting a logical curve to the dataset. The main objective behind logistic regression is to apply a nonlinear sigmoidal function on a set of linear set of features. It can perform both binary and multinomial classifications [27].

Deep Learning model consists of multiple layers of perceptrons, with each layer learning concepts from the data on the previous layer. These approaches have been used in a variety of applications, including authorship attribution, and have resulted in improved performance [43, 44, 58].

¹<https://rapidminer.com/>

Decision Tree is an algorithm that builds tree-structured models to perform classification and regression. It employs a set of if-then rules that are mutually exclusive and exhaustive. Using the training data, the rules are learned one by one. Each time a rule is established, the tuples covered by the rules are eliminated, and the process is repeated until a termination condition is met. The decision tree is one of the techniques that has been used frequently in Authorship Attribution [36].

Random Forest is a combination of multiple decision tree predictors in which these trees vote on the referred class, with the highest voted class being considered the final predicted class [9].

Support Vector Machine is a widely used algorithm for classifying data based on super finite degrees of polarity. It finds the hyperplane that best distinguishes the instances. The best hyperplane is the one with the greatest distance between each instance. The larger the data, the more accurate the prediction is [10].

3 DATA COLLECTION

There is currently no reliable corpus available for the Sinhala authorship attribution task. To perform authorship attribution we require a dataset containing text samples, where each sample is labelled with the name of the true author. To perform experiments, we created a new Sinhala authorship attribution corpus extracted from an online newspaper website, *Lankadeepa*². Python was used to write the scraper, and data was extracted in two steps: (i) retrieve all URLs on each page of the website, and (ii) extract the text and the name of the author from the website based on the retrieved URLs. Our corpus contains 980 Sinhala documents from 140 authors in four domains. Furthermore, each candidate author has seven writing samples available, and on average, there are 245 samples per Topic, which is a more realistic scenario in which many writing samples may not be available. Table 1 summarizes the dataset used to investigate the study.

Table 1. The summary of the Dataset

Topic	# Authors	# Documents
Courts	35	245
Business	35	245
Politics	35	245
General	35	245
Total	140	980

4 METHODOLOGY AND EXPERIMENTAL SETUP

4.1 Methodology

In this subsection, we discuss our solution. Figure 1 shows the overview of our solution. We perform authorship attribution in three steps: (i) Feature extraction, (ii) similarity search, and (iii) authorship attribution.

Features Extraction. In the first step, we extract function words from each text sample. Function words, also known as grammatical words or structure words, are words that have little semantic content on their own but play important roles in the structure of sentences and convey grammatical relationships between other words. Examples of function words include articles (e.g., "the", "a"), conjunctions (e.g., "and", "but"), prepositions (e.g., "in", "on"), pronouns (e.g., "he", "she", "it"), and auxiliary verbs (e.g., "is", "has", "will"). Function words are considered reliable stylometric features for authorship attribution tasks for several reasons: (I) Function words tend to occur frequently

²<http://www.lankadeepa.lk/>

in texts, and their distribution can vary significantly between authors. Since function words are fundamental to the structure of language, authors may have distinctive preferences in their usage of function words, leading to unique patterns that can be used to distinguish one author’s style from another. (II) Unlike content words (words carrying significant semantic content such as nouns, verbs, and adjectives), the usage of function words tends to be more stable across different topics and genres. This stability makes them particularly useful for authorship attribution tasks where the focus is on identifying an author’s unique style rather than the content of the text. (III) Authors may consciously alter their use of content words to mimic the style of another author or to disguise their own style. However, function words are typically used subconsciously and are less likely to be intentionally manipulated. This makes them more reliable indicators of an author’s natural writing style. (IV) Function words are common across languages, and while their specific usage patterns may vary between languages, the concept of function words and their importance in conveying grammatical relationships remains consistent. This language-independence makes function words valuable features for authorship attribution tasks in multilingual contexts. (V) Function words often exhibit distinctive statistical properties such as Zipfian distributions, where a few function words are very common while the majority are relatively rare. These statistical properties can be exploited for feature selection and machine learning algorithms in authorship attribution tasks. As a result, they can be useful in determining authorship. Function words are commonly considered articles, prepositions, and conjunctions. In the Sinhala Language, there are 191 function words. The list of Sinhala function words can be accessed using the following link³.

Table 2. Function words in the Sinhala Language

සමග	සිට	වෙතින්	ඉක්බිති	එම්බල	මෙන්	මගින්	සහ	මිස	සේක
සමඟ	දී	වෙතට	දැන්	බොල	සේ	හෝ	දක්වා	මුත්	ගැන
අහා	මහා	වෙනුවෙන්	යලි	නම්	වැනි	ඉතා	ට	කිම	අනුව
ආප්	මහ	වෙනුවට	පුන	වනාහි	බදු	ඒ	ගේ	කිමි	පරිදි
ආ	පමණ	ඉතින්	ඉතින්	කලී	වත්	එම	එ	අයි	විට
මිහෝ	පමණක්	ගැන	සිට	ඉදුරා	අයුරු	ද	ක	මන්ද	තෙක්
අනේ	පමන	නැ	සිටත්	අන්ත	අයුරින්	අතර	ක්	හෙවත්	මෙතෙක්
අදේ	වන	අනුව	පටත්	ඔත්ත	ලෙස	විසින්	බවත්	නොහොත්	මේතෙක්
අපොයි	විට	නව	තෙක්	මෙන්ත	වැඩි	සමග	බවද	පතා	තුරු
අපෝ	විටින්	පිළිබද	දක්වා	උදෙසා	ශ්‍රී	පිළිබදව	මන	පාසා	තුරා
අයිසෝ	මේ	විශේෂ	සා	පිණිස	හා	පිළිබද	අතුලු	ගානෙ	තුරාවට
ආයි	මෙලෙස	දැනට	තාක්	සඳහා	ය	තුළ	අතුළු	නව	තුලින්
ඌයි	මෙයින්	එහෙත්	තුටුක්	අරබයා	නිසා	බව	මෙසේ	ඉතා	නමුත්
වී	ඇති	මෙහෙත්	පවා	නිසා	නිසාවෙන්	වැනි	වඩා	බොහෝ	එනමුත්
වින්	ලෙස	එහේ	ද	එනිසා	බවට	මහ	වඩාත්ම	වහා	වස්
වික්	සිදු	මෙහේ	හෝ	එබැවින්	බව	මෙම	නිති	සෙද	මෙන්
හෝ	වගයෙන්	ම	වත්	බැවින්	බවෙන්	මෙහි	නිතින්	සැතියන්	ලෙස
දෝ	යන	නවත්	විනා	හෙයින්	නම්	මේ	නිතොර	හනික	පරිදි
දෝහෝ	සඳහා	නව	හැර	සේක්	වැඩි	වෙන	නිතර	එම්බා	එහෙත්

Similarity Search. After extracting the function words, we calculated the frequencies of function words and used them as features of all conducting experimental studies. In the second step, we perform a similarity search to identify a small set of candidate authors. Recall that most existing authorship attribution solutions designed for other languages drop the accuracy as the number of candidate authors increases in the candidate author set. To handle such a situation, we perform a

³<https://github.com/nlpcuom/Sinhala-Stopword-list>

similarity search to identify a small set of candidate authors. We then apply different classification methods on the small set of candidate authors to identify the original author of the query text. The similarity between two text samples (feature vectors) is computed using a cosine similarity measure. We vary the number of text samples to be considered for the small candidate author set. We denote it with σ . Specifically, we empirically tested different values of σ , and 5 resulted in the best accuracy. We note that we do not have control over the number of candidate authors appearing in the small set. However, the value of σ indicates the maximum number of authors that can be included in the small set of candidate authors. For example, when we set the value of σ to 5 the maximum number of candidate authors that may appear in the small subset of candidate authors is five. As a result, we have reduced the number of candidate authors from 140 to 5 only, and it significantly increases the accuracy of the authorship attribution task.

The main advantages of using similarity search before the classification method are as follows:

(i) Similarity search results in a small number of candidates and classification can be used to further refine the list, eliminating candidates that might have high similarity scores but are not relevant to the specific task or problem at hand. (ii) Similarity search is effective at identifying candidates that are similar to a given input. However, it might not account for variations or distinctions that could be crucial in certain contexts. By employing a classification step after narrowing down the candidates using Similarity search, we can ensure that the final selection includes candidates that not only are similar but also fulfil specific classification criteria. (iii) By employing a classification step, we can reduce the likelihood of false positives that might arise from relying solely on similarity-based selection. Classification can help filter out candidates that might have high similarity scores but do not belong to the intended category. (iv) Similarity Search can be broad and might retrieve candidates that are somewhat related but not precisely aligned with the desired criteria. Classification helps enhance precision by ensuring that the final candidates are not only similar but also meet specific classification guidelines.

Authorship Attribution. After we obtain a small set of candidate authors, we then try different classifiers on these authors set, such as (i) k-Nearest Neighbors, (ii) Naive Bayes, (iii) Logistic Regression, (iv) Deep Learning, (v) Decision Tree, (vi) Random Forest, and (vii) Support Vector Machine.

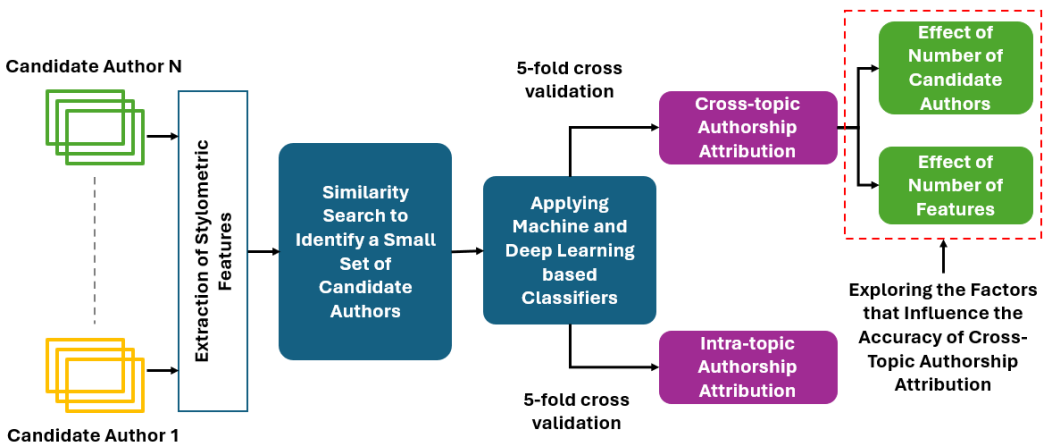


Fig. 1. The illustration of Sinhala authorship attribution framework.

4.2 Experimental Setup

In this subsection, we discuss the evaluation measures to evaluate our solution's performance and the evaluation strategy to conduct several experimental studies.

- **Evaluation measures.** We used accuracy as an evaluation measure. The accuracy of the authorship attribution is computed as follows. A prediction is considered correct if the true author of the test sample is identified correctly.
- **Evaluation strategy.** All the experimental studies are conducted using five-fold cross-validation unless stated otherwise.

4.3 Experimental Studies

We have conducted the following studies to achieve this investigation's objectives and answer the research questions listed in the Introduction section of this paper.

- **Intra-topic authorship attribution.** In this context, we used intra-topic authorship attribution, which is the simplest (yet most unrealistic) scenario in which the training and testing data are from the same topic (i.e. train on courts and test on courts). The author's style is more likely to be the most important factor in distinguishing between texts. Five-fold cross-validation was used to divide the dataset into training and testing subsets to generalize the model.
- **Cross-topic authorship attribution.** Next, we adopt *k-Nearest Neighbours* to conduct Cross-topic authorship attribution. In Cross-topic authorship attribution, we train the model on one topic and test on a different topic (i.e., Train on Courts and test on Business, Politics, and General). Therefore, in this case, the distribution of test texts across candidate authors differs from that of training texts.

The experiment for cross-topic authorship attribution was further investigated to evaluate the classification model's performance, using two factors that could influence the model's behaviour.

(i) **The effect of the number of candidate authors.** In this study, we vary the number of candidate authors by generating four different datasets containing 35, 70, 105, and 140 candidate authors.

(ii) **The effect of the number of features on the full dataset.** In this study, we vary the number of features on the full dataset. We varied the number of features in this experiment by using the most frequently used function words (i.e., we try the most frequent 50 function words, then the most frequent 100 function words, and so on). The features chosen were 50, 100, 150, and 191.

5 EXPERIMENTAL RESULTS AND DISCUSSION

In this section, we present the findings of various experiments that we conducted. We will begin by discussing the findings of Intra-topic authorship attribution, followed by the findings of intra-topic and cross-topic authorship attribution tasks.

5.1 Intra-topic Authorship Attribution

Table 3 shows the classification accuracy results for each topic domain using seven different classifiers including K-Nearest Neighbors, Naive Bayes, Logistic Regression, Deep Learning, Decision Trees, Random Forests, and Support Vector Machines. As can be seen, the k-NN classifier achieved the highest classification accuracy in almost every topic domain except for courts. As for Courts, Business, Politics, and General, the k-NN classifier achieved a classification accuracy of 84.03%, 84.84%, 85.44%, and 86.09%, respectively. The main reason for k-NN's achievements could be that it

performs classification through a comparison with instances stored in memory instead of building a generalized model. In addition to this: (i) there is no information loss through generalization [3]; (ii) it can learn from a limited set of examples [7]; (iii) little or no training is required to perform classification task [3]; (iv) it can incrementally add new information at runtime [7]; (v) it is a non-parametric method and does not require a prior knowledge relating to probability distributions for the classification problem [34].

Furthermore, the Support Vector Machine achieved the second-best results yielding a classification accuracy of 84.11%, 84.06%, 85.12%, and 85.72% for Courts, Business, Politics, and General, respectively. When compared to other classification models, both models performed significantly better.

Table 3. Intra-topic authorship attribution using Function words (i.e., training on Courts, Test on Courts)

Classifier	Accuracy			
	Courts	Business	Politics	General
k-Nearest Neighbors	84.03%	84.84%	85.44%	86.09%
Naive Bayes	68.27%	69.76%	70.45%	70.14%
Logistic Regression	55.87%	60.01%	58.75%	59.65%
Deep Learning	67.77%	68.84%	65.76%	68.92%
Decision Tree	79.80%	77.28%	75.88%	78.36%
Random Forest	53.56%	56.09%	57.89%	58.87%
Support Vector Machines	84.11%	84.06%	85.12%	85.72%

5.2 Cross-topic Authorship Attribution

Table 4 provides classification accuracy for each topic domain across multiple topic domains. We chose k-NN as the classification algorithm because it performed well in terms of Intra-topic authorship attribution. The highest accuracy was obtained when the model was trained on Courts and tested on General, which was 85.78%. The overall performance of Cross-topic authorship attribution was found to be effective. This supports our hypothesis and answers one of our research questions: documents written on one topic can be reliably predicted using a model developed using documents from multiple other topics. This demonstrates that authors employ a consistent writing style across topics. Furthermore, the effectiveness of using function words for Cross-topic authorship attribution is demonstrated in our results. This is an interesting discovery because we discovered that by using function words, the cross-topic authorship attribution problem can be solved just as effectively as the Intra-topic authorship attribution problem.

The Effect of Number of Candidate Authors. We wanted to demonstrate the effect of varying the number of candidate authors. To observe the performance, we created four different datasets and K-NN was used for classification. As can be seen in Table 5, the accuracy increases as the number of candidate authors reduces. The best accuracy is achieved when the number of candidate authors is set to 35. However, the model performed generally well when the number of candidate authors was 140 (total number of authors). The model achieved an accuracy of 85.81% when we experimented using all the authors from every topic domain.

The Effect of Number of Features. Next, we investigated the effect of varying the number of features on the full dataset. Table 6 illustrates the results we obtained. k-NN performance increases as the number of features increases. The highest accuracy, 85.81% was obtained when the number of features was set to be 191 while the lowest accuracy was obtained when the number of features was set to 50.

Table 4. Cross-topic authorship attribution: using k-nearest neighbours

Test	Training			
	Accuracy			
	Courts	Business	Politics	General
Courts	84.03%	84.01%	83.91%	83.55%
Business	84.03%	84.84%	83.70%	83.15%
Politics	84.86%	85.16%	85.44%	84.96%
General	85.78%	84.67%	85.65%	86.09%

Table 5. Cross-topic authorship attribution: Effect of number of candidate authors on the performance of the cross-topic authorship attribution using k-nearest neighbours

Number of Candidate Authors	35	70	105	140
Accuracy	87.66%	86.84%	86.48%	85.81%

Table 6. Cross-topic authorship attribution: Effect of number of features on the performance of the cross-topic authorship attribution using k-nearest neighbours

Number of Features on Full Dataset	50	100	150	191
Accuracy	55.16%	66.81%	69.65%	85.81%

6 CONCLUSIONS

This is the first investigation to explore Sinhala text’s authorship attribution task in intra- and cross-topic authorship attribution settings. Our experiments used seven machine learning classifiers for intra-topic authorship attribution. These experiments revealed that the most effective algorithm is k-Nearest Neighbors, which we used to perform cross-topic authorship attribution and achieved high accuracy. We also discovered that function words, combined with an appropriate classification method, are a sufficient style marker for distinguishing between authors. We also discovered that our solution can handle more candidate authors and that increasing the number of features enhances the accuracy of the authorship attribution task. We used function words as style markers because, unlike other types of stylometric features, they can be reliably extracted for Sinhala texts. More research is required to determine the impact of other types of features and compare performance on authorship attribution for the Sinhala language. In future, we plan to apply our approach to different types of texts and different scenarios such as open-set authorship attribution. We also plan to investigate the impact of text length on attribution accuracy, and the applicability of the proposed solution to other languages with limited language processing tools.

REFERENCES

- [1] Ahmed Abbasi and Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20, 5 (2005), 67–75.
- [2] Muhammad Kashif Afzal, Matthew Shardlow, Suppawong Tuarob, Farooq Zaman, Raheem Sarwar, Mohsen Ali, Naif Radi Aljohani, Miltiades D Lytras, Raheel Nawaz, and Saeed-Ul Hassan. 2023. Generative image captioning in Urdu using deep learning. *Journal of Ambient Intelligence and Humanized Computing* 14, 6 (2023), 7719–7731.
- [3] Ethem Alpaydin. 2021. *Machine learning*. MIT press.
- [4] Alaa Saleh Altheneyan and Mohamed El Bachir Menai. 2014. Naïve Bayes classifiers for authorship attribution of Arabic texts. *Journal of King Saud University-Computer and Information Sciences* 26, 4 (2014), 473–484.
- [5] Bagher BabaAli. 2021. Online writer identification using statistical modeling-based feature embedding. *Soft Computing* 25 (2021), 9639–9649.

- [6] Abu Bakar, Raheem Sarwar, Saeed-Ul Hassan, and Raheel Nawaz. 2023. Extracting Algorithmic Complexity in Scientific Literature for Advance Searching. *Journal of Computational and Applied Linguistics* 1 (2023), 39–65.
- [7] Stephen D Bay. 1999. Nearest neighbor classification from multiple feature subsets. *Intelligent data analysis* 3, 3 (1999), 191–209.
- [8] Mudit Bhargava, Pulkit Mehndiratta, and Krishna Asawa. 2013. Stylometric analysis for authorship attribution on twitter. In *Big Data Analytics: Second International Conference, BDA 2013, Mysore, India, December 16-18, 2013, Proceedings 2*. Springer, 37–47.
- [9] Leo Breiman. 2001. Random forests. *Machine learning* 45 (2001), 5–32.
- [10] Na Cheng, Rajarathnam Chandramouli, and Koduvayur P Subbalakshmi. 2011. Author gender identification from text. *Digital investigation* 8, 1 (2011), 78–88.
- [11] Malcolm W Corney. 2003. *Analyzing e-mail text authorship for forensic purposes*. Ph.D. Dissertation. Queensland University of Technology.
- [12] Nisansa De Silva. 2019. Survey on publicly available sinhala natural language processing tools and research. *arXiv preprint arXiv:1906.02358* (2019).
- [13] Olivier De Vel, Alison Anderson, Malcolm Corney, and George Mohay. 2001. Mining e-mail content for author identification forensics. *ACM Sigmod Record* 30, 4 (2001), 55–64.
- [14] Joachim Diederich, Jörg Kindermann, Edda Leopold, and Gerhard Paass. 2003. Authorship attribution with support vector machines. *Applied intelligence* 19 (2003), 109–123.
- [15] Wilhelm Geiger. 1995. *A grammar of the Sinhalese language*. Asian educational services.
- [16] Muhammad Umair Hassan, Saleh Alaliyat, Raheem Sarwar, Raheel Nawaz, and Ibrahim A Hameed. 2023. Leveraging deep learning and big data to enhance computing curriculum for industry-relevant skills: A Norwegian case study. *Heliyon* 9, 4 (2023).
- [17] Muhammad Umair Hassan, Xiuyang Zhao, Raheem Sarwar, Naif R Aljohani, and Ibrahim A Hameed. 2024. SODRet: Instance retrieval using salient object detection for self-service shopping. *Machine Learning with Applications* 15 (2024), 100523.
- [18] David I Holmes. 1998. The evolution of stylometry in humanities scholarship. *Literary and linguistic computing* 13, 3 (1998), 111–117.
- [19] Fatma Howedi and Masnizah Mohd. 2014. Text classification for authorship attribution using Naive Bayes classifier with limited training data. *computer engineering and intelligent systems* 5, 4 (2014), 48–56.
- [20] Hamed Jelodar, Yongli Wang, Gang Xiao, Mahdi Rabbani, Ruxin Zhao, Seyedvalyallah Ayobi, Peng Hu, and Isma Masood. 2021. Recommendation system based on semantic scholar mining and topic modeling on conference publications. *Soft Computing* 25 (2021), 3675–3696.
- [21] Patrick Juola. 2006. Authorship attribution for electronic documents. In *IFIP International Conference on Digital Forensics*. Springer, 119–130.
- [22] Patrick Juola et al. 2008. Authorship attribution. *Foundations and Trends® in Information Retrieval* 1, 3 (2008), 233–334.
- [23] Ravneet Kaur, Sarbjeet Singh, and Harish Kumar. 2021. An intrinsic authorship verification technique for compromised account detection in social networks. *Soft Computing* 25 (2021), 4345–4366.
- [24] 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. 255–264.
- [25] Mike Kestemont. 2014. Function words in authorship attribution. From black magic to theory?. In *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL)*. 59–66.
- [26] George Kingsley Zipf. 1932. *Selected studies of the principle of relative frequency in language*. Harvard university press.
- [27] David G Kleinbaum and Mitchel Klein. 1996. *Survival analysis a self-learning text*. Springer.
- [28] Moshe Koppel, Shlomo Argamon, and Anat Rachel Shimoni. 2002. Automatically categorizing written texts by author gender. *Literary and linguistic computing* 17, 4 (2002), 401–412.
- [29] Moshe Koppel, Jonathan Schler, and Shlomo Argamon. 2011. Authorship attribution in the wild. *Language Resources and Evaluation* 45 (2011), 83–94.
- [30] Moshe Koppel, Jonathan Schler, and Elisheva Bonchek-Dokow. 2007. Measuring differentiability: Unmasking pseudonymous authors. *Journal of Machine Learning Research* 8, 6 (2007).
- [31] Yun Lin and Jie Wang. 2014. Research on text classification based on SVM-KNN. In *2014 IEEE 5th International Conference on Software Engineering and Service Science*. IEEE, 842–844.
- [32] David Madigan, Alexander Genkin, David D Lewis, Shlomo Argamon, Dmitriy Fradkin, and Li Ye. 2005. Author identification on the large scale. In *Proceedings of the 2005 Meeting of the Classification Society of North America (CSNA)*.
- [33] Tanya Malhotra and Anjana Gupta. 2023. A systematic review of developments in the 2-tuple linguistic model and its applications in decision analysis. *Soft Computing* 27, 4 (2023), 1871–1905.
- [34] Chengsheng Mao, Bin Hu, Philip Moore, Yun Su, and Manman Wang. 2015. Nearest neighbor method based on local distribution for classification. In *Advances in Knowledge Discovery and Data Mining: 19th Pacific-Asia Conference*,

- PAKDD 2015, Ho Chi Minh City, Vietnam, May 19-22, 2015, *Proceedings, Part I* 19. Springer, 239–250.
- [35] Iliia Markov, Efstathios Stamatatos, and Grigori Sidorov. 2018. Improving cross-topic authorship attribution: The role of pre-processing. In *Computational Linguistics and Intelligent Text Processing: 18th International Conference, CICLing 2017, Budapest, Hungary, April 17–23, 2017, Revised Selected Papers, Part II* 18. Springer, 289–302.
- [36] Rangsipan Marukatat, Robroo Somkiadcharoen, Ratthanan Nalintasnai, and Tappasarn Aramboonpong. 2014. Authorship attribution analysis of thai online messages. In *2014 International Conference on Information Science & Applications (ICISA)*. IEEE, 1–4.
- [37] Thomas Corwin Mendenhall. 1887. The characteristic curves of composition. *Science* 214s (1887), 237–246.
- [38] Emad Mohamed, Raheem Sarwar, and Sayed Mostafa. 2023. Translator attribution for Arabic using machine learning. *Digital Scholarship in the Humanities* 38, 2 (2023), 658–666.
- [39] Sarana Nutanong, Chenyun Yu, Raheem Sarwar, Peter Xu, and Dickson Chow. 2016. A scalable framework for stylometric analysis query processing. In *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, 1125–1130.
- [40] Dimas Wibisono Prakoso, Asad Abdi, and Chintan Amrit. 2021. Short text similarity measurement methods: a review. *Soft Computing* 25 (2021), 4699–4723.
- [41] Joseph Rudman. 1997. The state of authorship attribution studies: Some problems and solutions. *Computers and the Humanities* 31 (1997), 351–365.
- [42] Hadeel Saadany, Emad Mohamed, and Raheem Sarwar. 2023. Towards a better understanding of Tarajem: creating topological networks for Arabic biographical dictionaries. *Journal of Data Mining & Digital Humanities* (2023).
- [43] Fahad Sabah, Yuwen Chen, Zhen Yang, Muhammad Azam, Nadeem Ahmad, and Raheem Sarwar. 2023. Model optimization techniques in personalized federated learning: A survey. *Expert Systems with Applications* (2023), 122874.
- [44] Fahad Sabah, Yuwen Chen, Zhen Yang, Abdul Raheem, Muhammad Azam, and Raheem Sarwar. 2023. Heart Disease Prediction with 100% Accuracy, Using Machine Learning: Performance Improvement with Features Selection and Sampling. In *2023 8th IEEE International Conference on Network Intelligence and Digital Content (IC-NIDC)*. IEEE, 41–45.
- [45] Upendra Sapkota, Tamar Solorio, Manuel Montes, Steven Bethard, and Paolo Rosso. 2014. Cross-topic authorship attribution: Will out-of-topic data help?. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*. 1228–1237.
- [46] Raheem Sarwar. 2022. Author Gender Identification for Urdu Articles. In *International Conference on Computational and Corpus-Based Phraseology*. Springer, 221–235.
- [47] Raheem Sarwar and Saeed-Ul Hassan. 2021. Urduai: Writeprints for Urdu authorship identification. *Transactions on Asian and Low-Resource Language Information Processing* 21, 2 (2021), 1–18.
- [48] Raheem Sarwar, Qing Li, Thanawin Rakthanmanon, and Sarana Nutanong. 2018. A scalable framework for cross-lingual authorship identification. *Information Sciences* 465 (2018), 323–339.
- [49] Raheem Sarwar and Emad Mohamed. 2022. Author verification of nahj al-balagha. *Digital Scholarship in the Humanities* 37, 4 (2022), 1210–1222.
- [50] Raheem Sarwar and Sarana Nutanong. 2016. The key factors and their influence in authorship attribution. *Res. Comput. Sci.* 110 (2016), 139–150.
- [51] Raheem Sarwar, Thanasarn Porthaveepong, Attapol Rutherford, Thanawin Rakthanmanon, and Sarana Nutanong. 2020. StyloThai: A scalable framework for stylometric authorship identification of thai documents. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, 3 (2020), 1–15.
- [52] 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 Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* 19, 4 (2020), 1–19.
- [53] Raheem Sarwar, Pin Shen Teh, Fahad Sabah, Raheel Nawaz, Ibrahim A Hameed, Muhammad Umair Hassan, et al. 2024. AGI-P: A Gender Identification Framework for Authorship Analysis Using Customized Fine-Tuning of Multilingual Language Model. *IEEE Access* (2024).
- [54] Raheem Sarwar, Norawit Uraileertprasert, 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.
- [55] Raheem Sarwar, Chenyun Yu, Sarana Nutanong, Norawit Uraileertprasert, Nattapol Vannaboot, and Thanawin Rakthanmanon. 2018. A scalable framework for stylometric analysis of multi-author documents. In *Database Systems for Advanced Applications: 23rd International Conference, DASFAA 2018, Gold Coast, QLD, Australia, May 21-24, 2018, Proceedings, Part I* 23. Springer, 813–829.
- [56] 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.

- [57] Jacques Savoy. 2013. Authorship attribution based on a probabilistic topic model. *Information Processing & Management* 49, 1 (2013), 341–354.
- [58] Prasha Shrestha, Sebastian Sierra, Fabio A González, Manuel Montes, Paolo Rosso, and Thamar Solorio. 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: Volume 2, short papers*. 669–674.
- [59] Kanishka Silva, Burcu Can, Frédéric Blain, Raheem Sarwar, Laura Ugolini, and Ruslan Mitkov. 2023. Authorship attribution of late 19th century novels using GAN-BERT. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*. 310–320.
- [60] Kanishka Silva, Ingo Frommholz, Burcu Can, Fred Blain, Raheem Sarwar, and Laura Ugolini. 2024. Forged-GAN-BERT: Authorship Attribution for LLM-Generated Forged Novels. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*. 325–337.
- [61] Efstathios Stamatatos. 2006. Authorship attribution based on feature set subsampling ensembles. *International Journal on Artificial Intelligence Tools* 15, 05 (2006), 823–838.
- [62] Ph D Stamatatos et al. 2013. On the robustness of authorship attribution based on character n-gram features. *Journal of Law and Policy* 21, 2 (2013), 7.
- [63] C Udney Yule. 2014. *The statistical study of literary vocabulary*. Cambridge University Press.