IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 7 December 2023, accepted 20 January 2024, date of publication 24 January 2024, date of current version 1 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3358199

RESEARCH ARTICLE

AGI-P: A Gender Identification Framework for Authorship Analysis Using Customized Fine-Tuning of Multilingual Language Model

RAHEEM SARWAR[1], LE AN HA[2], PIN SHEN TEH[1], FAHAD SABAH[3], RAHEEL NAWAZ[4], IBRAHIM A. HAMEED[5], (Senior Member, IEEE), AND MUHAMMAD UMAIR HASSAN[5]

[1]Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, M15 6BH Manchester, U.K.
[2]Research Group in Computational Linguistics, RIILP, University of Wolverhampton, WV1 1LY Wolverhampton, U.K.
[3]Faculty of Information Technology, Beijing University of Technology, Beijing 100021, China
[4]Executive Office, Staffordshire University, ST4 2DE Stoke-on-Trent, U.K.
[5]Department of ICT and Natural Sciences, Norwegian University of Science and Technology, 6009 Ålesund, Norway

Corresponding author: Muhammad Umair Hassan (muhammad.u.hassan@ntnu.no)

This work was supported by the Norwegian University of Science and Technology (NTNU), Norway.

ABSTRACT In this investigation, we propose a solution for the author's gender identification task called AGI-P. This task has several real-world applications across different fields, such as marketing and advertising, forensic linguistics, sociology, recommendation systems, language processing, historical analysis, education, and language learning. We created a new dataset to evaluate our proposed method. The dataset is balanced in terms of gender using a random sampling method and consists of 1944 samples in total. We use accuracy as an evaluation measure and compare the performance of the proposed solution (AGI-P) against state-of-the-art machine learning classifiers and fine-tuned pre-trained multilingual language models such as DistilBERT, mBERT, XLM-RoBERTa, and Multilingual DEBERTa. In this regard, we also propose a customized fine-tuning strategy that improves the accuracy of the pre-trained language models for the author gender identification task. Our extensive experimental studies reveal that our solution (AGI-P) outperforms the well-known machine learning classifiers and fine-tuned pre-trained multilingual language models with an accuracy level of 92.03%. Moreover, the pre-trained multilingual language models, fine-tuned with the proposed customized strategy, outperform the fine-tuned pre-trained language models using an out-of-the-box fine-tuning strategy. The codebase and corpus can be accessed on our GitHub page at: https://github.com/mumairhassan/AGI-P

INDEX TERMS Business analytics, gender identification, language models, tourism industry.

I. INTRODUCTION

Author gender identification (AGI) is a task that involves determining the gender of an author based on their writing and has several real-world applications across different fields [1]. For example, in marketing and advertising, understanding the gender of authors can help tailor marketing strategies and advertisements to specific demographics [2]. It assists in creating content that resonates better with particular gender groups, leading to more effective advertising campaigns. In legal cases, determining the gender of an author based on written content can aid in forensic investigations. It might help identify potential suspects or verify the authenticity of documents [3]. For content curation and recommendation systems, platforms like social media, news aggregators, and recommendation engines can use author gender identification to personalize user content recommendations based on their preferences [4]. In natural language processing (NLP), AGI contributes to developing machine learning algorithms and models [5]. These models can aid sentiment analysis, chatbots, and other language-related AI applications.

The associate editor coordinating the review of this manuscript and approving it for publication was Agostino Forestiero.

The author's gender identification task can be considered a binary classification problem. Several features are suggested to perform the author gender identification task such as the most frequent function words, most frequent character n-grams, most frequent word n-grams, most frequent part-of-speech (POS) categories and their n-grams, sentiment lexicon, stylistic markers such as percentage of capital letters or punctuation, and mean sentence length [1], [5], [6], [7], [8], [9], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24]. To define the target category (male or female), many machine learning classifiers such as Decision Trees, Support Vector Machine, Logistic Regression, K Nearest Neighbors, and Random Forest have been recommended [5], [6], [7], [8], [9], [14], [25], [26], [27], [28], [29], [30].

The AGI task has never been investigated for the Punjabi texts. Punjabi belongs to the Indo-Aryan language family, mainly spoken by the Punjabi people in the Punjab region of two countries, Pakistan and India. As of 2017, Punjabi is the most spoken language in Pakistan, with approximately 80.5 million people. It is the 11^{th} most spoken language in India with 31.1 million speakers (as of 2011) and has official status in the Indian state of Punjab. A significant overseas diaspora speaks Punjabi in the United States, Canada, and the United Kingdom. The Punjabi language has approximately 113 million native speakers.[1]

Due to the lack of or inadequacy of several vital resources, such as gold standard datasets and fundamental natural language processing (NLP) toolkits, Punjabi can be classified as a low-resource language. However, the focus of our discussion is on the limitations of Punjabi in the context of author gender identification. The following are some major limitations.

A. LIMITATIONS OF PUNJABI IN CONTEXT OF AUTHOR GENDER IDENTIFICATION AND OUR RESEARCH OBJECTIVES

Limitation 1: Lack of Research: The author gender identification task has been extensively explored for resource-rich languages such as English [6] and Spanish [9], similarly to other NLP tasks such as part-of-speech (POS) tagging, text categorization, and named entity recognition (NER). However, this task has never been investigated for the Punjabi texts. Therefore, one of the main objectives of this paper is to fill this research gap and present the first thorough investigation of the author's gender identification task on Punjabi texts. We present an author gender identification solution that outperforms the well-known classifiers and the state-of-the-art fine-tuned pre-trained multilingual language models such as DistilBERT [31], mBERT [32], multilingual DeBERTa [33], and XLM-RoBERTa [34]. The findings of this investigation add new insights to existing knowledge (see Section III for more details).

Limitation 2: Unavailability of Reliable NLP Resources: Gender identification of authors is a crucial NLP task. However, as previously stated, this work has never been conducted on Punjabi texts, and there is no current dataset to perform this task with. To perform the author gender identification task, we require a dataset with each text sample associated with the gender label. As a result, in this paper, we built a new dataset containing 1944 samples where the length of each text sample is fixed to 250 tokens to evaluate the performance of our author gender identification solution (AGI-P), which would be made publicly available to scholars in this field (see Section III for more details).

Limitation 3: Inapplicable Features: As previously stated, a comprehensive set of features was used in resource-rich Western languages to perform the author's gender identification task. However, many of these features, such as the number of capital letters, the number of sentences that begin with capital letters, the number of sentences that begin with a lowercase alphabet, etc., cannot be extracted from texts written in Punjabi. Furthermore, due to the scarcity of credible NLP toolkits, several features are challenging to extract from Punjabi texts, such as the presence of sentiment, the frequency of POS tags, and the type of emotion, to name a few examples. In addition, the Punjabi's morphological complexity and diversity make the feature extraction more challenging. One of the main objectives of this paper is to identify the best features for the author's gender identification task for the Punjabi texts (the findings of an ablation study are given in Table 4).

Identifying the most compelling features for author gender identification in low-resource languages holds immense significance for several reasons:

1) Low-resource languages often lack extensive labeled datasets for model training. Selecting the right features is critical due to the scarcity of data, ensuring that the chosen features contribute meaningfully to gender identification.
2) Low-resource languages possess unique linguistic characteristics, different from widely studied languages.
3) Pinpointing features that reflect gender-specific linguistic nuances in these languages is essential for accurate identification.
4) Selecting optimal features directly impacts the model's performance in predicting gender accurately.
5) Identifying effective features aids in creating models adaptable to varying linguistic contexts and languages with limited resources.
6) Focusing on the most influential features maximizes the utility of limited resources available for feature extraction and model development.
7) Narrowing down feature sets minimizes computational overhead, especially in resource-constrained settings.
8) Identifying effective features paves the way for replicating successful approaches in similar low-resource language scenarios.

[1] https://en.wikipedia.org/wiki/Punjabi_language

9) Optimal feature selection contributes to scalable solutions adaptable to other under-resourced linguistic domains.

Limitation 4: Missing Application of Deep Learning: Fine-tuning pre-trained language models has achieved state-of-the-art results for various NLP tasks. However, despite compelling evidence from the literature, no study has evaluated the performance of these models to perform the author's gender identification task for Punjabi. In this investigation, we fine-tune the state-of-the-art pre-trained multilingual language models and compare their performance against our solution and well-known machine learning classifiers. We note that the proportion of the Punjabi data used to train the language model is less. To make sure that these pre-trained multilingual models are fully adapted to Punjabi, we propose a new customized fine-tuning strategy for the pre-trained multilingual language models, which improves their accuracy (see Section IV-A for more details).

B. RESEARCH QUESTIONS

In addition to addressing the aforementioned limitations, we answer the following research questions, adding new insights to the existing knowledge.

- **RQ 1:** Do well-known machine learning classifiers outperform the fine-tuned pre-trained language models for the author gender identification task on texts written in a low-resource language such as *Punjabi*?
- **RQ 2:** Do well-known machine learning classifiers outperform the fine-tuned pre-trained language models for the author gender identification task on texts written in a resource-rich language such as *English*?
- **RQ 3:** What are the most important features that discriminate the texts written by a male and a female?

C. SUMMARY OF OUR CONTRIBUTIONS

The following are the main contributions of this paper.

- We propose an author gender identification solution that can outperform well-known classifiers as well as the fine-tuned pre-trained language models such as multilingual BERT, DistilBERT, XLM-RoBERTa, and multilingual DeBERTa [35], [36].
- We propose a new fine-tuning strategy for pre-trained multilingual language models, improving their accuracy for the author gender identification task.
- As stated earlier, no current dataset exists for the author's gender identification task on Punjabi texts. Therefore, in this paper, we built the first dataset to perform this task, which will be publicly available.
- We present the first study on the author's gender identification task for the Punjabi texts, adding new insights to the existing knowledge.
- Given the limited availability of reliable NLP toolkits, we identify the best features to perform the author gender identification task for Punjabi. We conducted extensive experimental studies on datasets from two languages, including Punjabi and English, to compare our solution against well-known machine learning models and fine-tuned pre-trained language models.

The rest of the paper is organized as follows. Section II reviews the existing author gender identification studies. Our proposed solution for AGI-P is briefly described in Section III. Section IV evaluates the experimental studies and discusses their findings. The concluding remarks and future research directions are available in Section V.

II. LITERATURE REVIEW (STATE-OF-THE-ART)

Different demographic groups consistently behave differently, even in terms of language use. A substantial amount of research by sociolinguists has shown that differences in language are related to sociological factors, including age, gender, and educational attainment. Similarly, language psychologists have discovered connections between psychological characteristics and language use [37]. The field of author profiling, which has numerous applications in business and society, is currently generating a lot of interest. These techniques can be applied to business intelligence to assess demographically distinct attitudes regarding brands and businesses on social media and in targeted marketing and advertising. By tailoring the chatbot's conversational style to this personality profile, they can also be utilized in customer relations to make educated guesses about customers' personalities that interact with a company's conversational agent. In forensics, language-based profiles can be examined, and the writers of letters, emails, and other documents used in an inquiry can be identified. Author gender identification is the initial step in author profiling investigations [38], [39], [40], [41], [42]. This task has been extensively investigated; however, automatically determining stylistic differences between men and women, on the other hand, is far from ideal [37], [43], [44], [45].

Since 2010, the CLEF PAN initiatives have been investigating numerous stylometric tasks, e.g., authorship identification, plagiarism detection, author profiling, etc [46], [47], [48]. The tasks proposed correspond to various languages, with English being the most popular. The author's gender must be determined for the author profiling tasks. The best-performing approaches for the author gender identification task have used different feature types, including the most frequent n-grams of words or letters, the bag-of-words, the POS categories or their n-gram, mean sentence or word length, percentage of capital letters or punctuations, percentage of emojis, etc.

A logistic regression (LR) classifier had reported the best accuracy in CLEF-PAN 2014 [49], whereas a support vector machine (SVM) classifier had reported the best performance in 2015 [50]. The best result obtained with the gender identification corpus from the 2016 campaign was based on the LR classifier, which was trained on the most frequent words, most frequent n-grams, and stylistic cues [51]. A linear SVM classifier based on the most frequent word uni-grams and most frequent word bi-grams with the most frequent character 3-grams and the most frequent

character 5-grams reported the top performance in CLEF-PAN 2017 [52]. Similarly, an SVM classifier was used to achieve first place in 2018 [53]. Finally, the best results were achieved in 2019 using a logistic regression strategy based on word and letter n-grams [43].

It has been reported that one gender uses some topical words more frequently than the other. Men employ more terms linked to technology and finance (e.g., software, game, Linux, money, sports), whereas women choose to write about their friends and social relationships (e.g., shopping, friends, cute, love, mom) [45], [54], [55]. Women have also been reported to employ emotions or certainty phrases (such as must, always) more frequently than men [56]. It is more challenging to extract these features than regularly used determiners or pronouns for all the languages [37]. Extracting these features may require the collection of terms provided by the Linguistic Inquiry and Word Count (LIWC) programme [57].

This study is focused on the author's gender identification of the Punjabi texts. As discussed earlier, several features used in existing studies are not applicable to the Punjabi language. Furthermore, several of these properties are difficult to extract from Punjabi texts due to the scarcity of credible NLP toolkits. The presence of sentiment, the frequency of POS tags, and the type of emotion are only a few examples. These feature extractions are difficult due to the Punjabi language's complexity, morphological diversity, and the unavailability of reliable NLP toolkits for Punjab. One of the main objectives of this paper is to identify the best features for the author's gender identification task for the Punjabi texts.

A. AUTHOR GENDER IDENTIFICATION FOR LOW RESOURCE LANGUAGES

This section presents research focusing on author gender identification (AGI) tasks conducted in resource-scarce languages. Baseer et al. conducted this task on the Romanized Urdu dataset using 15 lexical features, visualizing the results through a two-dimensional graph [58]. The study revealed that male authors tend to use single characters more frequently. Conversely, females exhibit a more definitive conversational style, supported by their higher use of special characters and abbreviations. Additionally, it was noted that females employ more words from a specified Urdu corpus than males, reducing the usage of candidate words. Khandelwal et al. [59] addressed the challenge of predicting author gender in code-mixed content by introducing an English-Hindi Twitter dataset annotated with gender labels. Their study utilized machine learning methods, considering character and word-level features to infer an author's gender from the text.

Moreover, Sarwar et al. [1] recently conducted the AGI task on Urdu texts, employing 600 frequent multiword expressions and 300 frequent words as features along with a support vector machine classifier, achieving an accuracy of 93.79%.

III. PROPOSED SOLUTION (AGI-P)

This section describes our proposed solution for author gender identification for Punjabi text (AGI-P). As can be seen from Figure 1, AGI-P consists of four stages: (i) data collection, (ii) features extraction, (iii) machine learning, and (iv) author gender identification. Each process is explained in the following subsections.

A. DATA COLLECTION

To perform the author gender identification task, we require a dataset with each text sample associated with the gender label. We created a new Punjabi dataset containing news articles (texts) extracted from a newspaper[2] and annotated the dataset with gender labels using author information retrieved from the website.

We begin with a seed URL, which is the website address of a newspaper. We send HTTP requests to a seed URL to retrieve the web page content. Once a page is fetched, we parse its HTML content, extracting various elements such as author biography and news article text. We identify and extract hyperlinks in the HTML that point to other web pages. We maintain a queue or list of URLs we discover during parsing. We then follow these extracted links, navigating to new web pages. This process continues recursively, crawling through multiple levels of linked pages, discovering new links, adding them to the queue, and extracting author information and new article text. After the data collection, we removed the emoji information from the texts because we aimed to build a solution based on textual information only. We would also like to highlight that the punctuations were not removed as they may contain linguistic cues of the author's genders.

We fixed the length of each text to 250 tokens, making this task more challenging. We also tested our solution on an English dataset to achieve all the research objectives. The English dataset was extracted from Blog Corpus.[3] To make a fair comparison, we also fixed the length of each text sample in the English corpus to 250 words. A summary of the datasets is given in Table 1. As can be seen, both of the datasets are balanced in terms of the number of text samples from each gender.

B. FEATURES EXTRACTION

After collecting data, we partitioned each dataset into two sets, including training and test sets. Specifically, we used 80% of the data to train the probabilistic Light Gradient Boosted Machine (LightGBM) classifiers and 20% of the data for testing purposes. We extracted two types of features from each sample. The first type of feature is the 1800 most frequent variable length character n-grams (V.L.C), where the values of n are in the range of 2 and 10. The second type of feature is 1800 most frequent words (W). As a result, each text sample results in two feature vectors.

[2]www.punjabijagran.com
[3]https://www.kaggle.com/datasets/rtatman/blog-authorship-corpus

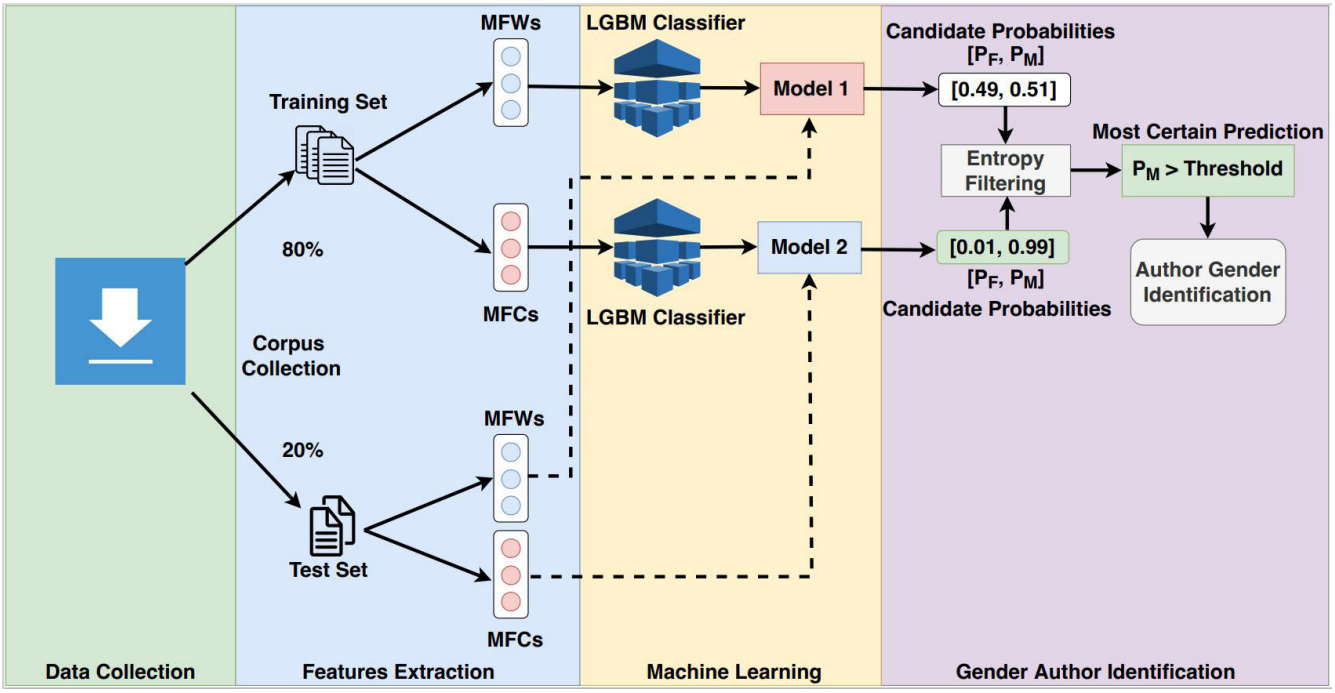


FIGURE 1. Overview of the proposed solution. News articles were gathered from a newspaper website and annotated with author gender information retrieved from the site. Feature extraction involved partitioning the dataset into training and test sets, utilizing 80% for training Light Gradient Boosted Machine (LightGBM) classifiers and 20% for testing. Feature vectors were generated, comprising frequent character n-grams and words. The best accuracy stemmed from 1800-character n-grams and words, each varying in length from 2 to 10 characters. Using probabilistic LightGBM, a classification model was trained on the feature sets, providing gender predictions for authors. LightGBM's efficiency, faster training, and higher accuracy due to its leaf-wise decision tree methodology played pivotal roles. To ensure confident predictions, entropy was employed as an uncertainty measure for identifying the most certain gender prediction per sample. A threshold value of 0.280 for Punjabi and 0.870 for English datasets was established to determine the gender of authors based on the final prediction.

TABLE 1. Summary of the datasets.

Language	Punjabi				English			
Gender	# Texts	# Words	# Characters	Text Len.	# Texts	# Words	# Characters	Text Len.
Male	972	243,000	123,302,3	250	972	243,000	132,356,5	250
Female	972	243,000	124,236,1	250	972	243,000	133,350,6	250
Total	1944	486,000	247,538,4	-	1944	486,000	265,707,1	-

The motivation behind using these features for the gender identification task is that we tried different types of word and character-based features and found that most frequent variable length character n-grams and most frequent words resulted in the best accuracy (see Table 4 for more details). We also varied the number of character and word-based features, and 1800 features resulted in the best performance (see Table 5 for more details). Moreover, we also tried different range values for n in character n-grams, and the values in range 2-10 resulted in the best accuracy (see Table 6 for more details). Once we identify the best word and character-based features, we move to the next step of our solution.

C. MACHINE LEARNING CLASSIFIER (LIGHTGBM)

After the features extraction process of our solution, we train a probabilistic LightGBM on each feature set, resulting in two author gender predictions. The motivation behind using probabilistic LightGBM is that it is a distributed, high-performance gradient-boosting framework for classification tasks. It is based on the decision tree method. It divides the tree leaf-wise with the best fit instead of other boosting algorithms that divide the tree depth- or level-wise. As a result, in LightGBM, when growing on the same leaf, the leaf-wise method can reduce more loss than the level-wise strategy, which leads to significantly superior accuracy that can only be sometimes attained by any of the existing boosting algorithms.

LightGBM has several advantages, such as higher efficiency and faster training (i.e., LightGBM uses a histogram-based approach, which accelerates training by grouping continuous feature values into discrete bins), reduced memory usage (i.e., discrete bins are used in place of continuous values, which uses less memory), better accuracy (i.e., by using a leaf-wise split strategy rather than a level-wise split approach, which is the primary element in getting higher accuracy; it produces far more complicated trees), huge dataset compatibility (i.e., compared to other tree-based

algorithms, it can handle large datasets while requiring significantly less training time), and it supports parallel learning.

D. AUTHOR GENDER IDENTIFICATION

Once we obtain two predictions for each test sample using probabilistic LightGBM, we use the entropy as the uncertainty measure to identify the most certain prediction for each test sample and use it as the final prediction for the author gender identification task. The average amount of "uncertainty" resulting from a random variable's potential outcomes is known as entropy in information theory. Given a discrete random variable X, which takes values in the alphabet $\mathcal{X}$ and is distributed according to $p : \mathcal{X} \rightarrow [0, 1]$:

$$H(X) = \sum_{c=1}^{n} P(x_i) \log P(x_i) \tag{1}$$

After we select the final prediction, we learn a threshold value to decide the gender of the text's author. The threshold value for the Punjabi dataset is 0.280, and for the English dataset, it is 0.870.

IV. PERFORMANCE EVALUATION

This section discusses the experimental setup and our extensive experimental studies for the author's gender identification task in two different languages.

A. EXPERIMENTAL SETUP

1) PARAMETER SETTINGS FOR OUR SOLUTION

We configured the LightGBM classifier primarily using its default settings, with specific adjustments made to select parameters to optimize performance for different language datasets. For the Punjabi dataset, we fine-tuned the *min_child_samples* parameter to a value of 10 and the *subsample_for_bin* to 100. In the case of the English dataset, we adjusted the *min_child_samples* to 35 and set the number of *n_estimators* to 1000. These particular settings were determined after experimenting with various values and were ultimately chosen because they yielded the highest accuracy during our testing.

2) PROPOSED FINE-TUNING STRATEGY AND PARAMETER SETTINGS FOR PRE-TRAINED MODELS

We have developed a specialized fine-tuning strategy for pre-trained multilingual language models tailored for specific linguistic contexts. This approach comprises a standard fine-tuning procedure and a proprietary, customized method.

Standard fine-tuning is implemented using the TensorFlow framework provided by Huggingface [60], with the hyper-parameters details available in Table 2. In contrast, our customized strategy begins with an initial adaptation phase (step 0), specifically for the Punjabi language. This involves using a Masked Language Model (MLM) task with the Punjabi subset of the CC-100 dataset to adjust the models to handle better Punjabi text, which is underrepresented in the CC-100 dataset [61].

TABLE 2. Parameter settings of the pre-trained multilingual language models.

Pre-trained Language Models	
Parameter	Value
No. of Epochs	5
Batch Size	8
Maximum Length	250
Optimizer	Adam
Learning Rate	$2e^{-5}$
Loss	BinaryCrossentropy

Subsequently, the gender identification datasets are divided into five segments for cross-validation (step 1). Each segment is used as a validation set in this phase while the model is fine-tuned over five epochs on the remaining data (step 2). The version of the model with the highest accuracy on the validation set is then selected for the final model ensemble.

For each input entry, the ensemble models generate a prediction probability, with the final prediction probability being the average of these five outputs (step 3). This method is applied to English without the initial priming step, as the volume of English data in pre-training is already substantial.

The ensemble approach aims to mitigate overfitting issues, ensuring robust model performance. The fine-tuned models for the custom strategy are indicated as $MODEL_C$ in Table 3.

The parameter values of fine-tuning of the pre-trained language models are given in Table 2. All models are base models, with 12 layers (with the exception of DistilBERT, which has 6 layers), a hidden size of 768, and 12 attention heads. DistilBERT has a vocabulary size of 31K tokens, mBERT has a vocabulary size of 120K tokens, and Multilingual DeBERTa and XLM-RoBERTa have a vocabulary size of 250K tokens.

3) EVALUATION MEASURES AND EVALUATION STRATEGY

We used accuracy as an evaluation measure for this task, which can be defined as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

where:

- a TP is an outcome where the classifier correctly predicts the positive class,
- a TN is an outcome where the classifier correctly predicts the negative class,
- an FP is an outcome where the classifier incorrectly predicts the positive class and
- an FN is an outcome where the model incorrectly predicts the negative class.

It is appropriate to rely on accuracy as our dataset is gender-balanced in terms of the number of text samples [62]. The train-test split ratio is fixed to 80%-20%, respectively.

TABLE 3. Accuracy Comparison of Our Solution (AGI-P), machine learning classifiers and fine-tuned pre-trained language models with different fine-tuning strategies.

Method	Punjabi	English
AGI-P (our solution)	0.9203	0.8454
Machine Learning Classifiers		
LightGBM	0.8972	0.8067
GBoost	0.8766	0.7861
RF	0.8560	0.7371
AdaBoost	0.8329	0.7835
KNN	0.8278	0.7062
DT	0.8355	0.6959
SVM	0.8406	0.7242
Fine-tuned Pre-trained Language Models (out-of-box fine-tuning strategy)		
DistilBERT	0.8020	0.8737
mBERT	0.8612	0.8652
Multilingual DeBERTa	0.8509	0.8180
XLM-RoBERTa	0.8380	0.8631
Fine-tuned Pre-trained Language Models (Customised fine-tuning strategy)		
$DistilBERT_C$	0.8046	0.8843
$mBERT_C$	0.9074	0.8869
$MultilingualDeBERTa_C$	0.8766	0.8637
$XLM-RoBERTa_C$	0.8534	0.8817

TABLE 4. Ablation Study: The performance of different feature types for the author gender identification task. For V.L.C and V.L.W, the values of n are in the range of 2-10.

Methods	Features Types									
	Character-Based Features				Word-Based Features				Combined	
	C		V.L.C		W		V.L.W		V.L.C+W	
	Punjabi	English	Punjabi	English	Punjabi	English	Punjabi	English	Punjabi	English
LightGBM	0.8792	0.7320	**0.8972**	**0.8067**	**0.8329**	**0.7990**	0.7404	0.7062	0.8895	0.6340
GBoost	0.8689	0.7474	0.8766	0.7861	0.8278	0.7603	0.7712	0.7088	0.7661	0.6598
RF	0.8612	0.7216	0.8560	0.7371	0.8252	0.7835	0.7686	0.7139	0.6067	0.5387
AdaBoost	0.8380	0.7010	0.8329	0.7835	0.7892	0.7526	0.7686	0.6753	0.7224	0.5670
KNN	0.8021	0.6881	0.8278	0.7062	0.7404	0.6546	0.5090	0.5567	0.8226	0.6959
DT	0.7943	0.6495	0.8355	0.6959	0.7584	0.6727	0.6915	0.6237	0.4036	0.5464
SVM	0.8303	0.6392	0.8406	0.7242	0.7918	0.7345	0.7918	0.6804	0.8252	0.5387

B. EXPERIMENTAL STUDIES, RESULTS DISCUSSION AND IMPLICATIONS

In this subsection, we present our extensive experimental studies, answer the research questions listed in Section I, and discuss the experimental findings.

1) ANSWER TO RQ 1 AND RQ 2

In this study, we compare our solution's (AGI-P) performance against the performance of well-known machine learning classifiers and fine-tuned pre-trained language models.

The machine learning classifiers include Light Gradient Boosted Machine Classifier (LightGBM), Gradient Boosting Classifier (GBoost), Random Forest Classifier (RF), Ada Boost Classifier (AdaBoost), K Nearest Neighbors Classifier (KNN), Decision Tree Classifier (DT) and Support Vector Machine Classifier (SVM). As for the well-known machine learning classifiers, we extract the 1800 most frequent variable length character n-grams (V.L.C) from each text sample where the values of n range between 2 and 10 and use them to train all the classifiers using their default parameter settings. The main reason to use only the V.L.C as the features for the machine learning classifiers is that they are the best features to perform the author gender identification task for both the languages (i.e., Punjabi and English, see experimental results given in Tables 5 and 6). We also compare the performance of our solution against the pre-trained language models such as DistilBERT, mBERT, XLM-RoBERTa and Multilingual DEBERTa. These models are fine-tuned using two different strategies: out-of-the-box and customized (proposed).

As can be seen from Table 3, our solution (AGI-P) outperforms the machine learning classifiers and the out-of-the-box fine-tuned pre-trained language models on Punjabi texts, which is the main focus of this paper. We also note that while using out-of-the-box fine-tuning strategy, machine learning classifiers outperform the pre-trained language models for the Punjabi texts, however, the performance of the fine-tuned pre-trained language models is higher than the machine learning classifiers for English.[4] When using customized fine-tuning, the performance of the best-fine-tuned models (an ensemble

[4]For comparison, we deliberately start fine-tuning from same pre-trained models for both Punjabi and English. Higher performance for English could be achieved using monolingual models rather than multilingual models, but this will make the comparison harder.

TABLE 5. Effect of varying the number of V.L.C features on the accuracy of the author gender identification task.

Method	Number of Features									
	100		600		1200		1800		2100	
	Punjabi	English	Punjabi	English	Punjabi	English	Punjabi	English	Punjabi	English
LightGBM	0.7455	0.6985	0.8869	0.7706	0.8869	0.8041	**0.8972**	**0.8067**	0.8843	0.8041
GBoost	0.7121	0.7320	0.8612	0.7732	0.8792	0.8093	0.8766	0.7861	0.8766	0.7887
RF	0.7275	0.6933	0.8329	0.7500	0.8509	0.7655	0.8560	0.7371	0.8509	0.7603
AdaBoost	0.6864	.6830	0.8252	0.7268	0.8406	0.7526	0.8329	0.7835	0.8123	0.7603
KNN	0.6941	0.6804	0.8303	0.7191	0.8303	0.7088	0.8278	0.7062	0.8355	0.7191
DT	0.6247	0.6314	0.7995	0.6572	0.7841	0.6856	0.8355	0.6959	0.8226	0.6856
SVM	0.7224	0.6753	0.8432	0.7216	0.8406	0.7242	0.8406	0.7242	0.8483	0.7294

TABLE 6. Effect of values of n for the V.L.C features on the accuracy of the author gender identification task.

Method	Values of the Range n of n-grams					
	2-5		2-10		2-15	
	Punjabi	English	Punjabi	English	Punjabi	English
LightGBM	0.8920	0.8144	**0.8972**	**0.8067**	0.8972	0.8067
GBoost	0.8792	0.7835	0.8766	0.7861	0.8766	0.7861
RF	0.8586	0.7577	0.8560	0.7371	0.8483	0.7500
AdaBoost	0.8329	0.7706	0.8329	0.7835	0.8329	0.7861
KNN	0.8226	0.7191	0.8278	0.7062	0.8278	0.7062
DT	0.8046	0.7036	0.8355	0.6959	0.8329	0.6727
SVM	0.8483	0.7371	0.8406	0.7242	0.8406	0.7242

of fine-tuned $mBERT_C$) outperforms the rest of the pre-trained language models.

For English, it is unsurprising that fine-tuning pre-trained language models produce higher accuracy than our proposed solution. This is in line with recent state-of-the-art [33]. These pre-trained models benefit from the information contained in the massive amount of textual data available to them in the pre-train phase and can utilize this information for the task. Furthermore, the vocabulary of these models is skewed heavily towards English as well. Nevertheless, these models are computationally expensive.

2) ANSWER TO RESEARCH QUESTION 3

To answer this question, we extracted character-based and word-based features from each text sample. Specifically, we extracted two types of character-based features including 1800 most frequent characters (C), and 1800 most frequent variable length character n-grams (V.L.C) from each text sample. Similarly, we extracted two types of word-based features from each text sample including 1800 most frequent words (W), and 1800 most frequent variable length word n-grams (V.L.W). We then apply well-known machine learning classifiers for the author's gender identification task. The experimental results are given in Table 4. As for the character-based features, the 1800 most frequent variable length characters n-grams (V.L.C) outperform the 1800 most frequent characters (C). On the other hand, 1800 most frequent words outperformed 1800 most frequent variable length word n-grams (V.L.W). We also combined the best character and word-based features and found that combined features perform poorly compared to the 1800 most frequent variable length character n-grams.

As it can be seen from Table 4 1800, the most frequent variable length character n-grams, where the values of n are between 2 and 10 range, results in the best accuracy. We also investigate the effects of varying the number of the variable length character n-grams (V.L.C) features and the values of n on the accuracy of the author gender identification task as follows.

3) EFFECT OF NUMBER OF FEATURES ON THE GENDER IDENTIFICATION TASK

In this subsection, we investigate the effect of the number of features on the accuracy of the author's gender identification task. We tried different number of features, including 100, 300, 600, 900, 1200, 1500, 1800, and 2100, while fixing the range value between 2-10. As can be seen from Table 5, the 1800 features result in the best accuracy for the author gender identification task for both languages.

4) EFFECT OF VALUES OF N ON THE PERFORMANCE OF THE GENDER IDENTIFICATION TASK

In this subsection, we investigate the effect of different n values on the accuracy of the gender identification task. Specifically, we tried different ranges of values, including 2-5, 2-10, and 2-15, using 1800 most frequent variable length character n-grams (V.L.C) as the features. As can be seen from Table 6, the values of n between the range of 2-10 resulted in the best accuracy using the LightGBM classifier. Therefore, we fixed the n values between 2 and 10 for the rest of the experimental studies.

V. CONCLUSION AND FUTURE WORKS

This paper proposes a solution (AGI-P) for the author's gender identification task for the short Punjabi texts. To test

our solution, we created a new dataset for the author gender identification task, which will be made publicly available. We conducted extensive experimental studies to show that our solution outperforms well-known machine learning classifiers and fine-tuned pre-trained language models. Specifically, despite the popularity of the fine-tuned pre-trained language models for achieving state-of-the-art performance for several NLP tasks, our solution (AGI-P) achieves the best accuracy level of 92.03% for the Punjabi dataset. We also proposed a new fine-tuning strategy for the pre-trained language models, outperforming the out-of-the-box fine-tuning strategy for low-resource language. Moreover, we found that 1800 of the most frequent variable length character n-grams are the best features to perform the author gender identification task for both Punjabi and English datasets using well-known machine learning classifiers. In the future, we plan to extend this study and identify the words and phrases that are more likely to be used by an author of a specific gender.

ACKNOWLEDGMENT

The authors thank the anonymous reviewers for their valuable comments and suggestions in improving their manuscript.

REFERENCES

[1] R. Sarwar and E. Mohamed, "Author verification of nahj al-balagha," *Digit. Scholarship Humanities*, vol. 37, no. 4, pp. 1210–1222, Oct. 2022.

[2] L. Hudders and S. De Jans, "Gender effects in influencer marketing: An experimental study on the efficacy of endorsements by same- vs. other-gender social media influencers on Instagram," *Int. J. Advertising*, vol. 41, no. 1, pp. 128–149, Jan. 2022.

[3] R. Meena, K. Krishan, A. Ghosh, and T. Kanchan, "Is it possible to estimate sex from signatures and handwriting? A review of literature, observations, and future perspectives," *Sci. Nature*, vol. 110, no. 4, p. 32, Aug. 2023.

[4] Y. Deldjoo, M. Schedl, P. Cremonesi, and G. Pasi, "Recommender systems leveraging multimedia content," *ACM Comput. Surveys*, vol. 53, no. 5, pp. 1–38, Sep. 2021.

[5] Y. HaCohen-Kerner, "Survey on profiling age and gender of text authors," *Exp. Syst. Appl.*, vol. 199, Aug. 2022, Art. no. 117140.

[6] C. Ikae and J. Savoy, "Gender identification on Twitter," *J. Assoc. Inf. Sci. Technol.*, vol. 73, no. 1, pp. 58–69, Jan. 2022.

[7] K. Alsmearat, M. Al-Ayyoub, R. Al-Shalabi, and G. Kanaan, "Author gender identification from Arabic text," *J. Inf. Secur. Appl.*, vol. 35, pp. 85–95, Aug. 2017.

[8] A. Mukherjee and B. Liu, "Improving gender classification of blog authors," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2010, pp. 207–217.

[9] M. A. Sanchez-Perez, I. Markov, H. Gómez-Adorno, and G. Sidorov, "Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same Spanish news corpus," in *Proc. Int. Conf. Cross-Language Eval. Forum Eur. Lang.* Cham, Switzerland: Springer, 2017, pp. 145–151.

[10] T. Kucukyilmaz, A. Deniz, and H. E. Kiziloz, "Boosting gender identification using author preference," *Pattern Recognit. Lett.*, vol. 140, pp. 245–251, Dec. 2020.

[11] S. Baxevanakis, S. Gavras, D. Mouratidis, and K. L. Kermanidis, "A machine learning approach for gender identification of Greek tweet authors," in *Proc. PETRA*, Corfu, Greece, F. Makedon, Ed. Jun. 2020, pp. 1–4.

[12] A. I. Al-Ghadir and A. M. Azmi, "A study of Arabic social media users—Posting behavior and author's gender prediction," *Cognit. Comput.*, vol. 11, no. 1, pp. 71–86, Feb. 2019.

[13] V. Simaki, C. Aravantinou, I. Mporas, M. Kondyli, and V. Megalooikonomou, "Sociolinguistic features for author gender identification: From qualitative evidence to quantitative analysis," *J. Quant. Linguistics*, vol. 24, no. 1, pp. 65–84, Jan. 2017.

[14] F. Safara, A. S. Mohammed, M. Yousif Potrus, S. Ali, Q. T. Tho, A. Souri, F. Janenia, and M. Hosseinzadeh, "An author gender detection method using whale optimization algorithm and artificial neural network," *IEEE Access*, vol. 8, pp. 48428–48437, 2020.

[15] K. Silva, B. Can, F. Blain, R. Sarwar, L. Ugolini, and R. Mitkov, "Authorship attribution of late 19th century novels using GAN-BERT," in *Proc. 61st Annu. Meeting Assoc. Comput. Linguistics*, 2023, pp. 310–320.

[16] H. Saadany, E. Mohamed, and R. Sarwar, "Towards a better understanding of tarajem: Creating topological networks for Arabic biographical dictionaries," *J. Data Mining Digit. Humanities*, vol. 2023, pp. 1–25, Jun. 2023.

[17] E. Mohamed and R. Sarwar, "Linguistic features evaluation for Hadith authenticity through automatic machine learning," *Digit. Scholarship Humanities*, vol. 37, no. 3, pp. 830–843, Aug. 2022.

[18] N. Trijakwanich, P. Limkonchotiwat, R. Sarwar, W. Phatthiyaphaibun, E. Chuangsuwanich, and S. Nutanong, "Robust fragment-based framework for cross-lingual sentence retrieval," in *Proc. Assoc. Comput. Linguistics (EMNLP)*, Nov. 2021, pp. 935–944.

[19] I. Safder, Z. Mahmood, R. Sarwar, S. Hassan, F. Zaman, R. M. A. Nawab, F. Bukhari, R. A. Abbasi, S. Alelyani, N. R. Aljohani, and R. Nawaz, "Sentiment analysis for Urdu online reviews using deep learning models," *Exp. Syst.*, vol. 38, no. 8, p. e12751, Dec. 2021.

[20] I. Safder, H. Batool, R. Sarwar, F. Zaman, N. R. Aljohani, R. Nawaz, M. Gaber, and S.-U. Hassan, "Parsing AUC result-figures in machine learning specific scholarly documents for semantically-enriched summarization," *Appl. Artif. Intell.*, vol. 36, no. 1, Dec. 2022, Art. no. 2004347.

[21] P. Limkonchotiwat, W. Phatthiyaphaibun, R. Sarwar, E. Chuangsuwanich, and S. Nutanong, "Handling cross and out-of-domain samples in Thai word segmentation," in *Proc. Assoc. Comput. Linguistics (ACL-IJCNLP)*, Aug. 2021, pp. 1003–1016.

[22] P. Limkonchotiwat, W. Phatthiyaphaibun, R. Sarwar, E. Chuangsuwanich, and S. Nutanong, "Domain adaptation of Thai word segmentation models using stacked ensemble," in *Proc. 2020 Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2020, pp. 3841–3847.

[23] S.-U. Hassan, N. R. Aljohani, M. Shabbir, U. Ali, S. Iqbal, R. Sarwar, E. Martínez-Cámara, S. Ventura, and F. Herrera, "Tweet coupling: A social media methodology for clustering scientific publications," *Scientometrics*, vol. 124, pp. 973–991, Aug. 2020.

[24] R. Sarwar, N. Urailertprasert, N. Vannaboot, C. Yu, T. Rakthanmanon, E. Chuangsuwanich, and S. Nutanong, "*CAG*: Stylometric authorship attribution of multi-author documents using a co-authorship graph," *IEEE Access*, vol. 8, pp. 18374–18393, 2020.

[25] B. Bassem and M. Zrigui, "Gender identification: A comparative study of deep learning architectures," in *Proc. Int. Conf. Intell. Syst. Design Appl.* Cham, Switzerland: Springer, 2018, pp. 792–800.

[26] M. K. Afzal, M. Shardlow, S. Tuarob, F. Zaman, R. Sarwar, M. Ali, N. R. Aljohani, M. D. Lytras, R. Nawaz, and S.-U. Hassan, "Generative image captioning in Urdu using deep learning," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 6, pp. 7719–7731, Jun. 2023.

[27] S. Mohammad, M. U. S. Khan, M. Ali, L. Liu, M. Shardlow, and R. Nawaz, "Bot detection using a single post on social media," in *Proc. 3rd World Conf. Smart Trends Syst. Secur. Sustainablity (WorldS)*, Jul. 2019, pp. 215–220.

[28] M. U. Hassan, S. Alaliyat, R. Sarwar, R. Nawaz, and I. A. Hameed, "Leveraging deep learning and big data to enhance computing curriculum for industry-relevant skills: A Norwegian case study," *Heliyon*, vol. 9, no. 4, Apr. 2023, Art. no. e15407.

[29] S.-U. Hassan, N. R. Aljohani, U. I. Tarar, I. Safder, R. Sarwar, S. Alelyani, and R. Nawaz, "Exploiting tweet sentiments in altmetrics large-scale data," *J. Inf. Sci.*, vol. 49, no. 5, pp. 1229–1245, Oct. 2023.

[30] E. Mohamed, R. Sarwar, and S. Mostafa, "Translator attribution for Arabic using machine learning," *Digit. Scholarship Humanities*, vol. 38, no. 2, pp. 658–666, May 2023.

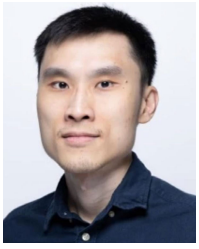
[31] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[32] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. Conf. North American Chapter Association Computational Linguistics, Human Language Technologies*. Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. [Online]. Available: https://aclanthology.org/N19-1423

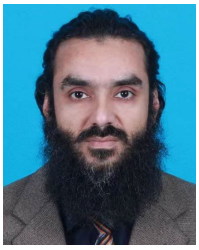
[33] P. He, J. Gao, and W. Chen, "DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing," 2021, *arXiv:2111.09543*.
[34] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8440–8451.
[35] S. Bano and S. Khalid, "BERT-based extractive text summarization of scholarly articles: A novel architecture," in *Proc. Int. Conf. Artif. Intell. Things (ICAIoT)*, Dec. 2022, pp. 1–5.
[36] S. Bano, S. Khalid, N. M. Tairan, H. Shah, and H. A. Khattak, "Summarization of scholarly articles using BERT and BiGRU: Deep learning-based extractive approach," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 35, no. 9, Oct. 2023, Art. no. 101739.
[37] J. W. Pennebaker, "The secret life of pronouns," *New Scientist*, vol. 211, no. 2828, pp. 42–45, Sep. 2011.
[38] R. Sarwar and S.-U. Hassan, "UrduAI: Writeprints for Urdu authorship identification," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 21, no. 2, pp. 1–18, Mar. 2022.
[39] R. Sarwar, A. T. Rutherford, S.-U. Hassan, T. Rakthanmanon, and S. Nutanong, "Native language identification of fluent and advanced non-native writers," *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 19, no. 4, pp. 1–19, Jul. 2020.
[40] A. Vashistha, A. Garg, R. Anderson, and A. A. Raza, "Threats, abuses, flirting, and blackmail: Gender inequity in social media voice forums," in *Proc. CHI Conf. Human Factors Comput. Syst.*, May 2019, pp. 1–13.
[41] R. López-Santillán, M. Montes-Y-Gómez, L. C. González-Gurrola, G. Ramírez-Alonso, and O. Prieto-Ordaz, "Richer document embeddings for author profiling tasks based on a heuristic search," *Inf. Process. Manag.*, vol. 57, no. 4, Jul. 2020, Art. no. 102227.
[42] M. Fatima, K. Hasan, S. Anwar, and R. M. A. Nawab, "Multilingual author profiling on Facebook," *Inf. Process. Manag.*, vol. 53, no. 4, pp. 886–904, Jul. 2017.
[43] F. Rangel and P. Rosso, "Overview of the 7th author profiling task at PAN 2019: Bots and gender profiling in Twitter," in *Proc. CEUR Workshop*, Lugano, Switzerland, 2019, pp. 1–36.
[44] J. Nerbonne, "The secret life of pronouns. What our words say about us," *Literary Linguistic Comput.*, vol. 29, no. 1, pp. 139–142, Apr. 2014.
[45] H. A. Schwartz, J. C. Eichstaedt, M. L. Kern, L. Dziurzynski, S. M. Ramones, M. Agrawal, A. Shah, M. Kosinski, D. Stillwell, M. E. P. Seligman, and L. H. Ungar, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS ONE*, vol. 8, no. 9, Sep. 2013, Art. no. e73791.
[46] J. Savoy, *Machine Learning Methods for Stylometry*. Berlin, Germany: Springer, 2020.
[47] P. Rosso, M. Potthast, B. Stein, E. Stamatatos, F. Rangel, and W. Daelemans, "Evolution of the PAN lab on digital text forensics," in *Information Retrieval Evaluation in a Changing World*. Berlin, Germany: Springer, 2019, pp. 461–485.
[48] S. Ashraf, O. Javed, M. Adeel, H. Iqbal, and R. M. A. Nawab, "Bots and gender prediction using language independent stylometry-based approach," in *Proc. CLEF*, 2019, pp. 1–11.
[49] F. Rangel, P. Rosso, I. Chugur, M. Potthast, M. Trenkmann, B. Stein, B. Verhoeven, and W. Daelemans, "Overview of the 2nd author profiling task at PAN 2014," in *Proc. CLEF Eval. Labs Workshop Work. Notes Papers*, Sheffield, U.K., 2014, pp. 1–30.
[50] P. Rosso, M. Potthast, B. Stein, and W. Daelemans, "Overview of the 3rd author profiling task at PAN 2015," in *Proc. CLEF*, 2015, pp. 461–485.
[51] F. Rangel, P. Rosso, B. Verhoeven, W. Daelemans, M. Potthast, and B. Stein, "Overview of the 4th author profiling task at PAN 2016: Cross-genre evaluations," in *Proc. CLEF*, 2016, pp. 750–784.
[52] M. Potthast, F. Rangel, M. Tschuggnall, E. Stamatatos, P. Rosso, and B. Stein, "Overview of PAN'17," in *Proc. Int. Conf. Cross-Language Eval. Forum Eur. Lang.* Cham, Switzerland: Springer, 2017, pp. 275–290.
[53] F. Rangel, P. Rosso, M. M.-Y. Gómez, M. Potthast, and B. Stein, "Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in Twitter," in *Proc. CLEF*, 2018, pp. 1–38.
[54] S. Argamon, M. Koppel, J. W. Pennebaker, and J. Schler, "Automatically profiling the author of an anonymous text," *Commun. ACM*, vol. 52, no. 2, pp. 119–123, Feb. 2009.
[55] J. Schler, M. Koppel, S. Argamon, and J. W. Pennebaker, "Effects of age and gender on blogging," in *Proc. AAAI Spring Symp., Comput. Approaches Analyzing Weblogs*, vol. 6, 2006, pp. 199–205.
[56] L. Young and S. Soroka, "Affective news: The automated coding of sentiment in political texts," *Political Commun.*, vol. 29, no. 2, pp. 205–231, Apr. 2012.
[57] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: LIWC and computerized text analysis methods," *J. Lang. Social Psychol.*, vol. 29, no. 1, pp. 24–54, Mar. 2010.
[58] F. Baseer, J. Jaafar, and A. Habib, "Gender and age identification through romanized Urdu dataset," in *Proc. 1st Int. Conf. Artif. Intell. Data Sci. (AiDAS)*, 2019, pp. 164–169.
[59] A. Khandelwal, S. Swami, S. S. Akhtar, and M. Shrivastava, "Gender prediction in english-hindi code-mixed social media content: Corpus and baseline system," *Computación y Sistemas*, vol. 22, no. 4, pp. 1241–1247, Dec. 2018.
[60] *Fine-tune a Pretrained Model*. Accessed: Dec. 7, 2023. [Online]. Available: https://huggingface.co/docs/transformers/training
[61] *Cc-100: Monolingual Datasets From Web Crawl Data*. Accessed: Dec. 7, 2023. [Online]. Available: https://data.statmt.org/cc-100/
[62] S. Khalid, S. Wu, and F. Zhang, "A multi-objective approach to determining the usefulness of papers in academic search," *Data Technol. Appl.*, vol. 55, no. 5, pp. 734–748, Oct. 2021.

RAHEEM SARWAR received the Ph.D. degree from the City University of Hong Kong. His research interests include technology development, artificial intelligence, NLP, data science, scientometrics, altmetrics, information retrieval, and text mining.

LE AN HA is working and publishing on a variety of subjects, including automatic terminology extraction, both monolingual and multilingual, multiple-choice question (MCQ) generation, analysis of multiple-choice test items, and multilingual preprocessing. He has extensive experience in developing commercial natural language processing (NLP) vertical solutions. He has acted as an Acting Coordinator of an EU Leonardo Project (TELLME), which has developed a range of products, including work-related language exercises and showcased NLP technologies, such as automatic term extraction and MCQ generation. He has developed a Computer-Aided Patient Notes Scoring System for a well-known U.S. medical examiner organization. Also, he has been leading research activities in the domain of applying NLP technologies for licensing testing, funded by an U.S. organization on a yearly rolling research contract, including American–British transliteration, information extraction, item difficulty prediction, item response time prediction, and item distractor prediction.



PIN SHEN TEH has been teaching for more than a decade, mainly with higher education institutions. He has experience in teaching ICT and coding to students aged 6–16. His teaching focuses on programming and database subjects. He is the ManMet Minecraft Project Pioneer and a Minecraft Certified Trainer. His research interests include practical machine learning applications, biometrics systems, and metaverse.



FAHAD SABAH received the M.S. degree in computer science from Information Technology University, Lahore, Pakistan. He is currently pursuing the Ph.D. degree with the Beijing University of Technology, China.

RAHEEL NAWAZ is a Pro VC of digital transformation with Staffordshire University. He is also a leading researcher in artificial intelligence and digital education. He holds several adjunct professorships and scientific directorships across Asia and North America. He has authored over 150 peer-reviewed research articles and his career grant capture stands at over £14 million. He has graduated 19 Ph.D. students. According to Google Scholar, he is among the top-10 most cited scholars in the world in the fields of digital transformations, applied artificial intelligence, and educational data science. He sits on the boards of research and charitable organizations, such as the National Centre for Artificial Intelligence, Pakistan; TechSkills, U.K.; and NTF, U.K. He has advised national policy organizations, including the Prime Minister's Task Force on Science and Technology, Pakistan.

IBRAHIM A. HAMEED (Senior Member, IEEE) received the first Ph.D. degree in AI from Korea University, South Korea, and the second Ph.D. degree in field robotics from Aarhus University, Denmark. He is a Professor and the Deputy Head of research and innovation with NTNU. He is the Elected Chair of the IEEE Computational Intelligence Society (CIS), Norway Section; and a Founder and the Head of the Social Robots Laboratory, Ålesund. His current research interests include artificial intelligence, machine learning, optimization, and robotics.

MUHAMMAD UMAIR HASSAN received the bachelor's degree in computer science from the University of Punjab, Pakistan, and the master's degree in computer science from the University of Jinan, China. He is currently pursuing the Ph.D. degree in computer science with the Norwegian University of Science and Technology (NTNU), Norway. He is also a University Lecturer with NTNU. He was a Research Assistant with the Shandong Provincial Key Laboratory of Network-Based Intelligent Computing, University of Jinan. His research interests include digital twins, machine learning, computer vision, deep learning, and image retrieval. Previously, he has worked in computer networking, cloud computing, and natural language processing.