

Author gender identification for Urdu articles

Item Type	Conference contribution
Authors	Sarwar, Raheem
Citation	Sarwar, R. (2022) Author gender identification for Urdu articles. Lecture Notes in Computer Science, 13528, pp. 221–235.
DOI	10.1007/978-3-031-15925-1_16
Publisher	Springer
Journal	Lecture Notes in Computer Science
Download date	2026-04-18 05:16:45
License	https://creativecommons.org/licenses/by-nc-nd/4.0/
Link to Item	http://hdl.handle.net/2436/624842

Author Gender Identification For Urdu

Raheem Sarwar¹[0000-0002-0640-807X]

Research Group in Computational Linguistics, RILP, University of Wolverhampton,
United Kingdom
R.Sarwar4@wlv.ac.uk

Abstract. In recent years, author gender identification has gained considerable attention in the fields of computational linguistics and artificial intelligence. This task has been extensively investigated for resource-rich languages such as English and Spanish. However, researchers have not paid enough attention to perform this task for Urdu articles. Firstly, I created a new Urdu corpus to perform the author gender identification task. I then extracted two types of features from each article including the most frequent 600 multi-word expressions and the most frequent 300 words. After I completed the corpus creation and features extraction processes, I performed the features concatenation process. As a result each article was represented in a 900D feature space. Finally, I applied 10 different well-known classifiers to these features to perform the author gender identification task and compared their performances against state-of-the-art pre-trained multilingual language models, such as mBERT, DistilBERT, XLM-RoBERTa and multilingual DeBERTa, as well as Convolutional Neural Networks (CNN). I conducted extensive experimental studies which show that (i) using the most frequent 600 multi-word expressions as features and concatenating them with the most frequent 300 words as features improves the accuracy of the author gender identification task, and (ii) support vector machines outperforms other classifiers, as well as fine-tuned pre-trained language models and CNN. The code base and the corpus can be found at: https://github.com/raheem23/Gender_Identification_Urdu.

Keywords: Multiword Expressions · Author Profiling · Author Gender Identification.

1 Background and Introduction

The Internet's rapid growth has given rise to a plethora of new ways to transmit information across time and geography. Online social networking platforms (such as Twitter, Myspace, and Facebook), e-commerce websites (such as eBay, Craigslist), and usenet newsgroups are all gaining popularity. However, this growth enables a variety of Internet abuses. Online communities are prone to deception, incorrect information, and other threats. The Internet Crime Complaint Center (IC3) has received an average of 552,000 complaints every year over the last five years. These complaints include a wide range of Internet frauds that affect

people all over the world. To date, IC3 has received 2,760,044 complaints, with a total loss of \$18.7 billion reported. Homeland security and law enforcement organisations have initiated programmes to avoid fraudulent attacks and trace the identity of offenders in order to safeguard against terrorism, predators, and other threats [6].

Anonymity is an important feature of online communities. In cyberspace, people may not need to reveal their genuine identities, such as their name, age, gender, or residence. In many cases of Internet crime, the criminals seek to conceal their real identities. As a result, developing an efficient and effective method for identity tracking in Internet forensics becomes critical. In this paper, I am primarily concerned in the *author gender identification* task, which can be defined as follows: given an article (a text), identify whether the article is written by male or female. The author gender identification task has been extensively investigated for resource-rich Western languages such as English [14], and Spanish [21]. However, researchers have not paid enough attention to perform this task for Urdu articles. After a thorough search of Urdu literature I found two relevant studies on texts written in the Roman Urdu (i.e., Urdu texts written in the Latin alphabet) [3,10]. I note that this paper focuses on Urdu articles written in Urdu alphabet. The author gender identification task can be considered as a binary classification problem [11,19]. To determine the gender of the articles authors', several types of features are suggested 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 sequences, as well as other stylistic markers such as percentage of capital letters or punctuations, mean sentence length, etc. [14,2,18,21,16,5,1,24]. Different machine learning classifiers such as Decision Trees, Logistic Regression, K Nearest Neighbors, Support Vector Machine, and Random Forest have been suggested to determine the target category [14,2,18,21,4]. This investigation explores the effectiveness of multi-word expressions (MWEs), word-based features and character-based features to perform the author gender identification task for the Urdu articles using machine learning classifiers. I compared the performance of the machine learning classifiers against the state-of-the-art fine-tuned pre-trained language models such as DistilBERT [22], mBERT [9], multilingual DeBERTa [12], and XLM-RoBERTa [7] as well as Convolutional Neural Networks (CNN).

Urdu is an Indo-Aryan language that borrowed a large percentage of its vocabulary from other languages such as Arabic and Persian. The Ethnologue, a well-known reference source that publishes statistics on living languages, has ranked Urdu as the 11th most spoken language in the world in 2020. It is also widely acknowledged as a major South Asian language, with 490 million native speakers worldwide [17]. It is the official language of five Indian states, including Bihar, Uttar Pradesh, and Jharkhand. It is the national language of Pakistan, which has a population of about 220 million people. According to the 2011 census of linguistic statistics conducted by the Indian government, India had 50,772,631 Urdu speakers. Urdu speakers can also be found in the United Kingdom, the

United States, Canada, Australia, the Middle East, and Europe. Urdu is often regarded as a low-resource language due to the lack of or inadequacy of various critical resources, such as gold standard datasets and fundamental natural language processing (NLP) toolkits, such as reliable tokenizers and stemmers [8]. My discussion, however, is focused on the limitations of Urdu in the context of the author gender identification task. Some key limitations are as follows.

- **Lack of attention.** Similar to other NLP tasks such as part-of-speech (POS) tagging, text categorisation, named entity recognition (NER), the author gender identification task has also been extensively investigated for resource-rich languages such as English [14], and Spanish [21]. However, this is the first study on the author gender identification task for the Urdu articles.
- **Unavailability of resources.** Author gender identification is an important NLP task. However, as mentioned earlier, this task has never been performed on the Urdu articles and there is no existing corpus available to perform this task. Therefore in this paper I introduced a new corpus to perform this task on the Urdu articles which can be accessed at: https://github.com/raheem23/Gender_Identification_Urdu.
- **Inapplicable features.** As mentioned earlier, a comprehensive set of features have been used to perform the author gender identification task on resource-rich Western languages. Many of these features, however, are not applicable to the Urdu language. The number of capital alphabets, the number of sentences that begin with capital alphabets, and the number of sentences that begin with a lowercase alphabet are among these features. Moreover, some of these features are difficult to extract from articles written in Urdu due to limited availability of the reliable NLP toolkits. The frequency of POS tags, the presence of sentiment, and the type of emotion are a few to mention. The intricacy and morphological richness of the Urdu language account for the hard nature of these aspects.
- **Missing application of deep learning.** Fine tuning pre-trained language models have achieve state-of-the-art results for various NLP tasks for resource-rich languages. However, despite the existence of compelling evidence from the literature, no study has evaluated the performance of these models to perform the author gender identification task for low-resource languages such as Urdu.

The main purpose of this paper was to introduce a new corpus that can be used to investigate the author gender identification task for Urdu articles and formulate a solution that outperforms the state-of-the-art methods. In addition to this, I aimed at answering the following research questions.

- **RQ 1:** What are the best features to perform the author gender identification task for Urdu articles?
- **RQ 2:** Can the performance of the author gender identification task be improved by concatenating the most frequent MWEs with the most frequent words?

- **RQ 3:** Can classical machine learning classifiers outperform state-of-the-art pre-trained language models to perform the author gender identification task for Urdu articles?

Summary of Contributions. My main contributions are as follows.

- I created the first corpus to perform the author gender identification task on Urdu articles.
- I explored, for the first time, the effectiveness of MWEs for the author gender identification task in Urdu.
- I evaluated the effectiveness of the state-of-the-art pre-trained language models for the gender identification task and compared its performance against the machine learning classifiers and CNN. The experimental findings adds new insights to existing knowledge.

The rest of the paper is organised as follows. Section 2 presents methodology. Section 3 describes the experimental studies and the findings. Section 4 contains the concluding remarks and future research directions.

2 Methodology

Overview. As can be seen in Figure 1, my methodology consists of three main processes. Firstly I created a data scraper to retrieve the articles from a newspaper website. These articles are published by male and female columnists. I then extracted features from each article and applied different well-known machine learning classifiers to these features to perform the author gender identification task. Each process is described in the following subsections.

2.1 Data Collection

I created a data scraper using the BeautifulSoup¹ and Newspaper² libraries in Python. These libraries are highly effective for extracting and curating articles from newspaper websites. I then used this scraper to collect the articles from the Dunya News Website. These articles were written by both male and female columnists. The summary of the dataset is given in Table 1. The dataset consists of 844 articles and there are equal number of articles from male and female authors. Moreover the size of each article is fixed to only 250 tokens, which makes this task more challenging to perform.

2.2 Features Extraction and Concatenation

Before I explain the features extraction process, I briefly define the multi-word expressions as follows.

¹ <https://beautiful-soup-4.readthedocs.io/en/latest/>

² <https://newspaper.readthedocs.io/en/latest/>

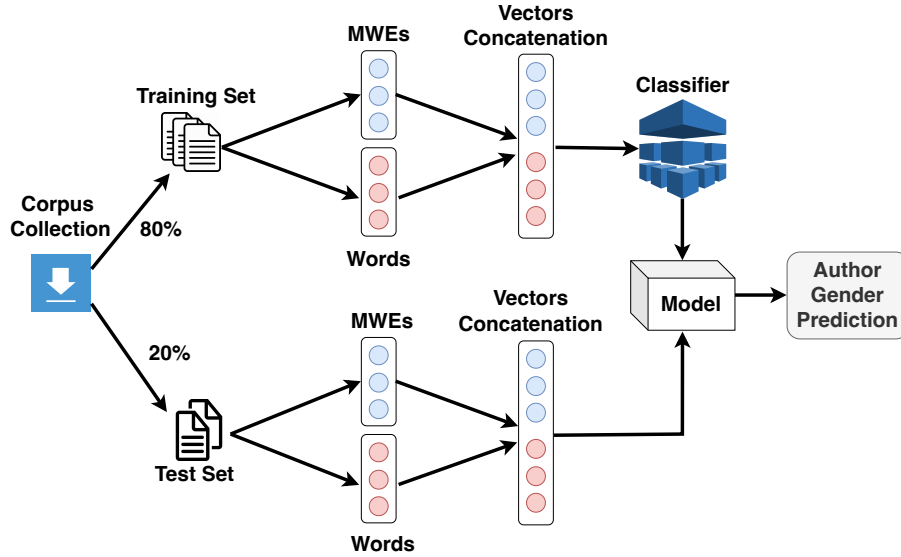


Fig. 1: Author gender identification system overview

Multi-word expressions (MWEs) (sometimes also known as word bigrams) are lexical entities made up of a number of orthographic words. Multi-word expressions make up a large component of any natural language’s lexicon. They are a varied group of structures with a wide range of properties, all of which are differentiated by their distinctive behaviour. In terms of morphology, some multi-word expressions enable some constituents to freely inflect while restricting the inflection of others. Multi-word expressions may allow constituents to go through non-standard morphological inflections that they would not go through in isolation in some instances. Some multi-word expressions behave like words, while others behave like phrases; some appear in a set pattern and order, while others allow for numerous syntactic changes. The semantic opacity of multi-word expressions is their most distinguishing feature, yet compositionality in multi-word expressions is incremental, ranging from entirely compositional to wholly idiomatic. [20].

Following the data collection process described in the above subsection, I extracted two types of features from each article including the most frequent 600 multi-word expressions and the most frequent 300 words. Firstly, I calculated the frequencies of all the multi-word expressions and words in the corpus. To extract multi-word expression from Urdu articles, I iterated through the corpus and identified common words occurring together using the following equation.

$$\frac{\text{count}(AB) - \text{count}_{\min}}{\text{count}(A) * \text{count}(B)} * N > \text{threshold} \quad (1)$$

Gender	Number of Articles	Number of Words	Number of Characters	Text Length
Male	422	110,500	518,628	250
Female	422	110,500	515,810	250
Total	844	221,000	103,443,8	250

Table 1: Summary of the dataset for the author gender identification task for Urdu

where:

- $\text{count}(A)$ is the number of times token A appeared in the corpus;
- $\text{count}(B)$ is the number of times token B appeared in the corpus;
- $\text{count}(A B)$ is the number of times the tokens A and B appeared in the corpus in order;
- N is the total size of the corpus vocabulary
- count_{min} is a user-defined parameter to ensure that accepted phrases occur a minimum number of times (5 in this paper)
- threshold is a user-defined parameter to control how strong is a relationship between two tokens (10 in this paper)

After I extracted all the multi-word expressions, I selected the most frequent 600 multi-word expressions from the corpus and used them as features for the author identification task. I also extracted the most frequent 300 words from the dataset and used them as features for the author gender identification task. After the features extraction process, each article was represented with two features vectors. I proposed to concatenate the 600 most frequent multi-word expressions and the 300 most frequent words vectors to obtain better performance in the author gender identification task, as shown in Figure 1.

2.3 Machine Learning Classifiers and Deep Learning Models.

After the features extraction and concatenation processes, I applied 10 well-known classifiers on these feature vectors to perform the author gender identification task with default parameter settings. These classifiers include Light Gradient Boosted Machine Classifier [15], Cat Boosted Classifier³, Extreme Gradient Boosted Classifier, Gradient Boosting Classifier, Random Forest Classifier, Extra Trees Classifier, Ada Boost Classifier, K Nearest Neighbors Classifier [23], Decision Tree Classifier, and Support Vector Machine Classifier [13]. I used Scikit-Learn⁴ library in Python for the implementation of these machine learning classifiers.

³ <https://catboost.ai/en/docs/>

⁴ <https://scikit-learn.org>

I also fine-tuned the state-of-the-art pre-trained language models such as DistilBERT [22], mBERT [9], multilingual DeBERTa [12], and XLM-RoBERTa [7] to perform the author gender identification task for Urdu articles and compared their performance against the machine learning classifiers and Convolutional Neural Networks (CNN). The fine-tuning processes can be defined as training a pre-trained language model such as multilingual BERT on the author gender identification task corpus. The fine-tuning was performed using Hugging library with Pytorch deep learning framework. The parameters values used to fine-tune the multilingual 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 768, and 12 attention heads. A summary of the CNN model is given in Table 3.

2.4 Evaluation Strategy and Evaluation Measure

I used 80% of the articles for training machine learning classifiers, CNN and fine-tuning pre-trained language models, and the rest of the articles were used for testing them. I used accuracy as an evaluation measure for this task which can be defined as follows.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

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.

Pre-trained Language Models	
Epochs	5
Batch Size	8
Maximum Length	250
Optimiser	Adam
Learning Rate	2^{-5}
Validation Split	0.2
Loss	Binary Crossentropy
load_best_model_at_end	True
seed	42

Table 2: Parameter settings for the fine-tuned pre-trained language models

Layer (type)	Output Shape	Param #
embedding_6 (Embedding)	(None, 250, 900)	16206300
conv1d_18 (Conv1D)	(None, 250, 128)	460928
max_pooling1d_18 (MaxPooling1D)	(None, 125, 128)	0
conv1d_19 (Conv1D)	(None, 125, 64)	32832
max_pooling1d_19 (MaxPooling1D)	(None, 62, 64)	0
conv1d_20 (Conv1D)	(None, 62, 32)	8224
max_pooling1d_20 (MaxPooling1D)	(None, 31, 32)	0
flatten_6 (Flatten)	(None, 992)	0
dense_12 (Dense)	(None, 256)	254208
dense_13 (Dense)	(None, 1)	257

Table 3: Summary of the CNN model

3 Experimental Studies

To achieve the objectives of this investigation and answer the research questions listed in Section 1, I conducted two main experimental studies. In the first study I evaluated the effectiveness of using multi-word expressions as features for the author gender identification task. I also evaluated the effectiveness of concatenating the 600 most frequent multi-word expressions (MWEs) with the 300 most frequent words and used them as features for the author gender identification task. In the second experimental study I compared the performance of the machine learning classifiers against the fine-tuned pre-trained language models and CNN.

In addition, I conducted five studies to investigate the effect of varying the number of: (i) the most frequent MWEs, (ii) the most frequent words, (iii) the most frequent variable length word n-grams, (iv) the most frequent characters, and (v) the most frequent variable length character n-grams, on the performance of the author gender identification task for Urdu.

3.1 Effect of most frequent MWEs, most frequent Words and their Concatenation on the Author Gender Identification Task

In this study I evaluated the effectiveness of: (i) the 600 most frequent multi-word expressions (MWEs) as features for the author gender identification task, (ii) the 300 most frequent words as features for the author gender identification task, and (iii) concatenating MWEs and words into one feature vector for the author gender identification task. As can be seen from Table 4, concatenating the 600 most frequent MWEs and the 300 most frequent words together reports the highest accuracy for the author gender identification task. It also implies that the information contained in these feature vectors is complementary and orthogonal. I also note that I have tried all combinations of the features vectors concatenation and “MWEs+Words” resulted in the best accuracy. In interest of brevity, the experimental results for the other features vectors concatenations are not given in Table 4

Table 4: Effect of the feature vector concatenation process where the 600 most frequent multi-word expressions (MWEs) and the 300 most frequent words are used as the features for the author gender identification task

Classifier	MWEs+Words	MWEs	Words
Light Gradient Boosted Machine Classifier	0.8814	0.7966	0.8531
Cat Boosted Classifier	0.8644	0.8701	0.8644
Extreme Gradient Boosted Classifier	0.8701	0.7910	0.8644
Gradient Boosting Classifier	0.8305	0.8475	0.8870
Random Forest Classifier	0.8983	0.8531	0.8531
Extra Trees Classifier	0.8418	0.8362	0.8701
Ada Boost Classifier	0.8249	0.7514	0.7853
K Nearest Neighbors Classifier	0.7571	0.6045	0.7119
Decision Tree Classifier	0.7288	0.7627	0.6949
Support Vector Machine Classifier	0.9379	0.8305	0.9209

3.2 Performance Comparison Among Machine Learning Classifiers, CNN and Fine-Tuned Pre-Trained Language Models

As mentioned earlier, fine-tuning the pre-trained language models such as BERT has reported state-of-the-art results for many natural language processing tasks for resource-rich languages. However, the performance of fine-tuned pre-trained language models has never been evaluated on the author gender identification task for the Urdu articles. Therefore, it would be interesting to compare the performance of the fine-tuned pre-trained multilingual language models against the performance of the machine learning classifiers for the author gender identification task for Urdu articles. The experimental results are given in Table 5. The experimental results for the machine learning classifiers are obtained using the most frequent the 600 multi-word expressions (MWEs) and the 300 most frequent words as features using default parameters values. As can be seen from Table 5, the support vector machine classifier outperformed the rest of the machine learning classifiers, CNN and fine-tuned pre-trained language models.

3.3 Effect of varying the number of MWEs as Features

As mentioned earlier, one of the main objectives of this investigation was to explore the effectiveness of using multi-word expressions (MWEs) as the features to perform the author gender identification task. Therefore, it would be interesting to investigate the effect of varying the number MWEs on the accuracy of the author gender identification task. Specifically, I varied the number of most frequent MWEs from 100 to 879⁵ and evaluated their performance using machine learning classifiers. As can be seen from Table 6 that using 600 MWEs as features reported the highest accuracy level and that they are effective to be used as features for the author gender identification task.

⁵ 879 are the total number of MWEs in the dataset

Table 5: Performance comparison among the machine learning classifiers, fine-tuned pre-trained language models and CNN: The experimental results for the machine learning classifiers are obtained using the most frequent the 600 multi-word expressions (MWEs) and the 300 most frequent words as features using default parameters.

Machine Learning Classifiers and Deep Learning Models	Accuracy
Light Gradient Boosted Machine Classifier	0.8814
Cat Boosted Classifier	0.8644
Extreme Gradient Boosted Classifier	0.8701
Gradient Boosting Classifier	0.8305
Random Forest Classifier	0.8983
Extra Trees Classifier	0.8418
Ada Boost Classifier	0.8249
K Nearest Neighbors Classifier	0.7571
Decision Tree Classifier	0.7288
Support Vector Machine Classifier	0.9379
CNN	0.8983
Multilingual BERT	0.8757
DistilBERT	0.5593
XML-RoBERTa	0.8531
Multilingual DeBERTa	0.9265

Table 6: Accuracy of the author gender identification task for Urdu articles using only most frequent MWEs as features

Classifier	100	300	600	879
Light Gradient Boosted Machine Classifier	0.7910	0.7966	0.7966	0.7966
Cat Boosted Classifier	0.8249	0.8305	0.8701	0.8362
Extreme Gradient Boosted Classifier	0.7797	0.8023	0.7910	0.8079
Gradient Boosting Classifier	0.8192	0.8249	0.8475	0.8305
Random Forest Classifier	0.7910	0.8249	0.8531	0.8079
Extra Trees Classifier	0.8192	0.8023	0.8362	0.8136
Ada Boost Classifier	0.7853	0.7458	0.7514	0.7514
K Nearest Neighbors Classifier	0.6723	0.5932	0.6045	0.5763
Decision Tree Classifier	0.7175	0.7514	0.7627	0.7458
Support Vector Machine Classifier	0.8192	0.8305	0.8305	0.8192

3.4 Effect of varying the number of the most frequent words as Features

In this study, I investigate the performance of different classifiers by varying the number of the most frequent words as the features for the author gender identification task for Urdu. As can be seen from Table 7, using 600 most frequent words as features reports 93.22% accuracy using the Support Vector Machine

classifier. However, I found that concatenating the 300 most frequent words with the 600 most frequent MWEs results in the best accuracy (see Table 4 for more details).

Table 7: Accuracy of the author gender identification task using the most frequent words only as features

Classifier	100	300	600	900
Light Gradient Boosted Machine Classifier	0.8475	0.8531	0.8870	0.9096
Cat Boosted Classifier	0.8701	0.8644	0.8927	0.8814
Extreme Gradient Boosted Classifier	0.8305	0.8644	0.8814	0.8870
Gradient Boosting Classifier	0.8305	0.8870	0.8644	0.8588
Random Forest Classifier	0.8192	0.8531	0.8983	0.8983
Extra Trees Classifier	0.8588	0.8701	0.8757	0.9040
Ada Boost Classifier	0.8475	0.7853	0.8814	0.8475
K Nearest Neighbors Classifier	0.6610	0.7119	0.7345	0.7345
Decision Tree Classifier	0.7232	0.6949	0.7119	0.6610
Support Vector Machine Classifier	0.8701	0.9209	0.9322	0.9266

3.5 Effect of varying the number of the most frequent variable length words n-grams as Features

In this study, I investigated the performance of different classifiers by varying the number of the most frequent variable length words n-grams as the features where the values of n are between 2 and 10. As can be seen from Table 8, using the 600 most frequent variable length word n-grams as features reported 87.57% accuracy using the Support Vector Machine classifier.

3.6 Effect of varying the number of the most frequent characters as Features

In this study, I investigated the performance of different classifiers by varying the number of the most frequent characters as the features for the author gender identification task for Urdu. As can be seen from Table 9, varying the number of the most frequent characters as features from 100 to 900 does not effect the accuracy and that Cat Boosted Classifier reported 82.49% accuracy.

3.7 Effect of varying the number of the most frequent variable length character n-grams as Features

In this study, I investigated the performance of different classifiers by varying the number of the most frequent variable length character n-grams as the features where the values of n are between 2 and 10. As can be seen from Table 10, using the 300 most frequent variable length word n-grams as features reported 92.66% accuracy.

Table 8: Accuracy of the author gender identification task for Urdu articles using the most frequent variable length words n-grams only as features where the values of the n are between 2 and 10

Classifier	100	300	600	900
Light Gradient Boosted Machine Classifier	0.8305	0.8362	0.8418	0.8475
Cat Boosted Classifier	0.8475	0.8475	0.8588	0.8475
Extreme Gradient Boosted Classifier	0.8362	0.8192	0.8588	0.8531
Gradient Boosting Classifier	0.8418	0.8192	0.7910	0.8079
Random Forest Classifier	0.8475	0.8475	0.8079	0.8475
Extra Trees Classifier	0.8644	0.8305	0.8305	0.8249
Ada Boost Classifier	0.8249	0.7627	0.7458	0.8079
K Nearest Neighbors Classifier	0.7119	0.6271	0.6554	0.6045
Decision Tree Classifier	0.7232	0.7119	0.7458	0.7345
Support Vector Machine Classifier	0.8701	0.8079	0.8757	0.8701

Table 9: Accuracy of the author gender identification task using the most frequent characters as features

Classifier	100	300	600	900
Light Gradient Boosted Machine Classifier	0.7853	0.7853	0.7853	0.7853
Cat Boosted Classifier	0.8249	0.8249	0.8249	0.8249
Extreme Gradient Boosted Classifier	0.8079	0.8079	0.8079	0.8079
Gradient Boosting Classifier	0.7627	0.7627	0.7627	0.7627
Random Forest Classifier	0.8192	0.8192	0.8192	0.8192
Extra Trees Classifier	0.8023	0.8023	0.8023	0.8023
Ada Boost Classifier	0.7910	0.7910	0.7910	0.7910
K Nearest Neighbors Classifier	0.6158	0.6158	0.6158	0.6158
Decision Tree Classifier	0.7119	0.7119	0.7119	0.7119
Support Vector Machine Classifier	0.6610	0.6610	0.6610	0.6610

Table 10: Accuracy of the author gender identification task for Urdu articles using the most frequent variable length characters n-grams as features

Classifier	100	300	600	900
Light Gradient Boosted Machine Classifier	0.8531	0.8757	0.8644	0.8757
Cat Boosted Classifier	0.8475	0.8983	0.9209	0.9040
Extreme Gradient Boosted Classifier	0.8870	0.8983	0.8814	0.8927
Gradient Boosting Classifier	0.8192	0.8588	0.8814	0.8757
Random Forest Classifier	0.7966	0.8927	0.8588	0.8983
Extra Trees Classifier	0.8531	0.8701	0.8588	0.8588
Ada Boost Classifier	0.8023	0.8475	0.8079	0.8362
K Nearest Neighbors Classifier	0.7571	0.7797	0.8192	0.8362
Decision Tree Classifier	0.6893	0.7401	0.7514	0.7458
Support Vector Machine Classifier	0.8870	0.9266	0.9209	0.9266

4 Conclusions and Future Works

Author gender identification is an important natural language processing task. This task has been extensively investigated for resource-rich languages. However, the applications of this task are not limited to resource-rich languages only. Therefore, I presented the first investigation on the author gender identification task for Urdu articles. I also explored, effectiveness of the multi-word expressions for the author gender identification task. I propose to use the multi-word expressions as the features for the author gender identification task by concatenating them with most frequent words. The experimental findings revealed that, despite the popularity of the pre-trained language models for the natural language processing tasks, they are unable to outperform the classical machine learning classifiers in gender prediction for low-resource languages. In future, I plan to investigate the effect of article size and the number of articles per class on the accuracy of the author gender identification task.

References

1. Al-Ghadir, A.R.I., Azmi, A.M.: A study of arabic social media users - posting behavior and author's gender prediction. *Cogn. Comput.* **11**(1), 71–86 (2019)
2. Alsmearat, K., Al-Ayyoub, M., Al-Shalabi, R., Kanaan, G.: Author gender identification from arabic text. *Journal of Information Security and Applications* **35**, 85–95 (2017)
3. Baseer, F., Jaafar, J., Habib, A.: Gender and age identification through romanized urdu dataset. In: 2019 1st International Conference on Artificial Intelligence and Data Sciences (AiDAS). pp. 164–169. IEEE (2019)
4. Bassem, B., Zrigui, M.: Gender identification: a comparative study of deep learning architectures. In: International Conference on Intelligent Systems Design and Applications. pp. 792–800. Springer (2018)
5. Baxevanakis, S., Gavras, S., Mouratidis, D., Kermanidis, K.L.: A machine learning approach for gender identification of greek tweet authors. In: Makedon, F. (ed.) PETRA '20: The 13th PErvasive Technologies Related to Assistive Environments Conference, Corfu, Greece, June 30 - July 3, 2020. pp. 57:1–57:4. ACM (2020)
6. Cheng, N., Chandramouli, R., Subbalakshmi, K.: Author gender identification from text. *Digital Investigation* **8**(1), 78–88 (2011)
7. Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V.: Unsupervised cross-lingual representation learning at scale. *CoRR* **abs/1911.02116** (2019), <http://arxiv.org/abs/1911.02116>
8. Daud, A., Khan, W., Che, D.: Urdu language processing: a survey. *Artificial Intelligence Review* **47**(3), 279–311 (2017)
9. Devlin, J., Chang, M., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR* **abs/1810.04805** (2018), <http://arxiv.org/abs/1810.04805>
10. Fatima, M., Hasan, K., Anwar, S., Nawab, R.M.A.: Multilingual author profiling on facebook. *Information Processing & Management* **53**(4), 886–904 (2017)
11. HaCohen-Kerner, Y.: Survey on profiling age and gender of text authors. *Expert Syst. Appl.* **199**, 117–140 (2022)

12. He, P., Gao, J., Chen, W.: Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *ArXiv* (2021)
13. Hearst, M.A., Dumais, S.T., Osuna, E., Platt, J., Scholkopf, B.: Support vector machines. *IEEE Intelligent Systems and their applications* **13**(4), 18–28 (1998)
14. Ikae, C., Savoy, J.: Gender identification on twitter. *Journal of the Association for Information Science and Technology* **73**(1), 58–69 (2022)
15. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., Liu, T.Y.: Lightgbm: A highly efficient gradient boosting decision tree. *Advances in neural information processing systems* **30** (2017)
16. Kucukyilmaz, T., Deniz, A., Kiziloz, H.E.: Boosting gender identification using author preference. *Pattern Recognit. Lett.* **140**, 245–251 (2020)
17. Malik, M.K.: Urdu named entity recognition and classification system using artificial neural network. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)* **17**(1), 1–13 (2017)
18. Mukherjee, A., Liu, B.: Improving gender classification of blog authors. In: *Proceedings of the 2010 conference on Empirical Methods in natural Language Processing*. pp. 207–217 (2010)
19. Safara, F., Mohammed, A.S., Potrus, M.Y., Ali, S., Tho, Q.T., Souri, A., Janenia, F., Hosseinzadeh, M.: An author gender detection method using whale optimization algorithm and artificial neural network. *IEEE Access* **8**, 48428–48437 (2020)
20. Sag, I.A., Baldwin, T., Bond, F., Copestake, A., Flickinger, D.: Multiword expressions: A pain in the neck for nlp. In: *International conference on intelligent text processing and computational linguistics*. pp. 1–15. Springer (2002)
21. Sanchez-Perez, M.A., Markov, I., Gómez-Adorno, H., Sidorov, G.: Comparison of character n-grams and lexical features on author, gender, and language variety identification on the same spanish news corpus. In: *International Conference of the Cross-Language Evaluation Forum for European Languages*. pp. 145–151. Springer (2017)
22. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv* **abs/1910.01108** (2019)
23. Sarwar, R., Hassan, S.U.: Urduai: Writeprints for urdu authorship identification. *Transactions on Asian and Low-Resource Language Information Processing* **21**(2), 1–18 (2021)
24. Simaki, V., Aravantinou, C., Mporas, I., Kondyli, M., Megalooikonomou, V.: Sociolinguistic features for author gender identification: From qualitative evidence to quantitative analysis. *J. Quant. Linguistics* **24**(1), 65–84 (2017)