Hallawi et al. Iraqi Journal of Science, 2024, Vol. 65, No. 11, pp: 6725- 6737
DOI: 10.24996/ijs.2024.65.11.42

ISSN: 0067-2904

Gender Classification Based on Iraqi Names Using Machine Learning

Huda Hallawi[1], Ahmed F Almukhtar[1*], Dhamyaa A. Nasrawi[2], Ali Durr Salah[2], Tariq Zaid Faisal[2]

[1]*Department of Information Technology, College of Computers Science & Information Technology, University of Kerbala, Karbala, Iraq*

[2]*Department of Computer Science, College of Computers Science & Information Technology, University of Kerbala, Karbala, Iraq*

Received: 19/1/2024 Accepted: 23/7/2024 Published: 30/11/2024

Abstract

In machine learning, the classification task is about building a model to predict a class of elements based on their attributes and set of examples. This work aims to classify people based on their names. Two models were developed; the former is based on a single feature that is represented by a name. Whereas the latter is built upon nine features derived from the name itself, which are: is_longname, is_vowelend, is_vowelbegin, 2_gramend, 2_grambegin, 1_gramend, 1_grambegin, is_contain_abo, and is_contain_abed. Furthermore, two datasets were utilized: the first was collected from the Ministry of Labor and Social Affairs, while the second was gathered from the Iraqi university website. There are a lot of strange IRAQI names in two datasets, as well as spelling errors, which represent a real challenge in the classification process. Five machine learning methods were applied and tested within the developed models, including Random Forest, Naive Bayes, Logistic Regression, Multilayer Perceptron, and Extreme Gradient Boost. Ultimately, the experimental results demonstrate an increase in accuracy when applying the model to the original dataset, which includes names and their frequencies. The Multilayer Perceptron has achieved 97% accuracy in one feature model, while the Extreme Gradient Boost has achieved 97% accuracy in the multi-feature model. On the other hand, the results do not exceed 79% when the models are applied to the unique dataset (names without their frequencies).

Keywords: Gender classification, machine learning techniques, Iraqi names, multi features, unique dataset.

تصنيف الجنس على أساس الأسماء العراقية باستعمال التعلم الآلي

هدى الحلاوي[1], احمد فائق المختار*[1], ضمياء نصراوي[2]،علي الدر صلاح[2]، طارق زيد فيصل[2]

[1]قسم تكنولوجيا المعلومات، كلية علوم الحاسوب وتكنولوجيا المعلومات، جامعة كربلاء، كربلاء، العراق

[2]قسم علوم الحاسوب، كلية علوم الحاسوب وتكنولوجيا المعلومات، جامعة كربلاء، كربلاء، العراق

* Email: ahmed.almukhtar@uokerbala.edu.iq

الخلاصة

تعد مهمة التصنيف في التعلم الآلي مهمة بناء نموذج للتنبؤ بصنف العناصر او البيانات بالاعتماد على سماتها ومجموعة من الأمثلة. يهدف هذا العمل إلى تصنيف الأشخاص اعتمادا على أسمائهم حيث تم تطوير نموذجين، الأول يعتمد على سمة واحدة تتمثل بالاسم والنموذج الثاني يعتمد على تسع ميزات مشتقة من الاسم نفسه هي: (هل الاسم طويل، هل ينتهي بحرف علة، هل يبدأ بحرف علة، اخر حرفين، اول حرفين، اخر حرف، اول حرف، هل يحوي على المقطع 'ابو'، هل يحوي على المقطع 'عبد').

استعملت مجموعتين من بيانات، حيث تم جمع إحداهما من وزارة العمل والشؤون الاجتماعية، والاخرى من مواقع الجامعات العراقية. تحوي البيانات على عدد كبير من الاسماء العراقية الغريبة بالإضافة الى الاخطاء الاملائية وهذا يعد تحديا حقيقيا في عملية التصنيف. تم تطبيق خمس طرائق للتعلم الآلي واختبار البيانات المجمعة وهي (Random Forest، Extreme gradient Boost، Naive Bayes، Logistic Regression، Multilayer Perceptron). اظهرت النتائج أن الدقة كانت أعلى عندما تم تطبيق النموذج على مجموعة البيانات الأصلية (الأسماء مع تكراراتها)، حيث نجحت تقنية Multilayer Perceptron في تحقيق دقة بنسبة 97% مع نموذج السمة الواحدة، في حين حققت تقنية Extreme gradient Boost دقة 97% في نموذج متعدد السمات. من ناحية أخرى، لم تتجاوز النتائج 79% عندما تم تطبيق النماذج على مجموعة البيانات الفريدة (الأسماء دون تكراراتها).

1. Introduction

In recent times, there has been a growing demand for inferring the gender of a person due to its vital and beneficial impacts in several fields such as legal interrogation, e-commerce platforms, investigating election polls, marketing, and forensics [1–8]. Conducting the gender of a person is essential in a wide range of research, such as investigations of psychological, anthropological, and sociological research. Therefore, gender classification techniques are highly considered an effective and proficient area of research [2], [4], [5].

Generally, gender recognition is a two-class problem in which any given entity should be classified into one of the two classes (male or female). This is a trivial mission for humans, but it is challenging for machines. Many potential solutions are applied to deal with the given problem, including the human-computer interaction approach [3], [9], text-based approach [7-8], voice-based approach [11], [12], and face recognition approach [9], [10], [13], and [14].

The text-based approach for gender detection depends on human natural texts, which an enormous number of experts are seeking. Some models have been established based on investigating various dialects posted in online web applications (Facebook, Twitter) to infer gender type [8], [10], [15], and [16]. Others were built by analyzing usernames from user profiles on social networks; almost the first name is used to predict the gender type [10], [17], [18], [19]. Names are varied based on the language in use; therefore, different languages were investigated in research, such as Bengali [10], Chinese [17], English [18], [20], India, Western countries, Sri Lanka, and Japanese [21]. This study is addressed to solve the problem of Arabic name-based gender recognition; the problem is characterized by a lack of data.

A wide range of machine learning, neural networks, and deep learning neural network techniques were applied for solving name-based gender detection, such as MLP, RNN, GRU, CNN, BiLSTM [18], LSTM, CNN, LeNet-5 [21], and Rectified Linear Unit (ReLU) [10].

In this work, two datasets were utilized: the first was collected from the Ministry of Labor and Social Affairs, while the second was gathered from the Iraqi university website. There are a lot of strange IRAQI names in two datasets, as well as spelling errors, which represent a real challenge in the classification process. The aim of this work is to develop two models to predict gender type based on Iraqi names. The first is based on a single feature that is represented by a name, whereas the second is built according to nine features: is_longname, is_vowelend, is_vowelbegin, 2_gramend, 2_grambegin, 1_gramend, 1_grambegin, is_contain_abo, and is_contain_abed. Five machine learning methods are applied (random forest, Naïve Bayes, logistic regression, multilayer perceptron, and extreme gradient boost). The main contributions to this work are:

1- Dealing with strange Iraqi dataset names (their frequency is less than 3).
2- Extracting multiple features to predict the gender type instead of the name itself.
3- Predicting the gender type with spelling errors in names is important because multi-feature extraction helps focus on significant parts of the data.

The rest of the sections are organized as follows: Section 2: Related Work, which discusses the most relevant works. Section 3 illustrates the proposed model of our work. Section 4 discusses an evaluation and the experimental results. Finally, Section 5 concludes this paper.

2. Related Work

This section is dedicated to investigating the most relevant research on text-based gender recognition problems using natural language processing (NLP), as shown in Table 1. In the past two decades, numerous techniques have been issued regarding the name-gender problem. These techniques are enormously diverse, but they almost all make use of machine learning and deep learning techniques.

In [8], the authors analyzed the Iraqi dialects from different Iraqi regions to detect differences in genders, ages, and regions. The proposed model is based on statistical techniques; the challenge was collecting the words of various ages and regions. One of the most important research findings is that gender is a leading factor in influencing the use of variant Iraqi dialects.

The authors in [10] proposed an advanced neural network model to categorize the gender of conventional Bangla names. Six of the traditional machine learning classifiers, including Logistic Regression (LR), Stochastic Gradient Descent (SGD), Decision Tree (DT), Naive Bayes (NB), Random Forest (RF), and K Nearest Neighbor (KNN), were also trained and compared with the proposed neural network model. From all the compared models, the neural network model shows better performance in terms of accuracy, precision, recall, and F1 score. It succeeded in achieving a 73.04% accuracy rate. However, the tested models come with one limitation, which is the data volume.

Several character-based techniques were conducted in [15] and are based on the character encoding of the first name spelling. The proposed techniques are constructed as linear models accompanied by numerous deep neural models such as character-based LSTM, BERT, embedding-based models, and baseline content-based models. All the applied machine learning models are built upon computing the probability of P(M|X), where X is the user feature All. The tested models show very good performance in detecting the gender of large datasets, including multilingual datasets. Furthermore, they succeeded in minimizing the binary entropy loss function.

In [16], the authors analyzed the characteristics of user profiles on Twitter, along with the user's tweets, to investigate gender types. The proposed approach is based on a combination of advanced word embedding techniques, including Term Frequency-Inverse Document Frequency (TF-IDF), Bag of Words (BOW), GloVE, BERT, and GPT2. The outcomes show that the proposed model achieves a 10% higher accuracy (about 67%) compared with the same methods applied to user tweets only.

Gender recognition using the first name is not a trivial problem, especially in the NLP area. The authors in [19] have applied and tested two groups of models to classify Brazilian names; the first group uses machine learning algorithms. In contrast, the second group applies deep neural network tools. Several machine learning procedures were implemented and tested, such as NB, SVM, RF, gradient boosting, light GBM, logistic regression, and ridge classifiers.

Five models were implemented and examined regarding the deep neural network, including MLP, RNN, GRU, CNN, and BiLSTM. The two groups of models were trained and analyzed using a Brazilian name data set against several objective functions encompassing accuracy, recall, precision, f1 score, and confusion matrix. The outcomes show that RNN, BiLSTM, and GRU outperform the other tested models in most cases. Primarily, the BiLSTM model comes with exceeding precision and a gender ratio of 0.9720 and 100%, respectively. The same data set of Brazilian names was used by the authors in [20] to compare the accuracy of two types of deep neural models: feed-forward (SVM, KNN) and recurrent models (RNN, BiLSTM, GRU). The results show that the recurrent models achieve better accuracy than the feed-forward models in this problem.

In [22], the authors identified a degradation problem and low performance when applying traditional machine learning techniques used for gender detection in English names with Asian languages. The traditional techniques have deteriorated with Chinese names; therefore, they proposed a new approach for word embedding based on simplifying Chinese characters and incorporating phonetic information. The new proposed approach was trained and tested using the BERT technique. The new model quantitatively outperforms other traditional machine learning techniques, including Naïve Bayes, GBDT, and Random Forest. The experimental results of the new model demonstrate that 93.45% accuracy was tested with enormous gender-labeled datasets ranging from one to over six million names.

The authors in [23] established a special machine-learning model for gender classification over a large-scale dataset of about 3 million Vietnamese names. Primarily, the N-gram algorithm is adopted to apply for the full name of the Vietnamese language. In most languages, such as English, German, Portuguese, Brazilian, and Arabic, the first name plays a crucial role in gender classification issues. However, the middle name in the Vietnamese language plays a major role in gender classification. It directly affects the classification accuracy; applying just the middle name, the model accuracy is 76.1%. A model based on N-gram for the full name, combined with the whole full name, achieves 90.9% accuracy.

In [24], the researchers look at what happens when you use a deep learning model (LSTM) and six machine learning algorithms to figure out the gender of Arabic names. The algorithms are Support Vector Machine, Multinomial Naive Bayes, Bernoulli Naive Bayes, Decision Tree, Random Forest, and Logistic Regression. Furthermore, a dataset is produced to examine how these algorithms perform when applied to Arabic names. Two different datasets exist: the English-language dataset is designated as AE (22,850 samples), while the Arabic-language dataset is designated as AA (21,320 samples). A deep learning model achieved 94% on the AA dataset and 96% on the AE dataset, together with the five machine learning methods used in the experiment.

Table 1: Overview of the Literature Review

N	Ref	Dataset	Method	Result	Strength	Weakness	Language
1	[8]	Iraqi corpus from TieQar University	Statistical technique		It succeeded in working with different Iraqi dialects	lacks AI techniques	Iraqi
2	[15]	Huge corpus from Yahoo Data and SSA Data	DLSTM and LSTM, EMB,	(90%) Accuracy	Character encoding of the first name	Probability	English
3	[16]	A huge corpus from users' profiles on Twitter	TF-IDF, BOW, GloVE, BERT, and GPT2	(67%) Accuracy	Two groups of data are applied user tweets and user profile	--------	English
4	[22]	Dataset of six million Chinese names.	BERT	93.45% Accuracy	high performance with Chinese names (new embedding by simplifying characters and incorporating phonetic information)		Chinese
5	[10]	Corpus of conventional Bangla names	Neural network model	73.04% Accuracy	Advanced neural network model	Data volume	Bangla
6	[19]	Brazilian names data set	[NB, SVM, RF, gradient boosting, light GBM, logistic regression, ridge classifier], [MLP, RNN, GRU, CNN, and BiLSTM]	BiLSTM (0.9720 Precision and 100% gender ratio)	It succeeded in working with Brazilian names	Special encoding characters	Brazilian
7	[20]	Brazilian names data set	CNN, BiLSTM, RNN, MLP, GRU	(92:67% ,95:89% ,93:85% ,86:92% 94:80%)	Deep network models succeed in capturing dependencies in the word vector.	Special encoding characters	Brazilian
8	[23]	Corpus of 3 million Vietnamese	N-gram model	90.9% Accuracy	It works efficiently with Vietnamese	It is directly affected by the middle name	Vietnam
8	[24]	English-language dataset is designated as AE and the Arabic-language dataset is designated as AA	deep learning model (LSTM) and six machine learning algorithms (SVM, NaiveBayes(Bernoulli and Multinomial), Decision Tree, Random Forrest, and Logistic Regression)	94% on the AA dataset and 96% on the AE F1 score	LSTM work efficiently with AA	The variation of spelling some Arabic names in English have played a role in the results in general.	Arabic written in Arabic and English

3. Methodology

The proposed model applies five methods, namely Random Forest, Extreme Gradient Boosting Naive Bayes, Logistic Regression, and Multilayer Perceptron, to conduct the gender classification task. The block diagram in Figure 1 primarily divides the proposed model into three main parts: data pre-processing, model training, and model prediction. Each part is further described below.

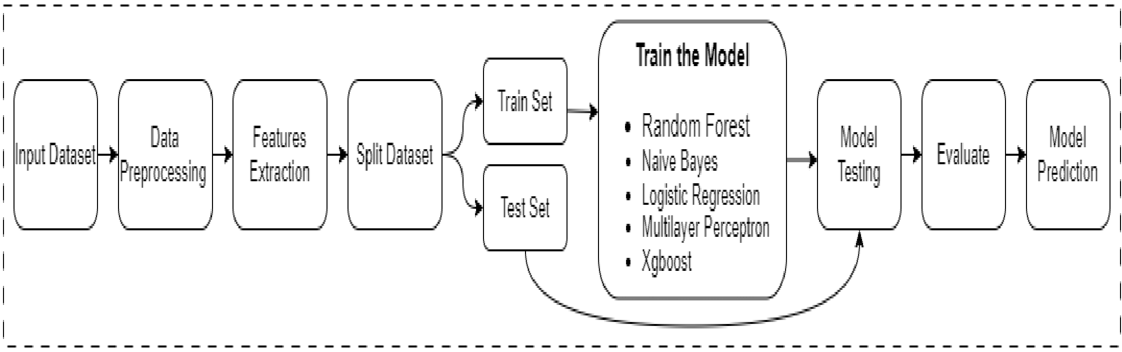

Figure 1: Steps of the gender classification model

3.1. Dataset

The Iraqi name dataset used in this work was collected from the Ministry of Labor and Social Affairs and Iraqi University (University of Mosul and University of Anbar) websites.

The original Iraqi name dataset (names with their frequencies) involves:

- 1759608 names (female: 815624, male: 943984), which means 46% female, and 54% male.
- The top three female reputed names are (‘فاطمه’: 16464, ‘زهره’: 16267, ‘زينب’:10147). On the other side, the top three male reputed names are (‘محمد’: 46753, ‘علي’: 33020, ‘حسين’: 32687).
- The unique Iraqi names dataset (names without their frequencies) consists of 49391 names (female: 23334, male: 26057). that means 47% female and 53% male, as shown in **Figure 2:** Count of genders in the unique Iraqi names dataset.
- The number of shared names in both genders is 4692, such as ‘صفاء’,’نور’, and ‘نضال’, about 9.5% in the unique Iraqi names dataset.

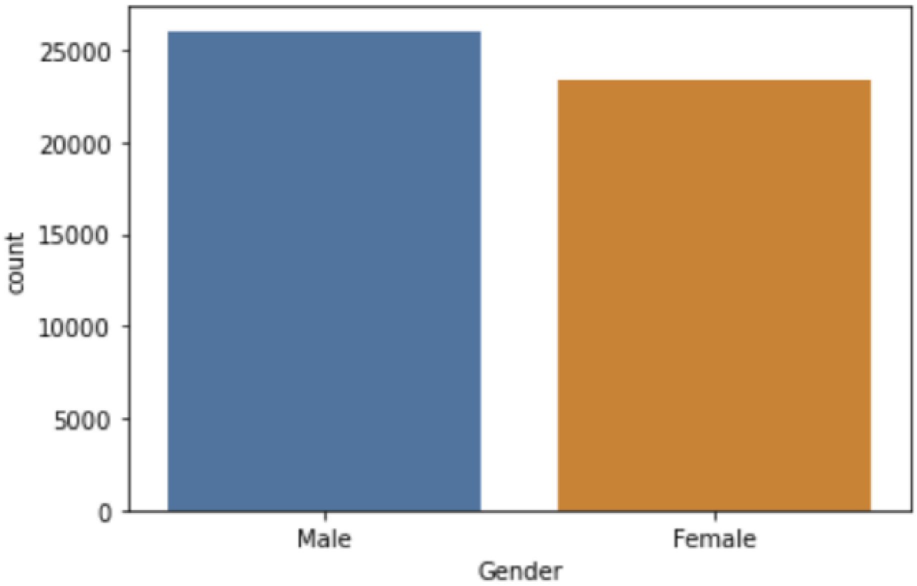

Figure 2: Count of genders in the unique Iraqi names dataset

3.2. Pre-processing

The collected dataset was noisy because it contained numerous typos and errors. As a result, they went through six steps for pre-processing and cleaning their data:

1- Eliminating special characters.
2- Eliminating unrequired spaces such as ‘ الا ء ’ or ‘ الاء’, ‘الا ء ’, to become ‘الاء’.
3- Correcting the spelling errors.
4- Eliminating all names with one character.
5- Tokenizing the concatenated names such as ‘احمدعمر’ to become ‘احمد’, and ‘عمر’.
6- Normalization of some letters such as ‘إأآٱ’ to ‘ا’, ‘ة’ to ‘ه’.

After performing the pre-processing steps on the original and unique datasets, feature extraction must be implemented to feed them into models for the training and testing processes.

3.3. Feature Extraction

Feature extraction helps to efficiently generate new features from the original data that reflect key information by focusing on the most significant parts of the data. Two models were implemented in this work; the first is a single feature model that utilizes the name to predict gender type. The second model relies on nine distinct features that aid in prediction without revealing the name itself. The nine features, as shown in Table 2, include: (is_longname, is_vowelend, is_vowelbegin, 2_gramend, 2_grambegin, 1_gramend, 1_grambegin, is_contain_abo, is_contain_abed).

Table 2: Sample of the Iraqi name dataset with feature extraction

N	Gender	Name	Long_ Name	Vowel_ End	Vowel_ Begin	N2_Gram _End	N2_Gram _Begin	N1_G ram _End	N1_Gr am _Begin	ABO	ABED
0	1	فجه	0	0	0	جه	فج	ه	ف	0	0
1	1	انتهاء	1	0	1	اء	ان	ء	ا	0	0
2	1	فاطمه	0	0	0	مه	فا	ه	ف	0	0
3	0	سعد	0	0	0	عد	سع	د	س	0	0
4	0	عبدالله	1	0	0	له	عب	ه	ع	0	1
5	1	رازقيه	1	0	0	يه	را	ه	ر	0	0
6	0	احمد	0	0	1	مد	اح	د	ا	0	0

3.4. Model Training and Testing

The dataset is split into two main parts: data training (80%) and testing (20%). A count vectorizer is employed for vectorization, and then the models are trained on the data and evaluated using five supervised machine learning methods. The characteristics of each method will be further discussed separately.

3.4.1. Random Forest (RF)

A well-known algorithm for solving the problem of overfitting in decision trees [25] aggregates the outcomes of multiple decision trees to produce a conclusive result. The idea behind the Random Forest algorithm is the grouping and bagging concept. First, it builds a group of sub-decision trees to train the generated subsets of the dataset, and then the final result is built through the bagging method by assembling the gathered output from multiple models. The key point of the method is reducing the correlation by choosing a random subset of features to create models of small decision trees. Implementing this method involves setting the following parameters:

$$criterion\ =\ 'gini', n\_estimators = 100, \max\ \_depth = 9$$

3.4.2. Naïve Bayes (NB)

A popular supervised machine learning algorithm is based on the assumption that all the predictions are constructed by self-dependent features. This means that it re-diffuses the inputs for a given class or category. The method is derived from the Bayes theorem to explore a class from unexplored datasets. The main factor of the technique is its kernel function, which permits the classifier to enhance its performance when assessing the probability of the density function for input data. This method is implemented with multinomial Naïve Bayes in the default parameters.

3.4.3. Logistic Regression (LR)

Logistic regression is a statistical approach that can be used for categorical classification, where the belongingness of the predicted object depends on different parameters. The technique follows the general regression concept, where the association between the dependent and independent variables is the subject of interest. The goal behind logistic regression is to describe the correlation between dependent and independent variables by approaching the best-fitting model. Generally speaking, the reason for using logistic regression rather than other models of regression is because it has more adaptive capabilities for real-world scenarios where the classification situation of the prediction for continuous value is unbounded and probabilistic. This method is implemented with the default parameters.

3.4.4. Multilayer Perceptron (MLP)

The method, a multilayer perceptron classifier, is designed to tackle the limitation of linearity in the hyperplane, considering the XOR gate, for example, where the input data are associated with the output in a non-linear way. Mainly, the algorithm consists of three or more layers, one for input and another for output, and the rest of the hidden layers are stacked between the input and output layers [26]. The results from the current layer are fed to the next layer based on a feedforward mechanism, whereas the backpropagation concept permits the network weights to be adjusted iteratively. Implementing this method requires the following parameters:

$$activation =' relu', learning\_rate = 'constant', early\_stopping = False, alpha = 0.0001, hidden\_layer\_sizes = (100, 3)$$

3.4.5. eXtreme Gradient-Boost (XGBoost)

XGBoost is an updated version of gradient-boosting decision trees; however, it is considered a fast and easy-to-use algorithm, does not require tuning or parameter optimization, and is suitable for large datasets [27]. The algorithm is used for supervised learning and accepts multiple features x_i for training to predict the output of variables y_i. The algorithm primarily consists of two terms: the regularization term and the loss function term. The former term is devoted to adjusting the model complexity, which can help in avoiding model overfitting. The latter, the loss function, specifies the prediction approach such that the model prediction concerns the trained data. This method is implemented with the default parameters.

4. Results and Evaluation

This section discusses the performance of gender classification models. We consider precision, recall, macro F1 score, and accuracy as metrics for the model's performance. Two models are applied to the original and unique datasets; the first model uses a single feature, and the second uses multiple features, as shown in Tables 3, 4, and 5.

Table 3: Results of machine learning methods using a single feature

Models	Precision		Recall		F1-score		Accuracy	
	Unique	**Original**	**Unique**	**Original**	**Unique**	**Original**	**Unique**	**Original**
Random Forest	76	89	75	89	75	89	75	89
Naïve Bayes	75	81	74	82	75	81	75	82
Logistic Regression	76	85	75	84	75	84	75	85
MLP Classifier	75	97	75	97	75	97	75	97
Xgboost	76	95	76	95	76	95	76	95

Table 4: Results of machine learning methods using multiple features

Models	Precision		Recall		F1-score		Accuracy	
	Unique	**Original**	**Unique**	**Original**	**Unique**	**Original**	**Unique**	**Original**
Random Forest	80	94	78	93	78	93	79	93
Naïve Bayes	68	72	68	71	67	70	67	70
Logistic Regression	70	77	70	77	70	77	70	77
MLP Classifier	79	95	77	95	78	95	78	95
Xgboost	79	97	78	97	79	97	79	97

Table 5: Summary of results with the best models of gender classification

Dataset	Features	Best Model	Accuracy
Original	Single	MLPClassifier	97
	Multi	Xgboost	97
Unique	Single	Xgboost	76
	Multi	Random Forest, Xgboost	79

According to the results, the accuracy was higher when the model was applied to the original dataset (names with their frequencies), Multilayer Perceptron achieved 97% accuracy in one feature model, and Extreme Gradient Boost achieved 97% accuracy in a multi-feature model. On the other hand, the results do not exceed 79% when using the models on a unique dataset (names without their frequencies). In one feature model, Extreme Gradient Boost achieves 76% accuracy, whereas in multi-feature models, Extreme Gradient Boost and Random Forest achieve 79% accuracy. The unique dataset's results are almost lower than the original dataset's because the training stage heavily relies on the frequency of names, and most machine learning methods based on probabilistic models have a greater impact on classification accuracy than the original dataset.
Moreover, a comparison of the performance measures between our work and related research is reported in Table 6. As shown in Table 6, the proposed model gains higher accuracy in classifying gender.

Table 6: Comparison of our work with related work

N	Ref	Dataset	Method	Result	Language
1	[15]	Huge corpus from Yahoo Data and SSA Data	DLSTM and LSTM, EMB,	(90%) Accuracy	English
2	[16]	A huge corpus from users' profiles on Twitter	TF-IDF, BOW, GloVE, BERT, and GPT2	(67%) Accuracy	English
3	[22]	Dataset of six million Chinese names.	BERT	93.45% Accuracy	Chinese
4	[10]	Corpus of conventional Bangla names	Neural network model	73.04% Accuracy	Bangla
5	[19]	Brazilian names data set	[NB, SVM, RF, gradient boosting, light GBM, logistic regression, ridge classifier], [MLP, RNN, GRU, CNN, and BiLSTM]	BiLSTM (0.9720 Precision and 100% gender ratio)	Brazilian
6	[20]	Brazilian names data set	CNN, BiLSTM, RNN, MLP, GRU	(92:67% ,95:89% ,93:85% ,86:92% 94:80%)	Brazilian
7	[23]	Corpus of 3 million Vietnamese	N-gram model	90.9% Accuracy	Vietnam
8	[24]	English-language dataset is designated as AE (22,850 samples), and the Arabic-language dataset is designated as AA (21,320 samples).	deep learning model (LSTM) and six machine learning algorithms (SVM, NaiveBayes(Bernoulli and Multinomial), Decision Tree, Random Forrest, and Logistic Regression)	94% on the AA dataset and 96% on the AE F1 score	Arabic written in Arabic and English
9	Our work	two datasets collected from the Ministry of Labor and Social Affairs and from the Iraqi university website	Two models based on a single and multi-feature with five machine learning methods (Random Forest, Naïve Bayes, Logistic Regression, Multilayer Perceptron, and Extreme gradient boost).	97% Accuracy	Arabic/IRAQI

5. Conclusion and Future Work

The problem of predicting gender based on strange Iraqi names with spelling errors is considered. Two models, single-feature and multi-feature, have been applied to five machine learning methods. The tests show that the original dataset does better than the unique dataset for all methods and both models. This is because the two datasets have different frequencies, resulting in changes in probabilities that improve classification accuracy in the original dataset.

Concerning the unique dataset, the performance of the tested methods for both single-feature and multi-feature models has achieved average accuracy (75.2 and 74.6), respectively. In terms of the best and worst methods, Table 4 explains the best accuracy using two methods (MLPClassifier and Xgboost) with the original dataset, while Naive Bayes and Logistic Regression achieve the worst accuracy using the unique dataset.
Comparing with the related work with English, Vietnam, Brazilian, Chinese, Bangla, and Arabic name datasets, the proposed model with the original Iraqi name dataset gains 97% accuracy.

The future work will be dedicated to enriching the dataset and applying deep learning instead of traditional machine learning techniques.

6. Acknowledgments

We express our gratitude to the Computer Science and Information Technology College at the University of Kerbala for their continuous support and encouragement.

References

[1] S. A. Khan, M. Ahmad, M. Nazir and N. Riaz, “A Comparative Analysis of Gender Classification Techniques,” *International Journal of Bio-Science and Bio-Technology*, vol. 5, no. 4, pp. 223-244, August, 2013.

[2] I. D. Raji and J. Buolamwini, "Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 429–435, Jan. 2019. Available: https://dl.acm.org/doi/10.1145/3306618.3314244.

[3] M. Bertrand, and S. Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," *NATIONAL BUREAU OF ECONOMIC RESEARCH, Cambridge, MA 02138*, pp. 1-40, Jul. 2003. Available: http://www.nber.org/papers/w9873.

[4] K. Vikas, A. V. Reddy, S. Kumar and H. Shanmugasundaram, "User Gender Classification Based on Twitter Profile Using Machine Learning," in *Proceedings of the International Conference on Applied Artificial Intelligence and Computing (ICAAIC)*, pp. 09-11 May 2022, IEEE, India. Available: https://doi.org/10.1109/ICAAIC53929.2022.9793185.

[5] B. A. Kansake, G. B. Sakyi-Addo and N. K. Dumakor-Dupey, "Creating a gender-inclusive mining industry: Uncovering the challenges of female mining stakeholders," *Resources Policy*, vol. 70, p. 101962, Mar. 2021. Available: https://doi.org/10.1016/j.resourpol.2020.101962.

[6] G. de Felice, J. Burrai and E. Mari et al, "How Do Adolescents Use Social Networks and What Are Their Potential Dangers? A Qualitative Study of Gender Differences," *International Journal Environ. Research Public Health,* vol. 19, no. 9, p. 5691. Available: https://doi.org/10.3390/ijerph19095691.

[7] I. Van Buskirk, A. Clauset and D. B. Larremore, "An Open-Source Cultural Consensus Approach to Name-Based Gender Classification," in *Proceedings of the Seventeenth International AAAI Conference on Web and Social Media (ICWSM),* vol. 17, pp. 866-877, Jan. 2. 2023. Available: https://doi.org/10.1609/icwsm.v17i1.22195.

[8] R. S. Jaber, "The Sociolinguistics of Iraqi Women: A New Perspective," *International Journal of Linguistics Studies*, vol. 2, no. 1, pp. 45-55, 2022. Available: https://doi.org/10.32996/ijls.2022.2.1.6.

[9] T. Zhao, "Mind the Gap: Understanding Gender Inequality in Movie Industry using Social Network Analysis and Machine Learning," in *Proceedings of the 2019 International Conference on Information Technology and Computer Application (ITCA)*, pp. 100-106, Dec. 2020. Available: https://doi.org/10.1109/ITCA49981.2019.00030.

[10] J.F. Ani, M. Islam, N.J. Ria, S. Akter and A.K.M. Masum, "Estimating Gender Based On Bengali Conventional Full Name with Various Machine Learning Techniques," in *Proceedings of the 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1-6, Jul. 6. 2021. Available: https://doi.org/10.1109/ICCCNT51525.2021.9579927.

[11] H.M. Ahmed, and H. H. Mahmoud," Effect of Successive Convolution Layers to Detect Gender," *Iraqi Journal of Science*, Vol. 59, no.3C, pp. 1717-1732, 2018.

[12] U. H. Jaid, and A. K. Abdul Hassan," Review of Automatic Speaker Profiling: Features, Methods, and Challenges," *Iraqi Journal of Science*, vol. 64, no. 12, pp. 6548-6571, Dec. 2023. Available: https://doi .org//10.24996/ijs.2023.64.12.36.

[13] H. H. K. Tin, "Perceived Gender Classification from Face Images," *Journal of Modern Education and Computer Science*, vol. 4, no. 1, pp. 12-18, Jan. 2012. Available: https://doi.org/10.5815/ijmecs.2012.01.02.

[14] A. Mustafa and K. Meehan "Gender Classification and Age Prediction using CNN and ResNet in Real-Time," in *Proceedings of the International Conference on Data Analytics for Business and Industry: Way Towards a Sustainable Economy (ICDABI)*, pp. 1-6, Oct. 2020. Available: https://doi.org/10.1109/ICDABI51230.2020.9325696.

[15] Y. Hu, C. Hu, T. Tran, T. Kasturi, E. Joseph and M. Gillingham, "What's in a name? – gender classification of names with character based machine learning models", *Data Mining and Knowledge Discovery*, vol. 35, no. 4, pp. 1537-1563, May 2021. Available: https://doi.org/10.1007/s10618-021-00748-6.

[16] B. Onikoyi, N. Nnamoko and I. Korkontzelos, " Gender prediction with descriptive textual data using a Machine Learning approach," *Natural Language Processing Journal*, vol 4, p. 10018, Sep. 2023. Available: https://doi.org/10.1016/j.nlp.2023.100018.

[17] S. Yuenyong and S. Sinthupinyo, "Gender Classification of Thai Facebook Usernames," *International Journal of Machine Learning and Computing*, vol. 10, no. 5, pp. 618-623, Sep. 2020. Available: https://doi.org/10.18178/ijmlc.2020.10.5.982.

[18] A. C. Gallagher and T. Chen, " Estimating Age, Gender, and Identity using First Name Priors," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 1-8, Jun. 2008. Available: https://doi.org/10.1109/cvpr.2008.4587609.

[19] R. C. B. Rego and V. M. L. Silva, "PREDICTING GENDER OF BRAZILIAN NAMES USING DEEP LEARNING," *arXiv preprint arXiv:2106.101556v1*. Jun. 21. 2021.

[20] R. C. B. Rego, V. M. L. Silva and V. M. Fernandes, "Predicting Gender by First Name Using Character-level Machine Learning," *arXiv preprint arXiv:2106.101556v2*. Oct. 2021.

[21] C. Bhagvati, "Word Representations For Gender Classification Using Deep Learning," in *Proceedings of the International Conference on Computational Intelligence and Data Science*, vol. 132, pp. 614-622, 2018, Elsevier Ltd. Available: https://doi.org/10.1016/j.procs.2018.05.015.

[22] J. Jia and Q. Zhao, "Gender Prediction Based on Chinese Name," in *Proceedings of the Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC*, vol 11839, pp. 676–683, 2019, springer. Available: https://doi.org/10.1007/978-3-030-32236-6_62.

[23] T. H. Huong, K. Tran-Trung and V. T. Hoang, "A Computational Linguistic Approach for Gender Prediction Based on Vietnamese Names," *Mobile Information Systems*, vol. 2022, Article ID 6570228, pp. 1-6. Feb. 27. 2022. Available: https://doi.org/10.1155/2022/6570228.

[24] M. Almelabi, and R. Baashirah, "Gender Prediction Based on Arabic Names with Machine Learning Techniques", in B. Alareeni, A.Hamdan,"Explore Business, Technology Opportunities and Challenges After the Covid-19 Pandemic" , LNNS, volume 495, pp. 1248–1255, 2023. https://doi.org/10.1007/978-3-031-08954-1_107

[25] A. Cutler, D. R. Cutler and J. R. Stevens, "Random forests," in Title: Ensemble machine learning: Methods and applications, *Springer, New York, NY*, 2012, ch. 5. Available: https://doi.org/10.1007/978-1-4419-9326-7.

[26] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing*, vol 2, no. 5-6, pp. 183-197, Jul. 1991. Available: https://doi.org/10.1016/0925-2312(91)90023-5.

[27] A. Jolicoeur-Martineau, K. Fatras and T. Kachman, "Generating and Imputing Tabular Data via Diffusion and Flow-based Gradient-Boosted Trees," *arXiv preprint arXiv:2309.09968*, Feb. 2024. Available: https://doi.org/10.48550/arXiv.2309.09968.