Predicting Gender by Vietnamese Names Using Machine Learning Techniques

1st Truong Pham Le
University of Information Technology
Vietnam National University, Ho Chi Minh City, Vietnam
Ho Chi Minh City, Vietnam
20522090@gm.uit.edu.vn

2nd Thao Tran Phuong
University of Information Technology
Vietnam National University, Ho Chi Minh City, Vietnam
Ho Chi Minh City, Vietnam
20521938@gm.uit.edu.vn

***Abstract*—Nowadays, much work has been done on gender classification based on people's names. However, predicting gender by Vietnamese name is not a simple task because this language has 6 tones corresponding to 6 voices, combined with many vowels and consonants. In this paper, we examined and implemented several machine learning algorithms, such as Naive Bayes, Logistics Regression, Support Vector Machine, K-Nearest Neighbors, Decision tree, Random Forest. A dataset which over 30,000 names is used to train and evaluate the models. We analyzed the accuracy, recall, precision and f1 score to measure the models' performances. As a result, the best F1-score that we have achieved is up to 95% and we generated a web local based on our trained model.**

I. INTRODUCTION

Gender prediction is one of the most critical problems in machine learning with various applications for marketing, advertising, e-commerce, security, and human behavior[1, 2]. (here are many studies on gender identity based on facial images [3, 4], gaits [5], social media [6–8], facial images [9], ear images [10], and text [11]). In recent years, the identification of gender based on people's names has been extensively focused on by some authors [12–14]. Gender identification based on the character is a subtopic of natural language processing and text mining research. It can be supported and applied in many areas such as contextual advertising, question, and answering system, chatbot, and machine translation. In marketing, identifying the exact gender of a customer allows one to propose products to the right audience. For example, users will reduce the time in systems while they need to fill in their information. To protect and avoid fraud in the declaration, gender prediction is beneficial for different systems such as customer management systems, e-commerce, and social websites.

II. RELATED WORK

In previous related work, we found that there are many studies on gender, especially English and Chinese, along with a variety of traditional machine learning methods, deep learning, ... are given, this is a very interesting topic because a name not only refers to the way we address it but also has a special meaning and also shows the gender, nationality of a country according to us learning about this topic in the land In my country, there are only a few articles mentioned [15, 16] and found in common that only research on the Kinh ethnic group (the majority of the ethnic group in VietNamese) but my country has 54 ethnic groups. Therefore, we have collected more data on ethnic minorities to enrich the characteristics of Vietnamese names.In this study, we will execute on traditional machine learning models such as Naive Bayes, Logistics Regression, Support Vector Machine, K-Nearest Neighbors, Decision tree, Random Forest to evaluate performance. use more Votingclassifier to improve model performance

As the languages differentiate among nations and regions, a study on the component of names are necessary to further understand how their impacts on genders. Since Chinese is a logo-syllabic language, the approaches for this task are focusing on the character itself. In contrast, Vietnamese language is based on Latin alphabet which is similar to English. Therefore we only consider and compare the characteristic Vietnamese names to English names. Le [17] defined that Vietnamese name has three components, respectively surname(family name), middle name and first name. According to Le's research, the official order of Vietnamese name is differ from English names. Specifically, in English, the order of name component is first name, middle name, and family name (surname). For example, John Doe consists of John (given/first name) and Doe (family name). In a comparison, family name is placed as last component in English name, while in Vietnamese name, it is place at first position.

III. DATASET AND PRE-PROCESSING

In this section, the dataset and the pre-processing techniques used are described.

A. Dataset Collection

The dataset consists of 30851 Vietnamese names, of which 45.13% are female names and 54.87% are male, based on UIT-ViNames [15] and 4,000 data samples collected from ethnic minority schools help to diversify data collection. The labels distribution considering the gender (0 for female and 1 for male) is depicted in Figure 1.

B. Data Analysis

- Vietnam is a country with 54 ethnic groups living, so Vietnamese surnames are very diverse. For example, the

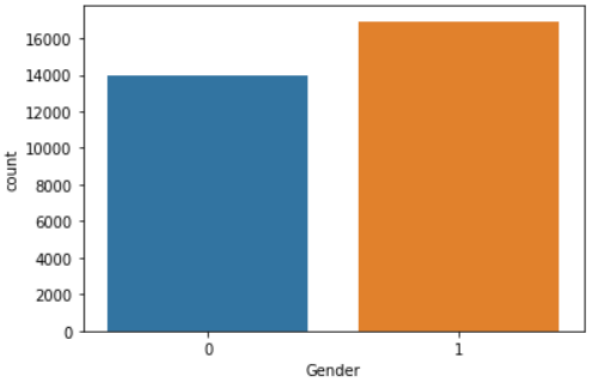

Fig. 1. The distribution male and female last names in Vietnamese..

Kinh's surname is Tran, Nguyen, Le,..., and the Khmer's surname is Kim, Son, Thach, etc. So we compared the distribution of male and female surnames for our dataset, shown in Figure 2.

Fig. 2. The distribution of male and female middle names in Vietnamese.

- Figure 2 illustrates that the most common surnames in both males and females are the same indicating that in ethnic Kinh names are the majority and surnames have no or little impact on gender determination. However, when we delve deeper into the names of the Vietnamese ethnic groups [18], we find that they use the first word in their names to distinguish between male and female like the Ede ethnic group, whose name is structured by gender-name-surname-genus, in Cham people are usually Ja(male) or M, Mn(female) and many other ethnic groups like Mang, Xa-Dieng,... Therefore, we consider last name, middle name and first name in predicting gender.

- Looking at Figure 3, it can be seen that the majority of men's middle names in Vietnam are Minh and Van; but in women, Thi is widely applied.

- Figure 4 illustrates some of the most common first names of Vietnamese males and females. In particularly, it points out that **Anh** is the one that appears frequently in both Male and Female names. For example, **Minh Anh** (Minh

Fig. 3. The distribution of male and female middle names in Vietnamese.

Fig. 4. The distribution of male and female first names in Vietnamese.

Anh) is more likely a Male while **Tú Anh** (Tu Anh) has higher percentage to be a Female. This observation suggests that a simple first name is not enough to robustly classify the genders. Therefore, middle names are necessary in order to rank up the possibility of detecting genders correctly.

C. Preprocessing

After having data, we perform data cleaning by removing extra spaces, extended characters, special characters, duplicate data, We also convert the names into lower case in order to strengthen the integrity for our data and of course as evaluated in the Data Analysis section we keep full name to go to predict gender.

IV. GENDER-PREDICTION MODELS

D. Naive Bayes

Naive Bayes is a probabilistic classifier based on the Bayes'theorem [19]. It is an algorithm belonging to a class of statistical algorithms that can predict the probability of a data element belonging to a class.

Multinomial Naive Bayes algorithm is a probabilistic learning method that is mostly used in Natural Language Processing (NLP). The algorithm is based on the Bayes theorem and predicts the tag of a text such as a piece of email or newspaper article. It calculates the probability of each tag for a given sample and then gives the tag with the highest probability as output.

As in this gender prediction problem, the Multinomial Naive Bayes (MNB) determines the frequency of occurrence of specific words in Vietnamese names that are more likely to

be present in a gender. For example, Thi is used for naming females while Van is usually for male names. Therefore, the MNB model is a reasonable choice for this classification task.

E. Logistic Regression

Logistic Regression is a "Supervised machine learning" algorithm that can be used to model the probability of a certain class or event. It is used when the data is linearly separable and the outcome is binary or dichotomous in nature.

F. Support Vector Machine

Support vector machine was one of the most popular and used algorithms in machine learning before artificial neural networks returned to deep learning models. It provides a superior method for text classification using kernel function to handle nonlinear spaces. The Vietnamese names have some words that both frequently appear in two labels, thus, SVM algorithm in this scenario is able to provide a better results on predictions.

G. Decision Tree

The decision tree algorithm was proposed by Quinlan [20] and it works for both classification [21] and regression problems. As our problem is a classification problem and each word in a name has a weighted influence towards one gender, this algorithm is well-suited.

H. Random Forest

The random forest algorithm decreases the risk of overfitting by building multiple trees and drawing conclusions with replacement [22]. The purpose of using this model in the problem is to compare it with the Decison model. Maruf et al. [19] described a new method on Random Forest and Feature Selection (FS) for text classification and achieved a macro-F1 score 73% higher than normal FS algorithm.

I. KNN

The KNN algorithm was proposed by Hodges et al. [23], and is used for both classification and regression [24]. This is a simple and intuitive model, yet highly effective due to its non-parametric nature. In this paper, we executed a K-nearest neighbors algorithm for binary classification.

V. EXPERIMENTS

In this Section, we describe the setup process testing using Naive Bayes, Logistic Regression, Support Vector Machine, K-Nearest Neighbor, Decision Tree, Random Fores as well as check the output of each technique by compare f1-score results then use Votingclassifier to improve model performance.

J. Data Prepration

First of all, we start by separating the data into a female-gender dataset and a male-gender dataset. Next, we randomly divide the above data set into two subsets: training and testing (performed on each dataset separately). We separate the corpus in the proportion of 70% and 30%, for the training set and test set respectively. Then, we merge two subsets by their gender label, for example, the train set for males is combined with a train set for females. After the data has been split up, the names (features) are needed to be converted into individual vectors. In our experiments, we evaluate Count Vector and TF-IDF features for vectorizing and tokenizing the data.

K. Result Analysis

We report the final results of six machine learning algorithms for both token methods in Table 1. Visual results specify that Support Vector Machine produces the best results with the counting vector method and TF-IDF. Table 2 also shows that measures such as precision, recall, and f1 are higher than the remaining 5 algorithms. In addition, we conducted experiments to remove components in names to understand their impact on gender prediction of Vietnamese people. In this test we use 3 models: Naive Bayes, Logistics Regression, Support Vector Machine for having the best output in table 1. then using voting classifier to improve model performance, table 3 is results when we run test on seven name combinations. The results show that the concatenation of the surname and the middle and first names gives the highest model performance with 95.57%

Table 1. F1-score of 6 machine learning techniques

Model	CountVectorizer			TfidfVectorizer		
	Female	Male	Average	Female	Male	Average
Naive Bayes	94.07	95.12	94.59	92.87	94.17	93.52
Support Vector Machine	**94.66**	**95.79**	**95.22**	**94.45**	**95.6**	**95.03**
Logistic Regression	94.54	95.67	95.11	94.25	95.46	94.85
K-Nearest Neighbors	91.45	93.41	92.43	89.68	91.41	90.54
Decision Tree	91.72	93.31	92.51	91.22	92.82	92.02
Random Forest	93.74	95.01	94.38	93.90	95.07	94.48

Table 2. Experimental performance of examined models on our dataset.

Model	Precision	Recall	F1-score
Naive Bayes	94.59	94.59	94.59
Support Vector Machine	**95.49**	**95.04**	**95.22**
Logistic Regression	95.34	94.94	95.11
K-Nearest Neighbors	92.89	92.17	92.43
Decision Tree	92.61	92.44	92.51
Random Forest	94.56	94.24	94.38

Table 3. F1-score of different combinations of name components.

Name Components	Naïve Bayes			Logistic Regression			Support Vector Machine			Votingclassifer
	Female	Male	Average	Female	Male	Average	Female	Male	Average	
Family Name (FaN)	21.68	70.07	45.88	21.68	70.07	45.88	23.76	70.00	46.88	45.88
Middle Name (MN)	88.75	91.13	89.94	89.68	92.45	91.07	89.53	92.29	90.91	91.05
First Name(FiN)	85.12	88.40	86.76	85.28	88.31	86.8	85.69	87.95	86.82	86.82
FaN + MN	88.47	90.84	89.66	89.42	92.29	90.85	89.09	92.04	90.56	90.79
FaN + FiN	85.04	88.03	86.54	85.30	88.36	86.83	85.06	87.74	86.40	86.61
MN + FiN	**94.17**	**95.22**	**94.70**	**94.57**	**95.69**	**95.13**	94.58	95.68	95.13	95.31
FaN + MN + Fi	94.07	95.12	94.59	94.54	95.67	95.11	**94.66**	**95.79**	**95.22**	**95.57**

VI. CONCLUSION AND FUTURE WORK

In this paper, we have presented several approaches to gender the task of classifying according to Vietnamese names. Our experiments are performed on six traditional machine learning techniques and then use Votingclassifier to improve the model performance for our dataset results, this rate is up to 95%. Therefore, we also create a simple web local based on this Voting Classifier model.

REFERENCES

[1] L. A. Alexandre, "Gender recognition: a multiscale decision fusion approach," Pattern Recognition Letters, vol. 31, no. 11, pp. 1422–1427, 2010.

[2] H. H. Saeed, M. H. Ashraf, F. Kamiran, A. Karim, and T. Calders, "Roman Urdu toxic comment classification," Language Resources and Evaluation, vol. 55, 2021.

[3] K. Khan, M. Attique, I. Syed, and A. Gul, "Automatic gender classification through face segmentation," Symmetry, vol. 11, no. 6, p. 770, 2019.

[4] A. Swaminathan, M. Chaba, D. K. Sharma, and Y. Chaba, "Gender classification using facial embeddings: a novel approach," Procedia Computer Science, vol. 167, pp. 2634–2642, 2020.

[5] L. Cai, J. Zhu, H. Zeng, J. Chen, C. Cai, and K.-K. Ma, "HOGassisted deep feature learning for pedestrian gender recognition," Journal of the Franklin Institute, vol. 355, no. 4, pp. 1991–2008, 2018.

[6] L. M. Lopez-Santamaria, J. C. Gomez, D. L. Almanza-Ojeda, and M. A. Ibarra-Manzano, "Age and gender identification in unbalanced social media," in Proceedings of the 2019 International Conference on Electronics, Communications and Computers (CONIELECOMP), pp. 74–80, IEEE, Cholula, Mexico, March 2019.

[7] P. I. Kiratsa, G. K. Sidiropoulos, E. V. Badeka, C. I. Papadopoulou, A. P. Nikolaou, and G. A. Papakostas, "Gender identification through facebook data analysis using machine learning techniques," in Proceedings of the Tewnty Second Pan-Hellenic Conference on Informatics - PCI '18, pp. 117–120, ACM Press, Athens, Greece, December 2018.

[8] A. Orita, "What is your "formal" name?: situational usage of surnames in Japanese social life," in Proceedings of the 4th Conference on Gender & IT - GenderIT '18, pp. 161–163, ACM Press, Heilbronn, Germany, May 2018.

[9] M. T. Vi, L. T. Dat, V. T. Hoang, and T. A. Nguyen-(i, "Unsupervised gender prediction based on deep facial features," in Proceedings of the 2021 Zooming Innovation in Consumer Technologies Conference (ZINC), pp. 1–4, Novi Sad, Serbia, May 2021.

[10] H. Nguyen-Quoc and V. T. Hoang, "Gender recognition based on ear images: a comparative experimental study," in Proceedings of the 2020 :ird International Seminar on Research of Information Technology and Intelligent Systems (ISRITI), pp. 451–456, Yogyakarta, Indonesia, December 2020.

[11] S. Kruger and B. Hermann, "Can an online service predict gender? On the state-of-the-art in gender identification from texts," in Proceedings of the 2019 IEEE/ACM 2nd International Workshop on Gender Equality in Software Engineering (GE), pp. 13–16, IEEE, Montreal, QC, Canada, May 2019.

[12] J. Jia and Q. Zhao, "Gender prediction based on Chinese name," in Natural Language Processing and Chinese Computing, J. Tang, M. Y. Kan, D. Zhao, S. Li, and H. Zan, Eds., vol. 11839, pp. 676–683, Springer International Publishing, Cham, New York, NY, USA, 2019.

[13] J. Mueller and G. Stumme, "Gender inference using statistical name characteristics in twitter," in Proceedings of the Third Multidisciplinary International Social Networks Conference on SocialInformatics 2016, Data Science 2016 - MISNC, SI, DS 2016, pp. 1–8, ACM Press, Union, NJ, USA, August 2016.

[14] A. A. Septiandri, "Predicting the Gender of Indonesian Names," 2017, https://arxiv.org/abs/1707.07129.

[15] Huy Quoc To et al, "Gender Prediction Based on Vietnamese Names with Machine Learning Techniques", 2021,

[16] Thien Ho Huong, "A Computational Linguistic Approach for Gender Prediction Based on Vietnamese Names" (2022)

[17] Trung Hoa Le. 2002. Ho va Ten nguoi Viet Nam

[18] Cac Dan Toc O Viet Nam Cach Dung Ho Va Dat Ten - Nguyen Khoi (2006)

[19] Z. Xue, J. Wei, and W. Guo, "A real-time naive bayes classifier accelerator on fpga," IEEE Access, vol. 8, pp. 40 755–40 766, 2020.

[20] J. R. Quinlan, "Induction of decision trees," Machine learning, vol. 1, no. 1, pp. 81–106, 1986.

[21] R. Potharst and J. C. Bioch, "Decision trees for ordinal classification," Intelligent Data Analysis, vol. 4, no. 2, pp. 97–111, 2000.

[22] I. Ahmad, M. Basheri, M. J. Iqbal, and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme

learning machine for intrusion detection," IEEE access, vol. 6, pp. 33 789–33 795, 2018.

[23] E. Fix and J. L. Hodges, "Discriminatory analysis. nonparametric discrimination: Consistency properties," International Statistical Review/Revue Internationale de Statistique, vol. 57, no. 3, pp. 238–247, 1989.

[24] F. Zhao and Q. Tang, "A knn learning algorithm for collusion-resistant spectrum auction in small cell networks," IEEE Access, vol. 6, pp. 45 796–45 803, 2018.