Research Article

A Computational Linguistic Approach for Gender Prediction Based on Vietnamese Names

Thien Ho Huong [ID]**, Kiet Tran-Trung, and Vinh Truong Hoang** [ID]

Ho Chi Minh City Open University, Ho Chi Minh City, Vietnam

Correspondence should be addressed to Vinh Truong Hoang; vinh.th@ou.edu.vn

Gender prediction is extensively studied in recent years since it is widely applied in many fields. Several factors have been investigated to determine a gender of male or female through facial images, voice, gait, finger print, etc. In this study, we present a machine learning approach for gender determination based on Vietnamese names. A model based on N-gram for the full name, combining its own middle name feature based on the specificity of Vietnamese language, is proposed. The experimental evaluation of gender prediction tasks is applied on GenderVN1.0 dataset (with 3 million Vietnamese names) that achieves 90.9% of accuracy.

1. Introduction

Gender prediction is one of the most important problems in machine learning with various applications for marketing, advertising, e-commerce, security, and human behavior [1, 2]. There 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 name has been extensively focused by a number of authors [12–14].

Gender identification based on name 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, chat-bot, and machine translation [15, 16]. In marketing, identifying the exact gender of a customer allows to propose products to the right audience. For example, users will reduce the time in systems while they need to fill their information. In order to protect and avoid fraud in declaration, the gender prediction is really helpful for different systems such as customer management systems, e-commerce, and social websites. Several online API services are proposed to predict gender based on English names such as Gender-API (https:// gender-api.com/) and Genderize (https://genderize.io/). Therefore, this gender identity will be really necessary for the

systems question and answering, chat-bot, or machine translation. It enables the interaction with customers becoming naturally as human beings.

Gender determination based on text is first investigated for authorship identification. For example, Cheng et al. [17] predicted the gender of content-free text's author by proposing 545 psycho-linguistic and gender-based features. These output features are then fed to different classifiers (Adaboost decision tree, Na¨ives Bayes, and SVM) for gender prediction. Different languages are considered for gender determination based on full name by machine learning methods, such as Russian [18,19], Indonesian [14], Chinese [12], Arabic [20], English [21–23], Kannada [24], Brazilian [25], Thai [26], and Bengali [27].

Several works of gender identification have been proposed such as using a dictionary of names [18], rule-based [28], and deep learning [14]. Some studies use machine learning models [7, 24] associated with the extraction of the full name feature. Since languages are different, extracting these features also depend on the characteristics of each language. In addition to using a dictionary of full names, Panchenko and Teterin [18] incorporates word ending features and characters N-gram for the Russian language. The results were quite satisfactory, with an accuracy of 96%. Tang et al. [29] introduce an approach for gender inference

and behavior on Facebook's username. They investigated nearly 1.7 million users in New York city by combining various attribute pairs. The experimental results achieve 95.2% of accuracy. Mueller and Stumme [13] use machine learning methods for gender identification based on statistics of name properties. The author has built a classification model called NamChar based on the features selected by the author such as number of syllables, number of consonants, number of vowels, vowel brightness, and ending character. The study proves that the NamChar model is more efficient than the use of the gender score with an accuracy of 70.9%, and this model is particularly efficient for unknown names. Jia and Zhao [12] have focused on simple Chinese features, combined with phonetic information (Pinyin and Hanzi). Then, these features are combined with the Chinese word embedding based on the pretrained BERT model. The results of this work achieved an accuracy of 93.5%.

The feature extraction using N-gram is quite common in text mining and natural language processing (NLP) tasks. In the gender identification problem, the use of N-gram has also been applied in many projects [14, 30, 31]. The authors have used the N-gram feature as basic features of the machine learning model. In this work, we also focused on using N-gram for feature extraction, combining with the term frequency (TF) features for Vietnamese middle name.

The rest of this paper is organized as follows. Section 2 presents the methodology for feature extraction of Vietnamese name. Section 3 describes datasets and our experiments in detail. Finally, the conclusion is discussed in section 4.

2. Methodology

2.1. Brief introduction of Vietnamese Name. Vietnam totally has 54 ethnic groups where the Kinh is the majority with nearly 86% of the population [32]. The people's names of each ethnic group are not the same due to their own language [33]. In this work, we focus on investigating the full name of the Kinh people. Vietnamese language is tonal. Therefore, names with the same spelling but with different tones represent different meanings. These phenomena can confuse people when the accent marks are dropped [34]. Vietnamese personal names generally consist of three parts as follows:

(1) A family name or last name

(2) One or more middle name(s) (one of which may be taken from the mother's family name)

(3) A given name or first name

Most Vietnamese have one middle name, but it is quite possible to have two or more of them or to have no middle name at all. A Vietnamese full name must be arranged in that order. This rule is officially used for administration and daily life. Moreover, a woman's name will be changed after marriage in contrast with other countries such as the UK, US, and China. Let us take an example of a full name: Ngô Đăng Hưng. In this case, Ngô is the family name or what we call the last name. Đăng is the individual's middle name, and Hưng is the given or first name. The given name, which appears last, is the name used to address someone, preceded by the appropriate title of Ngô Đăng Hưng, for example. In formal usage, he is referred to by his given name ("Mr. Hưng"), not by his family name ("Mr. Ngô"). To better understand the structure of this name, we present and analyze several names in Table 1.

It is estimated that there are around 100 family names in common use, but some are far more common than others. The name Nguyn is estimated to be used by almost 40% of the Vietnamese population [35]. Naming of Vietnamese people is also rich and varied. Names can be given with deep connotations such as Nguyễn Hòa Bình and Trần Hạnh Phúc (showing peace, happiness); or it can simply rhyme with the parent's name or even just named after a flower (Phạm Cúc (daisy), Xuân Lan (orchid), Thu Hồng (rose)...). Although Vietnamese names are not restricted, there are several things that need to be limited, such as avoiding names of someone in relatives (both the previous generation and the next generation), not naming men for women and vice versa to make it easy to distinguish, or not too giving a bad name, as well as superstitious, fanatic, and superficial, such as a champion and hero.

Additionally, there are some names that can only be used to designate people's gender as male and female. In these cases, gender will be clearly identified. However, there are a number of names used for both male and female names. So, if we only use the names only, sometimes we cannot distinguish between the gender of male or female. When these names are combined with middle names, the gender can be easily identified. For example the name "An" can be used to denote both male and female genders or combined with the middle name, "Thành An," "Thiên Ân" reward for female gender, and "Bình An," "Mạnh Ân" reflect for male gender. In addition to combining names with middle names in gender identification, in middle names there are also words indicating male or female gender. Based on this appearance, we can also identify gender through middle name. For example, in middle name, where the words "thị," "thúy," and "thúy" appear, the gender is female. The middle name appears in the word related to "literary" and "strong," and the gender is male. However, the formula for naming "văn for man" and "thị for women" seems to have changed a little nowadays.

2.2. Classifiers. Previous works [36–39] showed a comparative study of classifiers to evaluate the performance of NLP tasks. So, we apply three well-known classifiers such as logistics regression, Naïve Bayes, and random forest.

(i) The Naïve Bayes classifier is based on the probability theory of Bayes's theorem. Therefore, this classifier relies on probability and statistics calculation to make predictions or classifications of data which are often used to solve problems with text classification, spam filtering, and emotional recognition [40, 41].

(ii) Decision tree is a structured hierarchy used to classify objects based on a series of rules. The

TABLE 1: Illustrate the structure of Vietnamese names.

No	Full name	Length	Surname	Middle name	First name
1	Phạm Hùng	2 words	Phạm	—	Hùng
2	Nguyễn Thành, Trung	3 words	Nguyễn	Thành	Trung
3	Lê Thị Thùy, Dương	4 words	Lê	Thị Thùy	Dương
4	Đinh, Hoàng Bảo Trân	5 words	Đinh	Nguyễn, Hoàng Bảo	Trân
5	Nguyễn Lê, Hoàng Đinh Thảo Mai	6 words	Nguyễn	Lê, Hoàng Đinh Thảo	Mai

results of the decision tree model are results based on the questions. It can be applied to both regression problems and classification problems. The algorithms ID3, C4.5, J48, and CART (classification and regression trees) are algorithms that are extended from the decision tree algorithm [42]. Random forest algorithm is an extension of bagging approaches; it also combines many decision trees into a single model. Each decision tree of the forest is built from a random subset of features and only has accessed a random set of the training data points.

(iii) Logictics regression analyzes the relationship between the dependent variable and one or more independent variables based on the probability by using the logistic/sigmoid. This model is like the linear regression model; they also estimate the coefficients from training dataset to minimize the errors which are different between the real output and the predicted output. This classifier is also applied to predict English names in [23].

2.3. Related Work. All names will be tokenized to individual words in order to characterize by a feature vector. The term frequency (*TF*) is a frequency of word that appears and the number of times of its appearance in a document, divided by the total number of words in that document [40], where t is a word in a document, $f(t, d)$ is a frequency of word occurring in that document, and T is a total number of words from this document:

$$\mathrm{TF}(t) = \frac{f(t, d)}{T}. \tag{1}$$

In the computational linguistics domain, an N-gram is a contiguous sequence of N items from a given sequence of text. These items can be considered as syllables, letters, words, or base pairs according. The N-grams typically are extracted from a text corpus. We can use different values of N and obtain the corresponding size as follows: size 1 is referred to as a "unigram," size 2 is for "bigram," and size 3 is a "trigram." Larger sizes are sometimes referred to by the value of N, e.g., "four-gram" and "five-gram".

Based on the above characteristics, we incorporate feature extraction by the N-gram method for the Vietnamese full name. Figure 1 illustrates a scheme of gender determination based on Vietnamese names. The feature extraction stage is applied by *TF* extraction with different strategies for middle name and given name. The three classifiers are considered to predict gender.

2.4. Evaluation Metric. In order to evaluate the effectiveness of the proposed approach, an accuracy metric will be used. This metric is the ratio of the model's correct testing over test data. It depends on the four parameters *TP (True Positive), TN (True Negative), FP (False Positive),* and *FN (False Negative)* and is calculated by the following formula:

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

3. Experiment and Results

We built a Vietnamese full name dataset for gender determination tasks, namely, GenderVN1.0. By collecting a list of students from high school to undergraduate, we obtain nearly 3 millions names for both genders. To our knowledge, this is the first large scale for this task. A data cleaning process is applied to remove duplicate names with the same gender annotation. The characteristic of GenderVN1.0 dataset is illustrated in Table 2. The dataset generated during this study is available from the corresponding author upon request.

We decompose the GenderVN1.0 dataset into two disjoint subsets such as training and testing set by a ratio 60 : 40. The feature extraction is applied on training data and feeding into classifiers to build models. A typical Vietnamese name consists of three words including one for family name, one for middle name, and one first name. Here, we apply two strategies to extract features from considered Vietnamese names such as a full name and without using family name. Vietnamese names usually consist of 3 words for male and 4 words for female. Three values of $N \in \{1, 2, 3\}$ are considered to extract N-gram features. Three values of $N \in \{1, 2, 3\}$ are considered to extract N-gram features. For each type of feature, we apply three classifiers to predict gender independently. The N-grams are basically a set of co-occurring words within a given window and when computing the N-grams by moving one word forward. In this problem, for a common female name, "Nguyễn Thị Bạch Tuyết." If we extract features by using bigrams ($N = 2$), then the N-grams should be

(i) Nguyễn Thị

(ii) Thị Bạch

(iii) Bạch Tuyết

Table 3 presents the prediction results on the testing set. We observe that the best accuracy is obtained by using 1-gram extraction with logistic regression for both strategies. Obviously, a family name cannot allow us to predict the
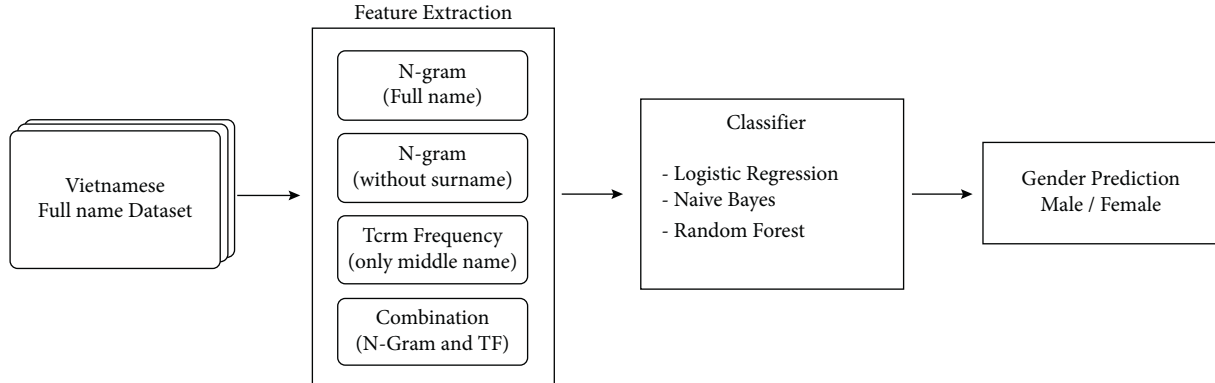
Figure 1: Scheme of gender prediction based on Vietnamese name.

Table 2: The detailed information of the GenderVN1.0 dataset.

	Original dataset	Filter duplicated names
Male	1,499,980	582,997
Female	1,530,988	564,654
Total of names	3,031,020	1,147,651
Total of words	10,272,108	4,015,734

Table 3: Gender prediction of N-gram by using full name and without family name on the GenderVN1.0 dataset.

N-gram	Using full name			Without family name		
	Logistics regression	Näive Bayes	Random forest	Logistics regression	Näive Bayes	Random forest
1-gram	90.4	90.0	88.2	**90.9**	90.3	90.5
2-gram	70.3	70.0	70.4	68.4	68.4	68.4
3-gram	54.8	54.8	54.8	53.0	53.0	53.0

Bold shows best accuracy obtained.

Table 4: Using middle name for gender prediction by 1-gram and/or *TF* extraction.

Classifier	Middle name term frequency	Combination (1-gram and term frequency)
Logistics regression	76.08	76.12
Näive Bayes	76.07	76.11
Random forest	76.12	76.11

gender of a person. The best accuracy is achieved at 90.9% in cases of using middle name and first name.

As we mentioned in Section 2.1, the middle name of Vietnamese people plays a major role for gender determination. Table 4 illustrates the prediction results by using only the middle name for feature extraction. We observe that the prediction results achieve around 76.0% for any considered classifiers or feature extraction methods. This result confirms again that the middle name of Vietnamese people allows predicting their gender accurately.

4. Conclusion

In this paper, we presented a method for gender prediction for Vietnamese names. We provided a first large-scale GenderVN1.0 dataset with more 3 millions Vietnamese corresponding to an annotated gender. The experimental results show the effectiveness of the proposed approach by achieving 90.9% for gender prediction on GenderVN1.0 dataset. The experiments also demonstrated that the sole middle name acts as a major role for gender prediction with an accuracy obtained by 76.1%. However, there are several limitations of the proposed approach. It only can recognize well the gender-based name of Kinh people. Secondly, it cannot predict the gender of Vietnamese name combined with foreigner words.

The first future of this work is now extended to improve the proposed approach by incorporating feature selection to remove irrelevant and combining with deep features. The second perspective is to represent a compact features descriptor for extracting Vietnamese names.

Data Availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

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, "HOG-assisted 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-Thi, "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 Third 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 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] R. Steinberger, "A survey of methods to ease the development of highly multilingual text mining applications," *Language Resources and Evaluation*, vol. 46, no. 2, pp. 155–176, 2012.

[16] H. Duong and V. T. Hoang, "Question answering based on ensemble classifier for university enrolment advising," in *Proceedings of the 2019 11th International Conference on Knowledge and Smart Technology*, pp. 35–39, KST), Phuket, Thailand, January 2019.

[17] N. Cheng, R. Chandramouli, and K. P. Subbalakshmi, "Author gender identification from text," *Digital Investigation*, vol. 8, no. 1, pp. 78–88, 2011.

[18] A. Panchenko and A. Teterin, "Detecting gender by full name: experiments with the Russian language," in *Communications in Computer and Information Science*, D. I. Ignatov, M. Y. Khachay, A. Panchenko, N. Konstantinova, and R. E. Yavorsky, Eds., vol. 436, pp. 169–182, Springer International Publishing, Cham, New York, NY, USA, 2014.

[19] A. Sboev, I. Moloshnikov, D. Gudovskikh, A. Selivanov, R. Rybka, and T. Litvinova, "Automatic gender identification of author of Russian text by machine learning and neural net algorithms in case of gender deception," *Procedia Computer Science*, vol. 123, pp. 417–423, 2018.

[20] S. A. Alanazi, "Toward identifying features for automatic gender detection: a corpus creation and analysis," *IEEE Access*, vol. 7, Article ID 111931, 2019.

[21] G. Ciccone, A. Sultan, L. Laporte, and M. Granitzer, *Stacked Gender Prediction from Tweet Texts and Images*, p. 11, 2020.

[22] L. Santamaría and H. Mihaljević, "Comparison and benchmark of name-to-gender inference services," *PeerJ Computer Science*, vol. 4, p. e156, 2018.

[23] 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, pp. 1–27, 2021.

[24] A. N. Myna, L. R. Swaroop, S. Hegde, U. Sourabh, and G. S. Rakshith Gowda, "Gender identification for Kannada names," in *Proceedings of the 2019 First International Conference on Advances in Information Technology (ICAIT)*, pp. 421–426, Chikmagalur, India, July 2019.

[25] R. C. B. Rego and V. M. L. Silva, "Predicting Gender of Brazilian Names Using Deep Learning," 2021, https://arxiv.org/abs/2106.10156.

[26] S. Yuenyong, S. Sinthupinyo, and S. Sinthupinyo, "Gender classification of Thai facebook usernames," *International Journal of Machine Learning and Computing*, vol. 10, no. 5, pp. 618–623, 2020.

[27] J. F. Ani, M. Islam, N. J. Ria, S. Akter, and A. K. Mohammad Masum, "Estimating gender based on Bengali conventional full name with various machine learning techniques," in *Proceedings of the 2021 Twelveth International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pp. 1–6, Kharagpur, India, July 2021.

[28] H. Liu and M. Cocea, "Fuzzy rule based systems for gender classification from blog data," in *Proceedings of the 2018 Tenth*

International Conference on Advanced Computational Intelligence (ICACI), pp. 79–84, IEEE, Xiamen, China, March 2018.

[29] C. Tang, K. Ross, N. Saxena, and R. Chen, "What's in a name: a study of names, gender inference, and gender behavior in facebook," in *Database Systems for Adanced Applications*, J. Xu, G. Yu, S. Zhou, and R. Unland, Eds., vol. 6637, pp. 344–356, Springer Berlin Heidelberg, Berlin, Germany, 2011.

[30] D. Ali, M. Muhammad, N. Akhtar, N. Salamat, H. Asmat, and A. Firdous, "Gender prediction for expert finding task," *Ijacsa*, vol. 7, no. 5, 2016.

[31] S. Daneshvar and D. Inkpen, *Gender identification in twitter using n-grams and lsa: notebook for pan at clef 2018*, CLEF, 2018.

[32] B. Baulch, K. T. T. Chuyen, D. Haughton, and J. Haughton, *Ethnic Minority Development in Vietnam: A Socioeconomic Perspective*, The World Bank, Wasington, DC, USA, 2002.

[33] P. H. Khương, *Đôi Nét V Đc Đim H Tên Ca Ngi Trung Quc Và Ngi Vit Nam*, vol. 7, no. 237, Ngôn Ngữ & Đời Sống, 2015.

[34] P. T. L. T. Hoa, "Họ và tên người Việt Nam," NXB Khoa Học Xã Hội, Beijing, China, 2005.

[35] H. M. Nguyen, B. H. Tran, T. D. Vuong, and Q. Vuong, *Colloquial Vietnamese: The Complete Course for Beginners*, Routledge, Oxfordshire, UK, 2012.

[36] F. Hemmatian and M. K. Sohrabi, "A Survey on Classification Techniques for Opinion Mining and Sentiment Analysis," *Artificial Intelligence Review*, vol. 52, 2017.

[37] K. Shuang, Z. Zhang, H. Guo, and J. Loo, "A sentiment information Collector-Extractor architecture based neural network for sentiment analysis," *Information Sciences*, vol. 467, pp. 549–558, 2018.

[38] A. A. Farisi, Y. Sibaroni, and S. A. Faraby, "Sentiment analysis on hotel reviews using Multinomial Naïve Bayes classifier," *Journal of Physics: Conference Series*, vol. 1192, Article ID 12024, 2019.

[39] H. T. Duong and V. Truong Hoang, "A survey on the multiple classifier for New benchmark dataset of Vietnamese news classification," in *Proceedings of the 2019 Eleventh International Conference on Knowledge and Smart Technology (KST)*, pp. 23–28, IEEE, Phuket, Thailand, January 2019.

[40] R. Ahuja, A. Chug, S. Kohli, S. Gupta, and P. Ahuja, "The impact of features extraction on the sentiment analysis," *Procedia Computer Science*, vol. 152, pp. 341–348, 2019.

[41] R. Othman, Y. Abdelsadek, K. Chelghoum, I. Kacem, and R. Faiz, "Improving sentiment analysis in twitter using sentiment specific word embeddings," in *Proceedings of the 2019 Tenth IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS)*, pp. 854–858, IEEE, Metz, France, (September 2019.

[42] C. Leistner, A. Saffari, J. Santner, and H. Bischof, "Semi-supervised random forests," in *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision*, pp. 506–513, IEEE, Kyoto, Japan, October 2009.