

# Gender Classification of Thai Facebook Usernames

Supitcha Yuenyong and Sukree Sinthupinyo

**Abstract**—This paper presents an application of machine learning to classify Facebook users' gender based on their username alone. User profile information on social networks is important in many studies, but occasionally no information is publicly available online, such as age or gender. Most studies only use textual information from the web page. Instead, we opted to study gender classification by username, in which the gender is inferred from the users first name and alias name. We focused only on Thai names which may have certain patterns that reveal the owner's gender. A combination of different models is proposed to classify gender based on Thai Facebook usernames. Each model was trained using a supervised learning approach. Furthermore, all the classification results were combined into a final model. Using this method, the model achieved 91.75% level of accuracy.

**Index Terms**—Gender classification, Facebook username, name analysis, social network, machine learning.

## I. INTRODUCTION

Social media is used extensively and has become a part of people's daily lives. It influences the lifestyles of people of all ages in almost every aspect of a person's life [1]. Facebook is the largest and most used social media platform globally and in Thailand [2]. Facebook provides a simple way for users to express their feelings, ideas, and opinions by enabling them to generate their own content. Information published online can be studied to analyze social processes from economics to public health. For example, online data can be used to forecast target groups by marketing teams, opinions about political events can be analyzed, and perspectives on social issues or mental health can be assessed [3]. Yet most social media sites do not require user demographic details. Moreover, it is not necessary for users to reveal their identities, and some users even disguise or conceal their private information to hide their true identities. Consequently, using user information derived from social media for analysis or communication can be inaccurate and can fail to match the target group. The correct gender classification of social media users is therefore important, since it can help solve issues derived from limited or inaccurate social user information and maximize benefits. According previous studies, the existing methods for classifying gender are often analyzed by text, but it is not always possible to find messages to analyze. The name of users is another way of classifying gender, especially in the Thai language.

Manuscript received November 14, 2019; revised May 22, 2020.

The authors are with the Department of Computer Engineering, Chulalongkorn University, 254 Phayathai Rd., Phatumwan Bangkok, 10330 Thailand. (e-mail: 6170973921@student.chula.ac.th, sukree.s@chula.ac.th).

In Thai, female and male names have different characteristics. For example, "พัชราภา" (Patcharapa), "วิชุลดา" (Wichulada), and "พรพรรณ" (Pornpan) are female names, while "สุชาติ" (Suchart), "สมเกียรติ" (Somkiat), and "ประวุฒิ" (Prawut) are male names. There are also names which can be used by either gender, such as "วิรัตน์" (Wirut), "สุวรรณ" (Suwan), and "สมพร" (Somporn). Further challenges were caused by many social media users utilizing an alias name rather than their actual name.

For the aforementioned reasons, we decided to study and solve such problems using machine learning techniques to classify gender based on Thai Facebook usernames. This was done using both first names and alias names and analyzing the features extracted from username-derived textual information.

## II. RELATED WORK

In recent years, several studies have proposed machine learning approaches to extract the demographic of social media users using text, names, images, location, or profile colors to classify gender, age, personality, education, marital status, ethnicity, geographic location, language, and race [4]-[11].

For instance, Schwartz, H.A. *et al.* [4] used post text and comment text to determine personality, gender, and age classification on Facebook. Alowibdi, J.S., U.A. Buy, and P. Yu [5] used first name, username, and profile colors (i.e., background color, text color, link color, sidebar fill color, and sidebar border-color) for gender classification on Twitter. Furthermore, Bergsma, S., Dredze, M., Van D.B., Wilson, T., and Yarowsky, D. [6] used first name, last name, and location to classify ethnicity, geographic location, gender, language, and race on Twitter. Additionally, Akbar, R. [7] and Septiandri, A.A. [8] used Indonesian names for gender classification, while Briediene, M. and Kapociute-Dzikiene, J. [9] used Lithuanian texts for gender, age, education, marital status, and personality type classification on Facebook. Moreover, Hirt, R., N. Kuhl, and G. Satzger [10] used German tweets, names, and profile pictures for gender classification on Twitter. Also, Vicente, M., F. Batista, and J.P. Carvalho [11] used username, screen name, description, tweet content, profile picture, and user activities for gender classification on Twitter.

These authors explored a variety of methods for gender classification from username. For example, Alowibdi, J.S., U.A. Buy, and P. Yu [5] ran experiments to compare three different classifiers: Naïve Bayes; Decision Tree; and Naïve Bayes Decision Tree hybrid. They trained classifiers with the

phoneme-based features set, with word-frequency based features, and with 1-gram through 5-gram features. The best result achieved was 82.5% accuracy in the case of the 3-gram phoneme-based features with Decision Tree classifier for first name, and 75.2% accuracy in the case of the 3-gram phoneme-based features using Decision Tree for usernames. Bergsma, S., Dredze, M., Van D.B., Wilson, T., and Yarowsky, D. [6] ran experiments using Support Vector Machine to classify gender by first name and surname, by making comparisons between five features: Token; character N-gram; cluster; token with N-gram; and all together features. The best result achieved was 90.2% accuracy with all features. Akbar, R. [7] compared Multinomial Naive Bayes with Random Forest to classify gender from Indonesian names using the frequency of characters, last character, and last two characters features. The classifiers yielded an accuracy of around 70% (Multinomial Naive Bayes) and 83% (Random Forest). Further, Septiandri, A.A. [8] classified Indonesian name genders using Character-Level Long-Short Term Memory compared with Naive Bayes, Logistic Regression, and XGBoost using n-grams as the features. The results showed that the best performance of Naive Bayes and Logistic Regression was obtained from 3-gram, and the best performance of XGBoost was from 2-gram. When using Character-Level Long-Short Term Memory techniques, they were able to classify gender more accurately than Logistic Regression was able to. The accuracy percentage rose from 85.28% to 92.25% in the full name case, while using first names only yielded a 90.65% level of accuracy. Vicente, M., F. Batista, and J.P. Carvalho [11] proposed methods based on the combination of different classifiers. They created four distinct classifiers, each of which considered a group of features extracted from four different sources by conducting performance comparisons among Logistic Regression, Multinomial Naive Bayes, Support Vector Machines, and Decision Tree classifier. The final classifier—combining the four previous individual classifiers—achieved the best performance, corresponding to 93.2% accuracy for English and 96.9% accuracy for Portuguese data.

TABLE I: TOTAL NUMBER OF USERNAME DATASETS IN THE STUDY

Gender	First name	Alias name	Total
Female	926	1121	2047
Male	1035	1235	2270

### III. METHODOLOGY

#### A. Dataset

We focused on Thai Facebook usernames collected between January to March 2019 using the Selenium library. Users were selected who had usernames only in Thai characters and had an open gender profile. After all the data was collected, the usernames were manually labelled into two types, namely first name and alias name. Of the 4317 usernames collected, 2047 were classified as female (47.42%) and 2270 were male (52.58%). Moreover, 1961 of them were first names while 2356 were alias names (see Table I). An

example of the dataset is given in Fig. 1.

Username	Gender	First /Alias Name
เทพบุตร ขาดาน	Male	Alias Name
หทัยรัตน์ โพธิ์ทอง	Female	First Name
เจ้าหญิงสายลม	Female	Alias Name
ผู้ชายคนหนึ่งที่แสนธรรมดา	Male	Alias Name
สุรพงษ์ พลใจจริง	Male	First Name
แม่หญิงแสนดี ที่สีหวัจ	Female	Alias Name
เอกพล จิตเจริญสมุทร	Male	First Name
หนุ่มหวาน ชิตชิต	Male	Alias Name
ก้อ กู้ก กู้กโก้	Female	Alias Name
เบญจพร สำเร็จกิจ	Female	First Name
พี่ เสือ	Male	Alias Name
ธีรยุทธ พลวงค์ษา	Male	First Name
นางมาราย ขอบเอาแต่ใจ	Female	Alias Name
กานต์สิณี เข็ดชู	Female	First Name
ไม่หล่อ แต่ ดูไม่เบื่อ	Male	Alias Name

Fig. 1. Dataset example.

#### B. Thai Word Tokenization Features

In the Thai language, words are not separated by a space, and spaces are instead used to denote a new sentence. Furthermore, most alias names can be separated into words, so we selected the PyThaiNLP library [12] using the Maximal Matching method to segment those Thai names. An example of Thai word tokenization is given in Fig. 2.

Username	Word tokenization
แม่พี่สาว เชียงใหม่	[แม่, พี่สาว, , , เชียง, ใหม่, , , เชียง, ใหม่]
แม่ห้องพระเจ้า กับน้องพระคุณ	[แม่, ห้อง, พระเจ้า, , , กับ, น้อง, พระคุณ]
พี่สาวที่ยิ้มแย้ม ตานตมฟ้ามีดอกฟ้าดาว	[พี่สาว, , , ยิ้มแย้ม, ตาน, ตม, ฟ้า, มี, ดอก, , , ฟ้า, ดาว]
เราทั้งคู่ชาย หัวใจสองชื่อ	[เรา, , , กับ, , , หัวใจ, สอง, , , ชื่อ]
แดน วิศุทธิ์ คนหล่อสุด	[แดน, , , วิศุทธิ์, , , คน, หล่อ, , , สุด]
พี่พี่ ไม่มีแม่กับแม่กับพี่พี่	[พี่พี่, , , ไม่, , , กับ, , , แม่, , , กับ, , , พี่พี่]
แมวโหด กระโดดข่วน	[แมว, , , โหด, , , กระโดด, , , ข่วน]
โซกุน เด็กครึ่งชาย	[โซกุน, , , เด็ก, , , ครึ่ง, , , ชาย]
เวลาสมัคร จะโหดและน่ากลัวมาก	[เวลา, , , สมัคร, , , จะ, , , โหด, , , และ, , , น่ากลัว, , , มาก]
เอ็มแอนด์เอ็ม ของพี่แดง	[เอ็มแอนด์, , , เอ็มแอนด์, , , ของ, , , พี่แดง]
เบบส มาสเตอร์ ไทยแลนด์	[เบบส, , , มาสเตอร์, , , ไทยแลนด์]
จะหายไปไหน ไม่ให้คิดถึง	[จะ, , , หายใจ, , , ยังไม่, , , ไม่, , , ให้, , , คิดถึง]
สิ่ง สบมาลีฟ้าหวิจขาว	[สิ่ง, , , สบ, , , มาลีฟ้า, , , หวิจขาว]
ครูป๊อป เจ้าชายสายลม	[ครู, , , ป๊อป, , , เจ้าชาย, , , สายลม]
ไอ้ ขำมันสูทฝรั่ง เสื้อสุพรรณ	[ไอ้, , , ขำ, , , มัน, , , สูทฝรั่ง, , , เสื้อ, , , สุพรรณ]

Fig. 2. Thai word tokenization example.

Fig. 3 presents words most highly distinguishing female and male alias names. Female word tokenization features are shown in pink, while male word tokenization features at in the figure in blue. The size of the word indicates the relative usage frequency, with larger words indicating more frequently used words.

The word tokenization analysis results show that each gender uses self-describing words. For example, females used "น้อง" (sister), "แม่" (mother), and "หญิง" (girl), while males used "พี่" (brother), "หนุ่ม" (boy), and "เสือ" (tiger).

#### C. Thai Speech Classification Features

In the Thai language, first and alias names use different parts of speech. First names often include nouns but do not include certain word types. For example, Thai first names do not include conjunctions such as "และ" (and), "หรือ" (or), and "แต่" (but), negators such as "ไม่" (no), and "ห้าม" (don't), determiners such as "นี้" (this) and "นั้น" (that).

Nonetheless, these may be found in the alias name. Virach S. *et al.* [13] classified Thai speech parts into 14 categories (noun, pronoun, verb, auxiliary, determiner, adverb, classifier, conjunction, preposition, injection, prefix, ending, negator, and punctuation) and divided them into 47 subcategories. In the present study, the PyThaiNLP library was used to extract speech parts with word tokenization. An example of Thai speech part classification is given in Fig. 4.



Fig. 3. Word clouds for alias name word tokenization. (female on top and male below).

Username	First /Alias Name	Part of speech
อุบลพร ดวงสุข	First Name	[[สุบพร, 'NPRP'], (ดวง, 'NCMN'), (สุข, 'NCMN')]
เนตรชนก วาณิชโส	First Name	[[เนตร, 'NCMN'], (ชนก, 'NCMN'), (วาณิช, 'NCMN'), (โส, 'NCMN'), (โส, 'VATT')]
นภัสศักดิ์ ทองสุวรรณ	First Name	[[นภ, 'NTTL'], (ณ, 'NPRP'), (ศักดิ์, 'NPRP'), (ทอง, 'DDBO'), (สุวรรณ, 'NPRP')]
ศิริคุณ อุบลเมษ	First Name	[[ศิริ, 'NCMN'], (คุณ, 'NPRP'), (เมษ, 'NCMN')]
เฉลิมชัย เรืองฮอยาง	First Name	[[เฉลิม, 'NCMN'], (ชัย, 'NCMN'), (เรือง, 'VSTA'), (ฮอยาง, 'ADVP')]
เมณฑล อิศริชญสุวรรณ	First Name	[[เมธ, 'NCMN'], (อศ, 'NCMN'), (อศ, 'NCMN'), (อศ, 'NCMN'), (อศ, 'NCMN'), (อศ, 'NCMN')]
อิทธิชัย ธรรมหาสมกุล	First Name	[[อิทธิ, 'NCMN'], (ชัย, 'NCMN'), (ธรรม, 'NCMN'), (หาสม, 'NCMN'), (กุล, 'NPRP')]
เจ้าแม่ผู้ขาย หัวใจทองดี	Alias Name	[[เจ้า, 'PPRS'], (แม่, 'JSBR'), (ผู้, 'VBM'), (ขาย, 'NCMN'), (หัวใจ, 'NCMN'), (ทอง, 'DCNM'), (ดี, 'CNIT')]
กบหรือเธอ ที่เปลี่ยนใจ	Alias Name	[[กบ, 'PPRS'], (หรือ, 'JCRG'), (เธอ, 'PPRS'), (ที่, 'PREL'), (เปลี่ยนใจ, 'VSTA')]
หัวใจ ไม่ซัดซัด	Alias Name	[[หัวใจ, 'NCMN'], (ไม่, 'NEG'), (ซัด, 'NCMN')]
เค สองหน้าสี่	Alias Name	[[เค, 'NCMN'], (สอง, 'DCNM'), (หน้า, 'PUNC'), (สี่, 'DCNM'), (หน้า, 'PUNC'), (สี่, 'NPRP')]
แสบ โจรใจร้าย	Alias Name	[[แสบ, 'NCMN'], (โจร, 'NCMN'), (ใจ, 'VBM'), (โหด, 'PNTR'), (ร้าย, 'VACT')]
ผู้ยิ้ม หัวใจไม่กลมแล้ว	Alias Name	[[ผู้, 'NCMN'], (ยิ้ม, 'VACT'), (หัว, 'PPRE'), (ใจ, 'PNTR'), (ไม่, 'VACT')]
เก็บแพชเชา เมฆาพรชช	Alias Name	[[เก็บ, 'NCMN'], (แพช, 'JSBR'), (ช, 'PPRS'), (เมฆา, 'VACT'), (พรช, 'JSBR'), (ช, 'PPRS')]
ผู้ชายคนน่ารัก ที่แสนธรรมดา	Alias Name	[[ผู้, 'NCMN'], (ชาย, 'CNIT'), (น่ารัก, 'DCNM'), (ที่, 'PREL'), (แสน, 'VACT'), (ธรรมดา, 'VATT')]

Fig. 4. Example of Thai speech part classification.

First name	Frequency character
เนตรชนก	(น: 1, น: 2, พ: 1, ช: 1, ก: 1)
เนตรสุวรรณ	(น: 1, น: 1, ว: 1, ช: 1, พ: 1, ร: 2, น: 1)
เมณฑล	(น: 1, พ: 2, ล: 1, ร: 2)
เฉลิมชัย	(น: 1, ช: 2, ค: 1, น: 1, ค: 1, ช: 1, พ: 1, ร: 1, ค: 1)
นภัสศักดิ์	(น: 1, น: 1, พ: 1, ค: 1, ค: 1, ค: 1, ค: 1, ค: 1, ค: 1)
นภัสชญ	(น: 2, น: 1, พ: 2, ค: 1, ช: 1, น: 1, ค: 1)
ชัยชาญ	(ช: 2, ค: 1, ช: 1, ร: 1, ญ: 1)
วรางคณา	(ว: 1, ร: 1, ง: 2, ค: 1, พ: 1, น: 1)
ชวลิต	(ช: 1, ล: 1, ช: 1, ค: 1, พ: 1, ช: 1, ค: 1)
ศุภนิสิน	(ศ: 1, ค: 1, น: 2, ค: 1, ช: 1, ค: 1, น: 1, ค: 1)
ชัยบุญวีร์	(ช: 1, ค: 1, ญ: 2, ร: 1, ร: 1, ค: 1, ค: 1)
กิตติพงษ์	(ก: 1, ค: 2, พ: 2, พ: 1, ช: 1, ช: 1, ค: 1)
อภิวัฒน์	(อ: 1, พ: 1, ค: 1, น: 2, ค: 1, พ: 1, ค: 1)
พัชร์นภ	(พ: 1, ค: 1, น: 2, ค: 1, น: 1, น: 1, พ: 1, ค: 1)
อุบลศักดิ์	(อ: 1, ค: 1, น: 1, ร: 1, ค: 1, ค: 1, น: 1, ค: 1)

Fig. 5. Example of Thai character frequency.

D. Thai Character Frequency Features

In the Thai language there are three main character types:

Consonants; vowels; and tones. These are located in the upper, middle and lower levels [14]. The frequency of each character was counted only from the first name. An example of Thai character frequency is given in Fig. 5.

The character occurrence was counted from female and male names, as shown in Fig. 6. The bar chart shows that the character 'ณ' occurs more often in female's first names than for males. The same result can be found in 'ญ', 'ร', 'ภ', and 'า'. Meanwhile, the characters 'ช', 'ง', 'ช', 'ต' and 'ิ' appear more frequently in male first names.

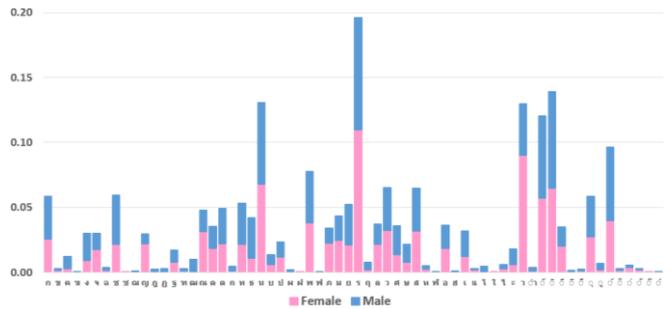


Fig. 6. Character occurrence frequency in first names.

To measure how disorganized character in the first name, we calculated the entropy of their gender for each character, with a high result denoting that character has a highly varied gender, and a low result pointing to the character being used in the names of the just one gender. In particular, entropy is defined as the sum of the probability of each gender, times by the log probability of that same gender. The 10-minimum entropy of character in the first name are 'ช', 'ณ', 'ญ', 'ร', 'ภ', 'า', 'ณ', 'แ', 'โ' and 'ค'.

E. Thai Substring Character Features

Thai male and female first names have different characteristics. For example, names starting with "พร-" are most likely to be female, while names starting with "พล-" are supposed to be male. Meanwhile, names ending with "-วรรณ" are most likely to be female, and names ending with "-วัฒน์" are always for males.

First name	Substring characters
เนตรชนก	(first2-char: 'นต', first3-char: 'นตช', first4-char: 'นตชช', last2-chars: 'นต', last3-chars: 'นตช', last4-chars: 'นตชช')
เนตรสุวรรณ	(first2-char: 'นต', first3-char: 'นตญ', first4-char: 'นตญช', last2-chars: 'นต', last3-chars: 'นตญ', last4-chars: 'นตญช')
เมณฑล	(first2-char: 'เมต', first3-char: 'เมตด', first4-char: 'เมตดล', last2-chars: 'เมต', last3-chars: 'เมตด', last4-chars: 'เมตดล')
เฉลิมชัย	(first2-char: 'เชล', first3-char: 'เชลช', first4-char: 'เชลชช', last2-chars: 'เชล', last3-chars: 'เชลช', last4-chars: 'เชลชช')
นภัสศักดิ์	(first2-char: 'นภ', first3-char: 'นภส', first4-char: 'นภสค', last2-chars: 'นภ', last3-chars: 'นภส', last4-chars: 'นภสค')
นภัสชญ	(first2-char: 'นภ', first3-char: 'นภช', first4-char: 'นภชญ', last2-chars: 'นภ', last3-chars: 'นภชญ', last4-chars: 'นภชญ')
ชัยชาญ	(first2-char: 'ชช', first3-char: 'ชชญ', first4-char: 'ชชญช', last2-chars: 'ชช', last3-chars: 'ชชญ', last4-chars: 'ชชญช')
วรางคณา	(first2-char: 'วร', first3-char: 'วรง', first4-char: 'วรงค', last2-chars: 'วร', last3-chars: 'วรงค', last4-chars: 'วรงคณ')
ชวลิต	(first2-char: 'ชช', first3-char: 'ชชล', first4-char: 'ชชลต', last2-chars: 'ชช', last3-chars: 'ชชล', last4-chars: 'ชชลต')
ศุภนิสิน	(first2-char: 'ศช', first3-char: 'ศชน', first4-char: 'ศชนิ', last2-chars: 'ศช', last3-chars: 'ศชนิ', last4-chars: 'ศชนิส')
ชัยบุญวีร์	(first2-char: 'ชช', first3-char: 'ชชญ', first4-char: 'ชชญว', last2-chars: 'ชช', last3-chars: 'ชชญ', last4-chars: 'ชชญว')
กิตติพงษ์	(first2-char: 'กต', first3-char: 'กตป', first4-char: 'กตปช', last2-chars: 'กต', last3-chars: 'กตป', last4-chars: 'กตปช')
อภิวัฒน์	(first2-char: 'อภ', first3-char: 'อภิ', first4-char: 'อภิม', last2-chars: 'อภ', last3-chars: 'อภิม', last4-chars: 'อภิมต')
พัชร์นภ	(first2-char: 'พช', first3-char: 'พชญ', first4-char: 'พชญน', last2-chars: 'พช', last3-chars: 'พชญ', last4-chars: 'พชญน')
อุบลศักดิ์	(first2-char: 'อภ', first3-char: 'อภช', first4-char: 'อภชค', last2-chars: 'อภ', last3-chars: 'อภช', last4-chars: 'อภชค')

Fig. 7. Example of Thai substring characters.

Features were created from substrings in the first name. Six feature types were used in this experiment: The first two characters; the first three characters; the first four characters;

the last two characters; the last three characters; and the last four characters. An example of Thai substring characters is given in Fig. 7.

Fig. 8 presents the top 10 most substring characters for first names. The female substring characters are shown on the left and the males are shown on the right.

Substring characters were found in the female first names, but not in the male first names, for instance, the last two characters; 'ภา', last two characters; 'ณี', last four characters; 'พย', and last three characters; 'ดา'. Meanwhile, substring characters were found in the male first names but not in the female first names, for example, the last three characters; 'ชัย', last two characters; 'พล', last four characters; 'พงษ์', and first two characters; 'ดี'.

The substring character analysis results show that female and male first names have different characteristics. These distinct properties could therefore be used as features for the model.

1	first2-characters: "ฤ"	1	first2-characters: "ฤ"
2	last2-characters: "ณ"	2	last3-characters: "ชัย"
3	last2-characters: "ดา"	3	last2-characters: "ช"
4	last2-characters: "พร"	4	last4-characters: "กั"
5	last4-characters: "ตณ"	5	last2-characters: "พล"
6	last2-characters: "ภา"	6	first2-characters: "ป"
7	last3-characters: "รรณ"	7	last4-characters: "พงษ์"
8	last4-characters: "กรณ"	8	first2-characters: "ธน"
9	first2-characters: "จ"	9	first2-characters: "สม"
10	first2-characters: "ภ"	10	last3-characters: "พน"

Fig. 8. Top 10 most common first name substring characters. (female on the left and male on the right).

### F. Process Design

In this study, four different feature groups were used, as follows:

- 1) Word tokenization features;
- 2) Speech part classification features;
- 3) Character frequency features;
- 4) Substring characters features;

All of these were used as the same input for four separate models which had different purposes. For example: A) classifying gender from only the first name; B) classifying gender from only the alias name; C) classifying username into first name or alias name; and D) classifying gender from username. The final model in Fig. 9 combines the output of all four models to produce a final model.

We used K-Nearest Neighbor, Support Vector Machine, Random Forest, Multinomial Naïve Bayes, and Neural Network in Models A, B, C, and D, while a Neural Network was used to combine the results from all the classifiers.

At the end of this process, the experiments were divided into three parts in the combined models, with the first part of the combined models to classify gender for first name (Model A) and gender for the alias name (Model B). The second part combined models to classify gender for the first name (Model A), gender for the alias name (Model B), and classify the first name and alias name (Model C). The third part combined all the models.

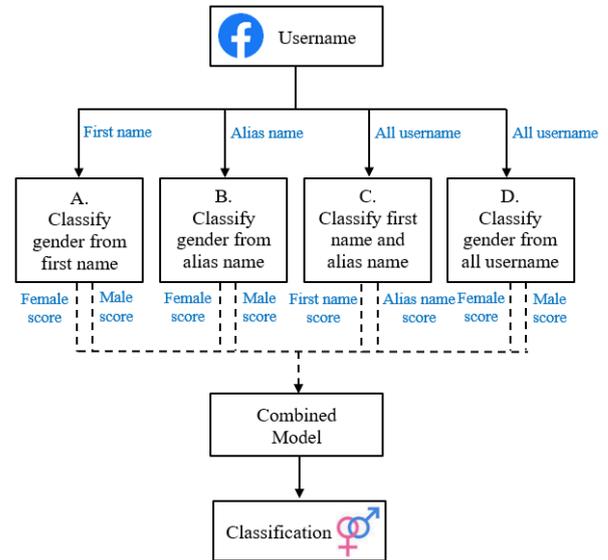


Fig. 9. Combined model that merges the output of the individual models. Four classifiers include: (a) Classify gender from first name; (b) classify gender from alias name; (c) classify first name and alias name; and (d) classify gender from username.

### G. Evaluation Approach

A program was developed in the Python 3 environment using an open source scikit-learn 0.21.3 [15].

All the experiments were performed using stratified 10-fold cross validation and evaluated performance based on accuracy. In k-fold cross-validation, the original samples were randomly partitioned into k equal sized subsamples. Each k subsamples was used as the test data, while the remaining k-1 subsamples were used as training data. The cross-validation process was then repeated k times, with each of the k subsamples used exactly once as the test data. The k results from the folds could then be averaged to produce a single estimation.

Accuracy was defined as the proportion of true results, either true positive or true negative (1).

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FN} \quad (1)$$

where  $TP$  is true positives,  $TN$  is true negatives,  $FP$  is false positives, and  $FN$  is false negatives. Accuracy was determined to be the proportion of true results, either true positive or true negative.

## IV. RESULTS

### A. Gender Classification from First Name

Each classifier was trained with character frequency and

substring characters of the first name only, not including surname. The model outputs were female and male scores. The best performance consistently achieved using Support Vector Machine reached an accuracy of 78.17%, followed by Multinomial Naïve Bayes (76.90%), Neural Network (76.09%), K-Nearest Neighbor (73.13%), and Random Forest (72.72%). The accuracy of the gender classification for first name is summarized in Table II.

TABLE II: ACCURACY OF GENDER CLASSIFICATION FROM FIRST NAME

Classifier	Accuracy
K-Nearest Neighbor	73.13%
Support Vector Machine	78.17%
Random Forest	72.72%
Multinomial Naïve Bayes	76.90%
Neural Network	76.09%

### B. Gender Classification from Alias Name

Each classifier was trained with word tokenization features from the alias name only. In this research, three gender groups were identified: Female; male; and unclassified. Female and male name had a probability of over 60%, while unclassified were less than 60%. The model outputs were female and male scores. The best performance was from Support Vector Machine at 69.45% of accuracy, followed by Multinomial Naïve Bayes (66.04%), Neural Network (65.09%), Random Forest (61.30%), and K-Nearest Neighbor (56.41%). The gender classification accuracy for alias name is summarized in Table III.

TABLE III: ACCURACY OF GENDER CLASSIFICATION FROM ALIAS NAME

Classifier	Accuracy
K-Nearest Neighbor	56.41%
Support Vector Machine	69.45%
Random Forest	61.30%
Multinomial Naïve Bayes	66.04%
Neural Network	65.09%

### C. First Name and Alias Name Classification

Each classifier was trained with word tokenization and speech part features. The purpose was to classify the first name from the alias name. The model outputs were first name and alias scores. Multinomial Naïve Bayes was the best classifier which achieved an accuracy of 83.18%, followed by Support Vector Machine (82.44%), Neural Network (80.12%), Random Forest (78.11%), and K-Nearest Neighbor (74.10%). The accuracy of the first and alias name classifications are shown in Table IV.

TABLE IV: ACCURACY OF FIRST NAME AND ALIAS NAME CLASSIFICATION

Classifier	Accuracy
K-Nearest Neighbor	74.10%
Support Vector Machine	82.44%
Random Forest	78.11%
Multinomial Naïve Bayes	83.18%
Neural Network	80.12%

### D. Gender Classification from All Usernames

Each classifier was trained with word tokenization

features for all usernames. The model outputs were female and male scores. The same as in the previous experiment, the Multinomial Naïve Bayes had the best performance at 65.81%, followed by Neural Network (63.70%), Random Forest (60.97%), Support Vector Machine (59.37%), and K-Nearest Neighbor (56.94%). The accuracy of the gender classification for all usernames is summarized in Table V.

TABLE V: ACCURACY OF GENDER CLASSIFICATION FROM ALL USERNAMES

Classifier	Accuracy
K-Nearest Neighbor	56.94%
Support Vector Machine	59.37%
Random Forest	60.97%
Multinomial Naïve Bayes	65.81%
Neural Network	63.70%

### E. Combining Classification Models

To determine the best final classifier, the performance of the five classifiers were compared, including K-Nearest Neighbor, Support Vector Machine, Random Forest, Multinomial Naïve Bayes, and Neural Network. We also compared the performance obtained from the three different input sets, that is all the models, the combination of three models, and the combination of two models. The performance from getting inputs from all the models had the best accuracy at 91.75% using the Neural Network, this was followed by the combination of three models and the combination of two models, respectively. The accuracy of the combined model is summarized in Table VI.

TABLE VI: COMBINED MODEL ACCURACY

Classifier	Two models	Three models	All models
K-Nearest Neighbor	74.91%	76.74%	81.72%
Support Vector Machine	75.40%	76.67%	84.04%
Random Forest	76.35%	77.65%	88.58%
Multinomial Naïve Bayes	79.68%	79.87%	86.89%
Neural Network	82.07%	83.46%	91.75%

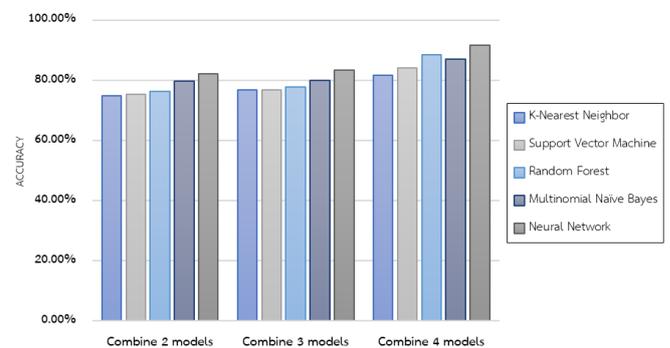


Fig. 10. The accuracy in combined models.

Fig. 10 summarizes the achieved accuracy per classifier for the three combination types. The combination of two models obtained an 82.07% level of accuracy with Neural network. For the combination of three models, an 83.46% level of accuracy was achieved with Neural network. Finally, the combination of four models achieved an accuracy of 91.75% using Neural network.

## V. CONCLUSION

Facebook data can be analyzed and studied for many benefits, but the data often lacks user demographic details which can result in the user information used in analysis or communication being incorrect and not matching the target group. This study therefore established the objective to classify gender from username only in order to solve the problem of correctly using user information and maximizing benefits.

The present study demonstrated that Thai names have certain patterns that can reveal the owner's gender. The authors also presented a gender classification method for Thai Facebook usernames using a combined model. Instead of applying the same model for all features, related features were grouped and used in separate models since first names and alias names have different characteristics. The output of each model was then used as inputs for the final model. The features proposed, including the word tokenization, speech part classification, character frequency, and substring characters can achieve a good result. The experimental results demonstrate that using word tokenization for all usernames achieved a baseline 65.81% level of accuracy, but the combined model achieved an improved performance with a 91.75% level accuracy.

## CONFLICT OF INTEREST

The authors declare no conflict of interest

## AUTHOR CONTRIBUTIONS

Supitcha, Y., Sukree, S. contributed to the design and implementation of the research, to the analysis of the results, to the writing of the paper and had approved the final version.

## REFERENCES

- [1] V. Pichit, "Social media: Future media," *Excusive Journal*, vol. 4, pp. 99-103, 2011.
- [2] We Are Social Ltd. (January 2019). Global and Thailand digital report 2019. [Online]. Available: <https://wearesocial.com/global-digital-report-2019>
- [3] N. Cesare *et al.*, "Demographics in Social Media Data for Public Health Research: Does it matter?" arXiv preprint arXiv:1710.11048, 2017.
- [4] H. A. Schwartz *et al.*, "Personality, gender, and age in the language of social media: The open-vocabulary approach," *PLoS One*, vol. 8, no. 9, p. e73791, 2013.

- [5] J. S. Alowibdi, U. A. Buy, and P. Yu, "Empirical evaluation of profile characteristics for gender classification on Twitter," in *Proc. the 2013 12th International Conference on Machine Learning and Applications*, vol. 1, pp. 365-369, 2013.
- [6] S. Bergsma *et al.*, "Broadly improving user classification via communication-based name and location clustering on twitter," in *Proc. the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2013.
- [7] R. Akbar, *Gender Classification of Indonesian Names Using Multinomial Naive Bayes and Random Forrest Classifiers*, 2016.
- [8] A. A. Septiandri, "Predicting the gender of Indonesian names," arXiv preprint arXiv:1707.07129, 2017.
- [9] M. Briedienė and J. Kapočiušė-Dzikienė, "An automatic author profiling from non-normative lithuanian texts," in *Proc. International Conference on Information Technologies*, Kaunas, Lithuania, 2018, vol. 2145.
- [10] R. Hirt, N. Kuhl, and G. Satzger, "Cognitive computing for customer profiling: Meta classification for gender prediction," *Electronic Markets*, vol. 29, no. 1, pp. 93-106 2019.
- [11] M. Vicente, F. Batista, and J. P. Carvalho, "Gender detection of Twitter users based on multiple information sources," *Interactions between Computational Intelligence and Mathematics*, part 2, pp. 39-54, 2019.
- [12] PyThaiNLP 2.0.7. (August 2019). Thai natural language processing in Python. [Online]. Available: <https://pypi.org/project/pythainlp/>
- [13] V. Sornlertlamvanich, T. Charoenporn, and H. Isahara, "ORCHID: Thai part-of-speech tagged corpus," National Electronics and Computer Technology Center Technical Report, pp. 5-19, 1997.
- [14] T. Theeramunkong *et al.*, "Character cluster based Thai information retrieval," in *Proc. the Fifth International Workshop on Information Retrieval with Asian Languages*, 2000.
- [15] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, pp. 2825-2830, 2011.

Copyright © 2020 by the authors. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited ([CC BY 4.0](https://creativecommons.org/licenses/by/4.0/)).



**Supitcha Yuenyong** was born in Suphanburi, Thailand, in 1990. She received the bachelor's degree in computer engineering at the Department of Computer Engineering, King Mongkut's University of Technology Thonburi. She is currently studying the master's degree in computer science at the Department of Computer Engineering, Chulalongkorn University. Her research interests are machine learning, natural language processing, and social network analysis.



**Sukree Sinthupinyo** was born in Bangkok, Thailand, in 1975. He received the bachelor's, master's, and doctoral degree in computer engineer at the Department of Computer Engineering, Chulalongkorn University. Now he is an associate professor at Chulalongkorn University. His research interests are artificial intelligence, machine learning, innovation, social network analysis, and social network mining.