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.

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. Kühl, 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

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).

doi: 10.18178/ijmlc.2020.10.5.982

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 Naïve 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.

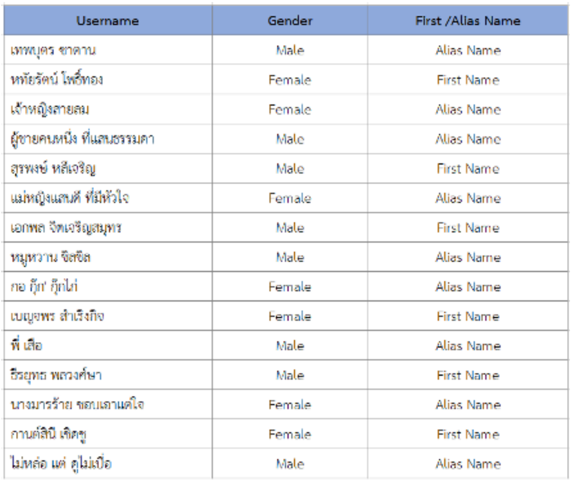
TABLE I: TOTAL NUMBER OF USERNAME DATASETS IN THE STUDY

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.

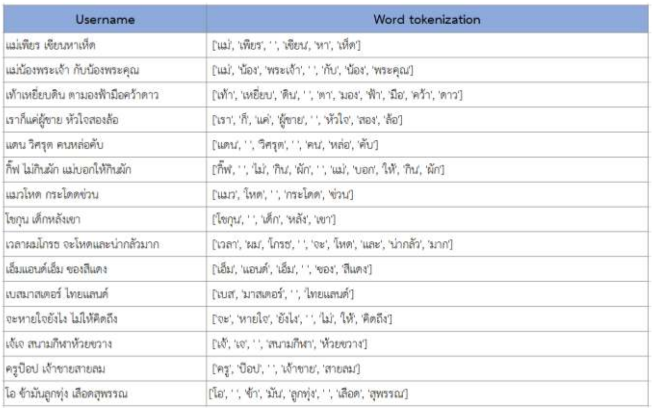


Username	Gender	First /Alias Name
เทพบุตร ซาตาน	Male	Alias Name
หทัยรัตน์ โพธิ์ทอง	Female	First Name
เจ้าหญิงสายลม	Female	Alias Name
ผู้ชายคนหนึ่ง ที่แสนธรรมดา	Male	Alias Name
สุรพงษ์ หลีเจริญ	Male	First Name
แม่หญิงแสนดี ที่มีหัวใจ	Female	Alias Name
เอกพล จิตเจริญสมุทร	Male	First Name
หมูหวาน จิ๊ดชิ๊ด	Male	Alias Name
กอ กุ๊ก' กุ๊กไก่	Female	Alias Name
เบญจพร สำเร็จกิจ	Female	First Name
พี่ เสือ	Male	Alias Name
ธีรยุทธ พลวงศ์ษา	Male	First Name
นางมารร้าย ชอบเอาแต่ใจ	Female	Alias Name
กานต์สินี เทิดชู	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
แม่เพียร เขียนหาเห็ด	['แม่', 'เพียร', ' ', 'เขียน', 'หา', 'เห็ด']
แม่น้องพระเจ้า กับน้องพระคุณ	['แม่', 'น้อง', 'พระเจ้า', ' ', 'กับ', 'น้อง', 'พระคุณ']
เท้าเหยียบดิน ตามองฟ้ามือคว้าดาว	['เท้า', 'เหยียบ', 'ดิน', ' ', 'ตา', 'มอง', 'ฟ้า', 'มือ', 'คว้า', 'ดาว']
เราก็แค่ผู้ชาย หัวใจสองล้อ	['เรา', 'ก็', 'แค่', 'ผู้ชาย', ' ', 'หัวใจ', 'สอง', 'ล้อ']
แดน วิศรุต คนหล่อดับ	['แดน', ' ', 'วิศรุต', ' ', 'คน', 'หล่อ', 'ดับ']
กิ๊ฟ ไม่กินผัก แม่บอกให้กินผัก	['กิ๊ฟ', ' ', 'ไม่', 'กิน', 'ผัก', ' ', 'แม่', 'บอก', 'ให้', 'กิน', 'ผัก']
แมวโหด กระโดดข่วน	['แมว', 'โหด', ' ', 'กระโดด', 'ข่วน']
ไขกุน เด็กหลังเขา	['ไขกุน', ' ', 'เด็ก', 'หลัง', 'เขา']
เวลาผมโกรธ จะโหดและน่ากลัวมาก	['เวลา', 'ผม', 'โกรธ', ' ', 'จะ', 'โหด', 'และ', 'น่ากลัว', 'มาก']
เอ็มแอนด์เอ็ม ของสีแดง	['เอ็ม', 'แอนด์', 'เอ็ม', ' ', 'ของ', 'สีแดง']
เบสมาสเตอร์ ไทยแลนด์	['เบส', 'มาสเตอร์', ' ', 'ไทยแลนด์']
จะหายใจยังไง ไม่ให้คิดถึง	['จะ', 'หายใจ', 'ยังไง', ' ', 'ไม่', 'ให้', 'คิดถึง']
เจ้เจ สนามกีฬาห้วยขวาง	['เจ้', 'เจ', ' ', 'สนามกีฬา', 'ห้วยขวาง']
ครูป๊อป เจ้าชายสายลม	['ครู', 'ป๊อป', ' ', 'เจ้าชาย', 'สายลม']
โอ ข้ามันลูกทุ่ง เลือดสุพรรณ	['โอ', ' ', 'ข้า', 'มัน', 'ลูกทุ่ง', ' ', 'เลือด', '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.

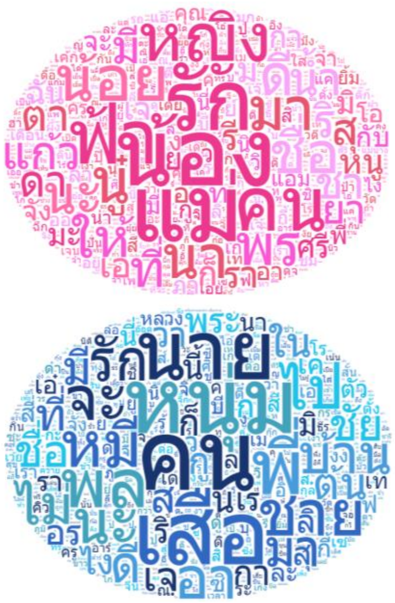
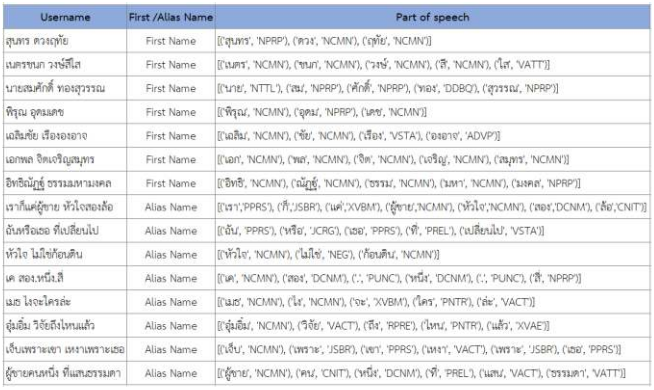
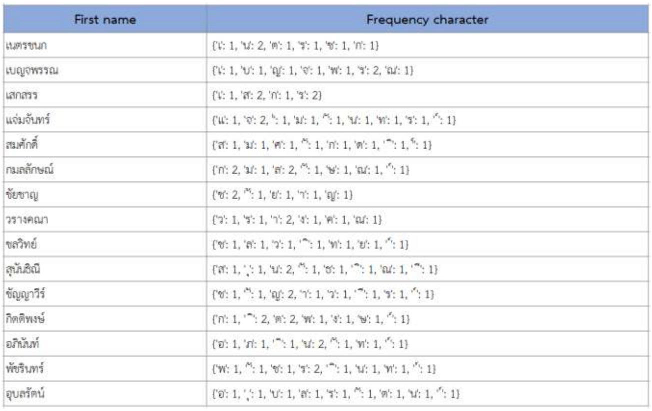


Fig. 3. Word clouds for alias name word tokenization. (female on top and male below).



Username	First /Alias Name	Part of speech
สุนทร ดวงฤทัย	First Name	[('สุนทร', 'NPRP'), ('ดวง', 'NCMN'), ('ฤทัย', 'NCMN')]
เนตรขนก วงษ์สีใส	First Name	[('เนตร', 'NCMN'), ('ขนก', 'NCMN'), ('วงษ์', 'NCMN'), ('สี', 'NCMN'), ('ใส', 'VATT')]
นายสมศักดิ์ ทองสุวรรณ	First Name	[('นาย', 'NTTL'), ('สม', 'NPRP'), ('ศักดิ์', 'NPRP'), ('ทอง', 'DDBQ'), ('สุวรรณ', 'NPRP')]
พิรุณ อุดมเดช	First Name	[('พิรุณ', 'NCMN'), ('อุดม', 'NPRP'), ('เดช', 'NCMN')]
เฉลิมชัย เรืององอาจ	First Name	[('เฉลิม', 'NCMN'), ('ชัย', 'NCMN'), ('เรือง', 'VSTA'), ('องอาจ', 'ADVP')]
เอกพล จิตเจริญสมุทร	First Name	[('เอก', 'NCMN'), ('พล', 'NCMN'), ('จิต', 'NCMN'), ('เจริญ', 'NCMN'), ('สมุทร', 'NCMN')]
อิทธิณัฏฐ์ ธรรมมหามงคล	First Name	[('อิทธิ', 'NCMN'), ('ณัฏฐ์', 'NCMN'), ('ธรรม', 'NCMN'), ('มหา', 'NCMN'), ('มงคล', 'NPRP')]
เราก็แค่ผู้ชาย หัวใจสองล้อ	Alias Name	[('เรา','PPRS'), ('ก็','JSBR'), ('แค่','XVBM'), ('ผู้ชาย','NCMN'), ('หัวใจ','NCMN'), ('สอง','DCNM'), ('ล้อ','CNIT')]
ฉันหรือเธอ ที่เปลี่ยนไป	Alias Name	[('ฉัน', 'PPRS'), ('หรือ', 'JCRG'), ('เธอ', 'PPRS'), ('ที่', 'PREL'), ('เปลี่ยนไป', 'VSTA')]
หัวใจ ไม่ใช่ก้อนดิน	Alias Name	[('หัวใจ', 'NCMN'), ('ไม่ใช่', 'NEG'), ('ก้อนดิน', 'NCMN')]
เค สอง.หนึ่ง.สี่	Alias Name	[('เค', 'NCMN'), ('สอง', 'DCNM'), ('.', 'PUNC'), ('หนึ่ง', 'DCNM'), ('.', 'PUNC'), ('สี่', 'NPRP')]
เมธ ไงจะใครล่ะ	Alias Name	[('เมธ', 'NCMN'), ('ไง', 'NCMN'), ('จะ', 'XVBM'), ('ใคร', 'PNTR'), ('ล่ะ', 'VACT')]
อุ๋มอิ๋ม วิจัยถึงไหนแล้ว	Alias Name	[('อุ๋มอิ๋ม', 'NCMN'), ('วิจัย', 'VACT'), ('ถึง', 'RPRE'), ('ไหน', 'PNTR'), ('แล้ว', 'XVAE')]
เจ็บเพราะเขา เหงาเพราะเธอ	Alias Name	[('เจ็บ', 'NCMN'), ('เพราะ', 'JSBR'), ('เขา', 'PPRS'), ('เหงา', 'VACT'), ('เพราะ', 'JSBR'), ('เธอ', 'PPRS')]
ผู้ชายคนหนึ่ง ที่แสนธรรมดา	Alias Name	[('ผู้ชาย', 'NCMN'), ('คน', 'CNIT'), ('หนึ่ง', 'DCNM'), ('ที่', 'PREL'), ('แสน', 'VACT'), ('ธรรมดา', 'VATT')]

Fig. 4. Example of Thai speech part classification.



First name	Frequency character
เนตรขนก	{'เ': 1, 'น': 2, 'ต': 1, 'ร': 1, 'ข': 1, 'ก': 1}
เบญจพรรณ	{'เ': 1, 'บ': 1, 'ญ': 1, 'จ': 1, 'พ': 1, 'ร': 2, 'ณ': 1}
เสกสรร	{'เ': 1, 'ส': 2, 'ก': 1, 'ร': 2}
แจ่มจันทร์	{'แ': 1, 'จ': 2, '่': 1, 'ม': 1, 'ั': 1, 'น': 1, 'ท': 1, 'ร': 1, '์': 1}
สมศักดิ์	{'ส': 1, 'ม': 1, 'ศ': 1, 'ั': 1, 'ก': 1, 'ด': 1, 'ิ': 1, '์': 1}
กมลลักษณ์	{'ก': 2, 'ม': 1, 'ล': 2, 'ั': 1, 'ษ': 1, 'ณ': 1, '์': 1}
ชัยชาญ	{'ช': 2, 'ั': 1, 'ย': 1, 'า': 1, 'ญ': 1}
วรางคณา	{'ว': 1, 'ร': 1, 'า': 2, 'ง': 1, 'ค': 1, 'ณ': 1}
ชลวิทย์	{'ช': 1, 'ล': 1, 'ว': 1, 'ิ': 1, 'ท': 1, 'ย': 1, '์': 1}
สุนันธิณี	{'ส': 1, 'ุ': 1, 'น': 2, 'ั': 1, 'ธ': 1, 'ิ': 1, 'ณ': 1, 'ี': 1}
ชัญญาวีร์	{'ช': 1, 'ั': 1, 'ญ': 2, 'า': 1, 'ว': 1, 'ี': 1, 'ร': 1, '์': 1}
กิตติพงษ์	{'ก': 1, 'ิ': 2, 'ต': 2, 'พ': 1, 'ง': 1, 'ษ': 1, '์': 1}
อภินันท์	{'อ': 1, 'ภ': 1, 'ิ': 1, 'น': 2, 'ั': 1, 'ท': 1, '์': 1}
พัชรินทร์	{'พ': 1, 'ั': 1, 'ช': 1, 'ร': 2, 'ิ': 1, 'น': 1, 'ท': 1, '์': 1}
อุบลรัตน์	{'อ': 1, 'ุ': 1, 'บ': 1, 'ล': 1, 'ร': 1, 'ั': 1, 'ต': 1, 'น': 1, '์': 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 'ษ', 'ง', 'ช', 'ต' andิ' appear more frequently in male first names.

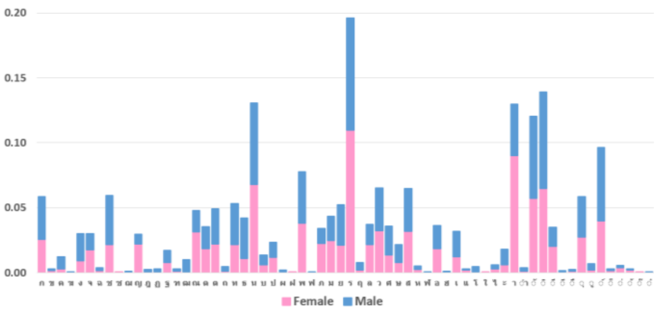
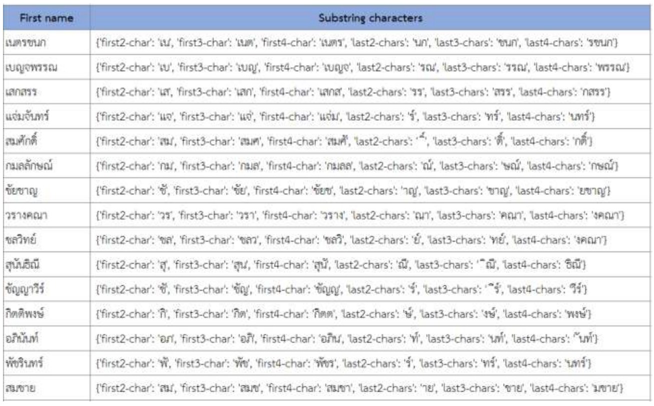


Fig. 6. Character occurrence frequency in first names.

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 'ฃ', 'ฝ', 'ฟ',๊', 'ใ',๋', 'ฒ', '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
เนตรขนก	{'first2-char': 'เน', 'first3-char': 'เนต', 'first4-char': 'เนตร', 'last2-chars': 'นก', 'last3-chars': 'ขนก', 'last4-chars': 'รขนก'}
เบญจพรรณ	{'first2-char': 'เบ', 'first3-char': 'เบญ', 'first4-char': 'เบญจ', 'last2-chars': 'รณ', 'last3-chars': 'รรณ', 'last4-chars': 'พรรณ'}
เสกสรร	{'first2-char': 'เส', 'first3-char': 'เสก', 'first4-char': 'เสกส', 'last2-chars': 'รร', 'last3-chars': 'สรร', 'last4-chars': 'กสรร'}
แจ่มจันทร์	{'first2-char': 'แจ', 'first3-char': 'แจ่', 'first4-char': 'แจ่ม', 'last2-chars': 'ร์', 'last3-chars': 'ทร์', 'last4-chars': 'นทร์'}
สมศักดิ์	{'first2-char': 'สม', 'first3-char': 'สมศ', 'first4-char': 'สมศั', 'last2-chars': 'ิ์', 'last3-chars': 'ดิ์', 'last4-chars': 'กดิ์'}
กมลลักษณ์	{'first2-char': 'กม', 'first3-char': 'กมล', 'first4-char': 'กมลล', 'last2-chars': 'ณ์', 'last3-chars': 'ษณ์', 'last4-chars': 'กษณ์'}
ชัยชาญ	{'first2-char': 'ชั', 'first3-char': 'ชัย', 'first4-char': 'ชัยช', 'last2-chars': 'าญ', 'last3-chars': 'ชาญ', 'last4-chars': 'ยชาญ'}
วรางคณา	{'first2-char': 'วร', 'first3-char': 'วรา', 'first4-char': 'วราง', 'last2-chars': 'ณา', 'last3-chars': 'คณา', 'last4-chars': 'งคณา'}
ชลวิทย์	{'first2-char': 'ชล', 'first3-char': 'ชลว', 'first4-char': 'ชลวิ', 'last2-chars': 'ย์', 'last3-chars': 'ทย์', 'last4-chars': 'งคณา'}
สุนันธิณี	{'first2-char': 'สุ', 'first3-char': 'สุน', 'first4-char': 'สุนั', 'last2-chars': 'ณี', 'last3-chars': 'ิณี', 'last4-chars': ธิณี'}
ชัญญาวีร์	{'first2-char': 'ชั', 'first3-char': 'ชัญ', 'first4-char': 'ชัญญ', 'last2-chars': 'ร์', 'last3-chars': 'ีร์', 'last4-chars': วีร์'}
กิตติพงษ์	{'first2-char': 'กิ', 'first3-char': 'กิต', 'first4-char': 'กิตต', 'last2-chars': 'ษ์', 'last3-chars': 'งษ์', 'last4-chars': 'พงษ์'}
อภินันท์	{'first2-char': 'อภ', 'first3-char': 'อภิ', 'first4-char': 'อภิน', 'last2-chars': 'ท์', 'last3-chars': 'นท์', 'last4-chars': 'ันท์'}
พัชรินทร์	{'first2-char': 'พั', 'first3-char': 'พัช', 'first4-char': 'พัชร', 'last2-chars': 'ร์', 'last3-chars': 'ทร์', 'last4-chars': 'นทร์'}
สมชาย	{'first2-char': 'สม', 'first3-char': 'สมช', 'first4-char': 'สมชา', 'last2-chars': 'าย', 'last3-chars': 'ชาย', 'last4-chars': 'มชาย'}

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; ั'พย์', 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; 'ธี'.

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.

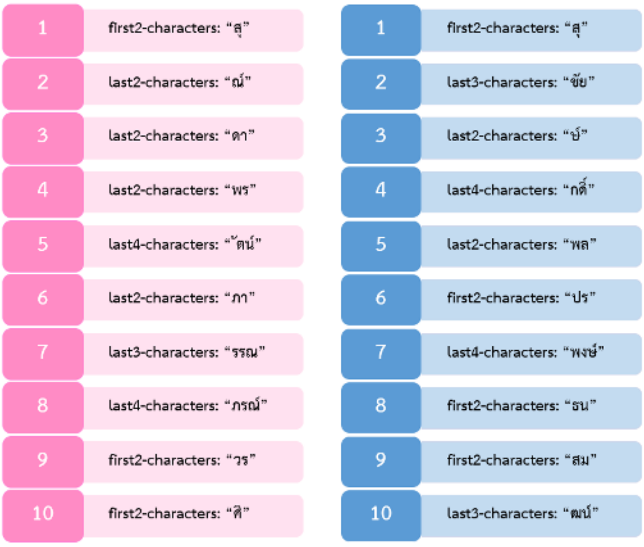


Fig. 8. Top 10 most common first name substring characters. (female on the left and male on the right).

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.

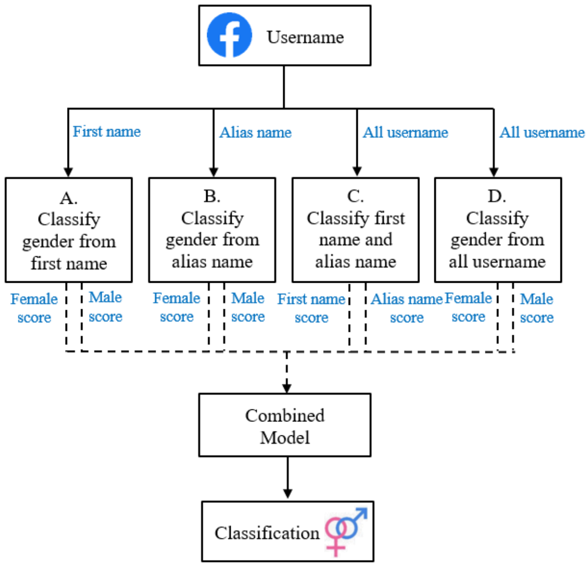


Fig. 9. Combined model that merges the output of the individual models. Four classifiers include: (a) Classify gender from first name; (b) classify gender from alias name; (c) classify first name and alias name; and (d) classify gender from username.

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).

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + TN + FN} \tag{1}$$

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, Supper 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%

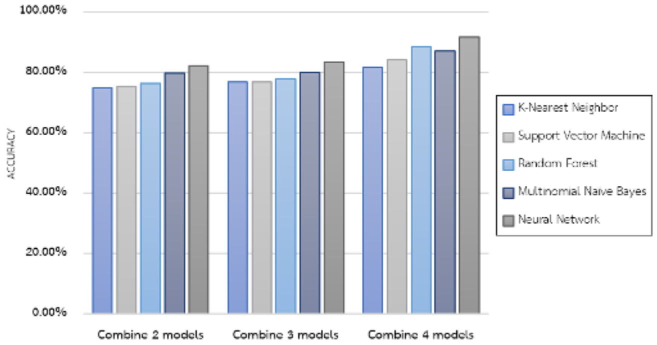


Fig. 10. The accuracy in combined models.

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. Kapočiutė-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. Kühl, 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.

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.