การจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊ก

น.ส.สุพิชชา ยืนยงค์

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต
สาขาวิชาวิทยาศาสตร์คอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2562

Gender Classification of Thai Username on Facebook

Miss Supitcha Yuenyong

A Thesis Submitted in Partial Fulfillment of the Requirements
for the Degree of Master of Science in Computer Science
Department of Computer Engineering
FACULTY OF ENGINEERING
Chulalongkorn University
Academic Year 2019

หัวข้อวิทยานิพนธ์	การจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊ก
โดย	น.ส.สุพิชชา ยืนยงค์
สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์
อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก	ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ

คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย อนุมัติให้นับวิทยานิพนธ์ฉบับนี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิทยาศาสตรมหาบัณฑิต

..... คณบดีคณะวิศวกรรมศาสตร์
(ศาสตราจารย์ ดร.สุพจน์ เตชวรสินสกุล)

คณะกรรมการสอบวิทยานิพนธ์

..... ประธานกรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.นัทที นิภานันท์)

..... อาจารย์ที่ปรึกษาวิทยานิพนธ์หลัก
(ผู้ช่วยศาสตราจารย์ ดร.สุกรี สินธุภิญโญ)

..... กรรมการ
(ผู้ช่วยศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเนศ)

..... กรรมการภายนอกมหาวิทยาลัย
(ผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ)

สุพิชชา ยืนยงค์ : การจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊ก. (Gender Classification of Thai Username on Facebook) อ.ที่ปรึกษาหลัก : ผศ. ดร.สุกรี สินธุภิญโญ

วิทยานิพนธ์นี้นำเสนอการนำการเรียนรู้ของเครื่องมาประยุกต์ในการจำแนกเพศผู้ใช้งานเฟซบุ๊กโดยใช้เพียงชื่อผู้ใช้งานเท่านั้น ซึ่งข้อมูลส่วนตัวของผู้ใช้งานของโซเชียลเนตเวิร์คมีความสำคัญในการนำมาวิเคราะห์ แต่บางครั้งไม่มีการเปิดเผยข้อมูล เช่น อายุ หรือเพศ โดยการศึกษาส่วนใหญ่มักจะนำเอาข้อความบนเว็บเพจมาวิเคราะห์ แต่การศึกษานี้เลือกใช้ชื่อผู้ใช้งานในการจำแนกเพศ โดยเพศสามารถอนุมานได้จากทั้งชื่อจริงและชื่อแฝงของผู้ใช้งาน โดยงานวิจัยนี้สนใจเฉพาะชื่อที่เป็นภาษาไทย ซึ่งชื่อของคนไทยจะมีรูปแบบที่สามารถแสดงตัวตนความเป็นเพศได้ การรวมกันของแบบจำลองสำหรับการจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊กที่แตกต่างกันในแต่ละแบบจำลองมีการเรียนรู้โดยใช้วิธีการเรียนรู้แบบจำลองเชิงทำนาย ได้แก่ การจำแนกเพศจากชื่อจริง การจำแนกเพศจากชื่อแฝง การจำแนกชื่อจริงและชื่อแฝง และการจำแนกชื่อทั้งหมด โดยผลการจำแนกทั้งหมดจะถูกรวมในแบบจำลองสุดท้าย เมื่อใช้วิธีนี้แบบจำลองมีความถูกต้องที่ 85.85% ซึ่งได้ผลลัพธ์ที่ดีกว่าเมื่อเปรียบเทียบกับการจำแนกเพศโดยคน ที่มีความถูกต้องที่ 77.03%

สาขาวิชา	วิทยาศาสตร์คอมพิวเตอร์	ลายมือชื่อนิสิต .....
ปีการศึกษา	2562	ลายมือชื่อ อ.ที่ปรึกษาหลัก .....

6170973921 : MAJOR COMPUTER SCIENCE

KEYWORD: Gender classification, Facebook username, Name analysis, Social Network, Machine learning

Supitcha Yuenyong : Gender Classification of Thai Username on Facebook. Advisor: Asst. Prof. SUKREE SINTHUPINYO, Ph.D.

This thesis 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 include gender classification from first name, gender classification from alias name, first name and alias name classification, and gender classification from all usernames. Furthermore, all the classification results were combined into a final model. Using this method, the model achieved 85.85% level of accuracy. Which has better results when compared to gender classification by the human that has accuracy is 77.03%

Field of Study: Computer Science Student's Signature .....

Academic Year: 2019 Advisor's Signature .....

กิตติกรรมประกาศ

วิทยานิพนธ์ฉบับนี้ไม่สามารถสำเร็จลุล่วงได้ด้วยผู้วิจัยเพียงคนเดียว ยังมีบุคคลที่ให้ความรู้ ความช่วยเหลือและให้การสนับสนุน

ขอขอบพระคุณอาจารย์ที่ปรึกษา ผู้ช่วยศาสตราจารย์ ดร.สุกรี สิญธุภิญโญ ที่ให้คำปรึกษาและช่วยเหลือ รวมทั้งเป็นแรงผลักดัน ชี้แนะแนวทางต่าง ๆ อันเป็นประโยชน์อย่างยิ่งในการทำวิจัย

ขอขอบพระคุณคณะกรรมการสอบวิทยานิพนธ์ซึ่งประกอบไปด้วย ผู้ช่วยศาสตราจารย์ ดร.นัททที นิภานันท์ ผู้ช่วยศาสตราจารย์ ดร.ณัฐพงศ์ ชินธเบศ และผู้ช่วยศาสตราจารย์ ดร.เด่นดวง ประดับสุวรรณ ที่ได้กรุณาให้เกียรติเป็นคณะกรรมการ รวมทั้งให้คำปรึกษาและข้อเสนอแนะอันเป็นประโยชน์อย่างมากต่อการทำวิจัยและวิทยานิพนธ์ฉบับนี้

ขอขอบพระคุณคณาจารย์ ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย ผู้ถ่ายทอดความรู้ รวมถึงให้คำแนะนำที่เป็นประโยชน์

สุดท้ายนี้ขอขอบพระคุณครอบครัว ผู้ให้กำลังใจ และผู้ที่อยู่เบื้องหลังความสำเร็จ ที่เป็นแรงขับเคลื่อนที่สำคัญ และสนับสนุนให้วิทยานิพนธ์ฉบับนี้สำเร็จลุล่วงไปด้วยดี

สุพิชชา ยืนยงค์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญภาพ

หน้า

บทที่ 1
บทนำ

1.1 ความเป็นมาและเหตุผลการวิจัย

สังคมออนไลน์ (Social network) ได้เข้ามามีอิทธิพลต่อการดำเนินชีวิตของผู้คนเกือบทุกด้าน คือ ช่วยให้ได้ข้อมูลข่าวสารรวดเร็วมากยิ่งขึ้น ทำให้ผู้คนสามารถสื่อสารกันได้สะดวกง่ายดาย ประหยัดเวลา ทั้งการได้รับรู้ข่าวสารจากทั่วทุกมุมโลก การจำหน่ายสินค้า การรับข้อมูล การโฆษณา การประชาสัมพันธ์ บทความด้านสุขภาพ การเมือง ซึ่งถือเป็นการเผยแพร่ข้อมูลหรือการแสดงความคิดเห็นออกมาให้สาธารณะได้รับทราบ สังคมออนไลน์เปิดโอกาสให้ผู้ใช้มีอิสระในการใช้บริการสูง โดยปราศจากข้อจำกัด ด้านเวลาและสถานที่ ทำให้เกิดการรวมตัวกันเป็นกลุ่มเพื่อแลกเปลี่ยนความรู้ ทัศนคติ เรื่องราว ที่สนใจ กลายเป็นชุมชนบนโลกออนไลน์ [1]

สื่อสังคมออนไลน์ (Social Media) เป็นสื่อที่ผู้ส่งสารแบ่งปันสารที่อยู่ในรูปแบบต่าง ๆ ไปยังผู้รับสารผ่านเครือข่ายออนไลน์ โดยสามารถโต้ตอบกันระหว่างผู้ส่งสารและผู้รับสารด้วยกันเอง ซึ่งสื่อสังคมออนไลน์ได้ถูกใช้งานอย่างแพร่หลายจนกลายเป็นส่วนหนึ่งในชีวิตของผู้คนในปัจจุบัน อีกทั้งเป็นแพลตฟอร์มที่สามารถตอบโจทย์การติดต่อสื่อสารได้ทุกเพศทุกวัย ไม่ว่าผู้ใช้บริการจะอยู่ในภาคธุรกิจหรือภาคส่วนใด โดยที่แต่ละบุคคลสามารถสร้างเครือข่ายทั้งระดับเล็กหรือระดับใหญ่ในสังคม ผ่านการรู้จักต่อไปเป็นทอด ๆ คล้ายเครือข่ายใยแมงมุม สร้างการพูดคุย บอกเล่า หรือนำเสนอข้อมูลส่วนตัว รูปภาพ และยังสามารถแสดงความคิดเห็นได้ โดยสื่อสังคมออนไลน์สามารถแบ่งออกเป็นประเภทต่าง ๆ เช่น บล็อก (เช่น เวิร์ดเพรส) ไมโครบล็อก (เช่น ทวิตเตอร์) เครือข่ายสังคมออนไลน์ (เช่น เฟซบุ๊ก) และการแบ่งปันสื่อทางออนไลน์ (เช่น ยูทูป) [2] สำหรับในประเทศไทยมีคนใช้งานสื่อสังคมออนไลน์มากถึง 51 ล้านคน โดยที่นิยมใช้มากที่สุด ได้แก่ เฟซบุ๊ก ยูทูป ไลน์ เฟซบุ๊กเมสเซนเจอร์ และอินสตราแกรม ตามลำดับ [3] ลักษณะโดยทั่วไปของการสื่อสารในเครือข่ายสังคมออนไลน์เกิดขึ้นผ่านทางข้อความสั้น ๆ มักใช้รูปแบบภาษาที่ไม่ได้มาตรฐาน และยังมีการสื่อสารโดยที่ไม่ใช้ข้อความ เช่น สัญลักษณ์ สติ๊กเกอร์ รูปภาพ เป็นต้น

จากการเผยแพร่ข้อมูลดังกล่าวบนโลกออนไลน์ สามารถนำไปวิเคราะห์ศึกษากระบวนการทางสังคม ตั้งแต่ด้านเศรษฐศาสตร์ไปจนถึงด้านสาธารณสุข ตัวอย่างเช่น การทำการตลาดที่ตรงกลุ่มเป้าหมาย การวิเคราะห์ความคิดเห็นเกี่ยวกับเหตุการณ์ทางการเมืองและประเด็นทางสังคม การวิเคราะห์ทัศนคติ สุขภาพจิต หรือพฤติกรรมการนอนหลับ การติดตามการระบาดของโรคและรายงานการเจ็บป่วย นอกจากนี้ ข้อมูลเหล่านี้ยังช่วยให้นักวิจัยสามารถประเมินคุณภาพของข้อมูลอื่น ๆ และพัฒนาวิธีทางสถิติเพื่อปรับข้อมูลที่มีความลำเอียง [4] แต่อย่างไรก็ตาม สื่อสังคมออนไลน์ส่วนใหญ่ไม่ได้มีรายละเอียดข้อมูลประชากรของผู้ใช้งาน อีกทั้งไม่จำเป็นที่จะต้องเปิดเผยตัวตน จึงมีการปลอมหรือปกปิด อายุ เพศ เพื่อซ่อนตัวตนที่แท้จริง อันส่งผลให้การนำข้อมูลของผู้ใช้งานบนสื่อสังคมออนไลน์ไปใช้เพื่อวิเคราะห์หรือสื่อสารเกิดความคลาดเคลื่อน และไม่ตรงกลุ่มเป้าหมาย ดังนั้น การจำแนกเพศที่ถูกต้องของผู้ใช้งานบนสื่อสังคมออนไลน์จึงเป็นเรื่องสำคัญ ที่จะช่วยแก้ปัญหาให้สามารถนำข้อมูลของผู้ใช้งานบนสื่อสังคมออนไลน์มาใช้ได้อย่างถูกต้อง และเกิดประโยชน์สูงสุด ซึ่งมีผู้ทำวิจัยที่เกี่ยวข้องไว้หลายท่าน ได้แก่ Schwartz, H.A. และคณะ [5], Alowibdi, J.S. และคณะ [6], Bergsma, S. และคณะ [7], Akbar, R. [8], Septiandri, A.A. [9], Briediene, M. และคณะ [10], Hirt, R. และคณะ [11], Vicente, M. และคณะ [12] โดยมีการจำแนกเพศจากการใช้ข้อมูลทั้งข้อความ ภาพ สี ตำแหน่ง หรือชื่อ และจากงานวิจัยดังกล่าว วิธีการที่มีอยู่มากที่สุดในการจำแนกเพศมักนิยมวิเคราะห์จากข้อความ แต่บางครั้งไม่สามารถหาข้อความเพื่อมาวิเคราะห์ได้ ชื่อของผู้ใช้งาน ไม่ว่าเป็นชื่อจริง หรือชื่อแฝงในสื่อสังคมออนไลน์ จึงเป็นอีกวิธีการหนึ่งที่สามารถนำมาวิเคราะห์เพศได้ และมีความท้าทายสำหรับการประมวลผลภาษาธรรมชาติเพื่อให้สามารถบ่งบอกเพศได้ ซึ่งชื่อของเพศชายและหญิงในภาษาไทยจะมีลักษณะที่แตกต่างกันออกไป ตัวอย่างเช่น เพศชาย ใช้ชื่อ สุชาติ วีรยุทธ สมเกียรติ ณรงค์ฤทธิ์ ประวุฒิ และเพศหญิง ใช้ชื่อ พัชราภา วิชุลดา พรพรรณ สมศรี วาสนา ดังนั้น ผู้วิจัยจึงเลือกศึกษาและหาแนวทางแก้ปัญหาดังกล่าว ด้วยการจำแนกเพศผู้ใช้เฟซบุ๊กจากชื่อผู้ใช้งาน ทั้งที่เป็นชื่อจริงและชื่อแฝงที่เป็นตัวอักษรภาษาไทย โดยวิเคราะห์คุณลักษณะที่มีผลต่อการจำแนกเพศจากชื่อผู้ใช้งาน และเปรียบเทียบประสิทธิภาพตัวจำแนก

1.2 วัตถุประสงค์ของการวิจัย

1. เพื่อจำแนกเพศผู้ใช้งานเฟซบุ๊กจากชื่อผู้ใช้งานทั้งชื่อจริงและชื่อแฝง ที่เป็นตัวอักษรภาษาไทยด้วยแบบจำลองการจำแนกกลุ่มชื่อจากผู้ใช้งานเฟซบุ๊ก
2. เพื่อสร้างแบบจำลองการจำแนกกลุ่มชื่อจากผู้ใช้งานเฟซบุ๊ก
3. เพื่อวิเคราะห์คุณลักษณะที่มีผลต่อการจำแนกเพศจากชื่อผู้ใช้งานเฟซบุ๊ก
4. เพื่อเปรียบเทียบประสิทธิภาพของแบบจำลองการจำแนกกลุ่มชื่อจากผู้ใช้งานเฟซบุ๊ก

1.3 ขอบเขตการวิจัย

1. จัดกลุ่มเพศจากชื่อผู้ใช้งานบนสื่อสังคมออนไลน์ ทั้งชื่อจริงและชื่อแฝง ที่เป็นตัวอักษรภาษาไทยเท่านั้น ไม่สามารถใช้กับตัวอักษรภาษาอื่น ๆ ได้
2. ชื่อผู้ใช้งานบนสื่อสังคมออนไลน์ที่นำมาใช้ในงานวิจัยนี้มาจากเฟซบุ๊ก
3. เพศที่อ้างอิงเพื่อวิเคราะห์การจำแนกเพศจากชื่อผู้ใช้งาน ได้แก่ เพศชาย และเพศหญิง

1.4 ขั้นตอนการวิจัย

การวิจัยเพื่อการจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊ก มีขั้นตอนการดำเนินงานดังต่อไปนี้

1. ศึกษางานวิจัยเกี่ยวกับการจำแนกเพศของผู้ใช้งานบนสื่อสังคมออนไลน์
2. ศึกษาหลักการตัดคำ การจำแนกชนิดของคำภาษาไทย และการวัดประสิทธิภาพการจำแนกแบบจำลอง
3. ศึกษาวิธีการเก็บรวบรวมข้อมูลชื่อผู้ใช้งานเฟซบุ๊ก
4. ศึกษาตัวจำแนกและเทคนิคการจำแนกชื่อ
5. ศึกษาเครื่องมือที่ใช้ในงานวิจัย เช่น เครื่องมือการตัดคำ เครื่องมือการสร้างแบบจำลอง
6. ออกแบบการทดลอง
7. เก็บรวบรวมข้อมูลชื่อผู้ใช้งานเฟซบุ๊ก
8. สร้างคุณลักษณะ และแบบจำลอง

9. ทดสอบและวัดผลความถูกต้องของการจำแนกเพศจากชื่อผู้ใช้งานเฟซบุ๊ก
10. วิเคราะห์ผลการทดลอง
11. สรุปผลและเรียบเรียงวิทยานิพนธ์

ตารางที่ 1 ตารางขั้นตอนการดำเนินการวิจัย

ขั้นตอนการดำเนินงาน	ระยะเวลาการดำเนินงาน											
	1	2	3	4	5	6	7	8	9	10	11	12
1. ศึกษางานวิจัยที่เกี่ยวข้อง												
2. ศึกษาหลักการตัดคำ ฯ												
3. ศึกษาวิธีการเก็บรวบรวมข้อมูล												
4. ศึกษาตัวจำแนกและเทคนิค												
5. ศึกษาเครื่องมือที่ใช้ในงานวิจัย												
6. ออกแบบการทดลอง												
7. เก็บรวบรวมข้อมูล												
8. สร้างคุณลักษณะ, แบบจำลอง												
9. ทดสอบและวัดผลความถูกต้อง												
10. วิเคราะห์ผลการทดลอง												
11. สรุปผล เรียบเรียง วิทยานิพนธ์												

1.5 ประโยชน์ที่คาดว่าจะได้รับ

ประโยชน์ที่ได้รับจากการวิจัยในครั้งนี้ มีดังนี้

1. สามารถสร้างแบบจำลองการจัดกลุ่มชื่อผู้ใช้งานบนสื่อสังคมออนไลน์ได้
2. สามารถเปรียบเทียบประสิทธิภาพตัวจำแนกต่าง ๆ เพื่อจำแนกเพศได้
3. สามารถนำกรอบงานนี้ไปประยุกต์ใช้กับภาษาอื่นนอกเหนือจากภาษาไทยได้

1.6 โครงสร้างเนื้อหาในวิทยานิพนธ์

บทที่ 1 กล่าวถึงบทนำ ซึ่งจะประกอบไปด้วยความเป็นมาและเหตุผลการวิจัย วัตถุประสงค์ของการวิจัย ขอบเขตการวิจัย ขั้นตอนการวิจัย ประโยชน์ที่คาดว่าจะได้รับ และโครงสร้างเนื้อหาในวิทยานิพนธ์

บทที่ 2 กล่าวถึงทฤษฎีที่เกี่ยวข้อง และงานวิจัยที่เกี่ยวข้อง

บทที่ 3 กล่าวถึงแนวคิดและวิธีการดำเนินงาน ซึ่งจะประกอบไปด้วยสภาพแวดล้อมและเครื่องมือ การรวมรวมชุดข้อมูล การสร้างคุณลักษณะ การออกแบบการทดลอง และการวัดประสิทธิภาพการจำแนกของแบบจำลอง

บทที่ 4 กล่าวถึงผลการทดลอง ซึ่งจะประกอบไปด้วยชุดข้อมูล และผลการทดลอง

บทที่ 5 กล่าวถึงบทสรุป ซึ่งจะประกอบไปด้วยสรุปและอภิปรายผลการทดลอง ปัญหาและอุปสรรค ข้อเสนอแนะ และผลงานวิจัยที่ได้รับการตีพิมพ์

บทที่ 2
ทฤษฎีและงานวิจัยที่เกี่ยวข้อง

ในบทนี้จะกล่าวถึงทฤษฎีและงานวิจัยที่เกี่ยวข้องที่นำมาประยุกต์ใช้กับการจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนสื่อสังคมออนไลน์ โดยได้ทำการค้นคว้า หลักการ ทฤษฎีและงานวิจัยที่เกี่ยวข้องเพื่อนำไปสู่แนวคิดและวิธีการดำเนินงาน ซึ่งทฤษฎีที่เกี่ยวข้อง ได้แก่ เฟซบุ๊ก เพศกับกลวิธีการตั้งชื่อในภาษาไทย การตัดคำภาษาไทย ชนิดของคำภาษาไทย เพื่อนบ้านที่ใกล้ที่สุด ซัพพอร์ตเวกเตอร์แมชชีน ป่าแบบสุ่ม นาอีฟเบย์ส โครงข่ายประสาทเทียม การวัดประสิทธิภาพการจำแนกของแบบจำลอง และงานวิจัยที่เกี่ยวข้องกับการจำแนกเพศจากชื่อผู้ใช้งานบนสื่อสังคมออนไลน์

2.1 ทฤษฎีที่เกี่ยวข้อง

2.1.1 เฟซบุ๊ก (Facebook)

ราชบัณฑิตยสถาน [13] ได้บัญญัติ คำว่า “เฟซบุ๊ก” หมายถึงเว็บไซต์ของบริษัทอเมริกันที่ชื่อว่า Facebook เฟซบุ๊กเป็นเว็บไซต์ที่ให้บริการสื่อสังคมออนไลน์เว็บไซต์หนึ่ง โดยเริ่มขึ้นในปีพุทธศักราช 2547 เมื่อนักศึกษามหาวิทยาลัยฮาร์วาร์ด (Harvard University) 3 คน เปิดบริการผ่านคอมพิวเตอร์ให้นักศึกษาในมหาวิทยาลัยของตนได้ติดต่อกัน ต่อมาสมาชิกก็ขยายวงออกไปเป็นนักศึกษาจากมหาวิทยาลัยอื่น นักเรียนชั้นมัธยม และในที่สุดก็เป็นประชาชนทั่วไป การเข้าเป็นสมาชิกเฟซบุ๊กไม่ต้องเสียค่าใช้จ่ายใด ๆ ผู้สมัครใหม่เพียงแต่สมัครเป็นสมาชิกตามเงื่อนไขที่กำหนด เว็บไซต์นี้มีบริการต่าง ๆ เช่น มีบริการเผยแพร่และรับข้อมูลส่วนบุคคลและข่าวสารต่าง ๆ และสามารถโต้ตอบกับสมาชิกรายอื่นได้

ผลการสำรวจของ “Global and Thailand digital report 2019” [3] พบว่า มีจำนวนผู้ใช้งานสื่อสังคมออนไลน์จำนวน 3,484 ล้านคนของจำนวนประชากรโลก โดยเป็นคนไทย 51 ล้านคน มีผู้ใช้งานกลุ่มใหญ่ คือ อายุ 18 – 24 ปี และ 25 – 34 ปี โดยคนไทยใช้เวลากับสื่อสังคมออนไลน์ เฉลี่ย 3 ชั่วโมง 11 นาที ต่อวัน ซึ่งเฟซบุ๊กเป็นสื่อสังคมออนไลน์ที่นิยมมากที่สุดในไทย

2.1.2 เพศกับกลวิธีการตั้งชื่อในภาษาไทย

ในสาขาวิชาอักษรศาสตร์ [14] [15] ได้วิเคราะห์เกี่ยวกับเพศกับกลวิธีการตั้งชื่อภาษาไทยโดยพบว่า นับตั้งแต่สมัยโบราณ ชื่อของคนไทยมีลักษณะของการแยกใช้ระหว่างเพศชายและเพศหญิง ในสมัยสุโขทัย พบว่า มักนิยมใช้ชื่อที่แสดงการลำดับเครือญาติ ตัวอย่างเช่น อ้าย ยี่ งั่ว ไส ดำ เป็นชื่อเพศชาย และ เอื้อย เป็นชื่อเพศหญิง คนไทยทางภาคเหนือและภาคตะวันออกเฉียงเหนือนิยมตั้งชื่อเพศชายและชื่อเพศหญิงให้แตกต่างกัน ตัวอย่างเช่น คำ หรือ จัน เป็นชื่อเพศชาย คำแว่น หรือ จันดี เป็นชื่อเพศหญิง โดยในปีพุทธศักราช 2554 จอมพล ป. พิบูลสงคราม ได้ออกประกาศสำนักนายกรัฐมนตรีเรื่องชื่อบุคคล เพื่อชักชวนให้ประชาชนนิยมตั้งชื่อให้สอดคล้องกับเพศ ธรรมเนียมการตั้งชื่อดังกล่าว จึงเป็นที่ยอมรับและเป็นแนวทางปฏิบัติเรื่อยมาจนปัจจุบัน

ลักษณะเฉพาะของการตั้งชื่อที่แตกต่างกันระหว่างเพศชายและเพศหญิง สามารถวิเคราะห์ได้ ดังนี้

1. ลักษณะทางพยางค์

จำนวนพยางค์ของชื่อเพศชายมีจำนวนพยางค์น้อยกว่าเพศหญิง แสดงถึงค่านิยมการใช้ชื่อเพศชายมีจำนวนพยางค์ที่น้อย เพราะมีลักษณะเรียบง่าย ไม่มีการตกแต่ง สมกับความเป็นชาย เช่น ศักดิ์ ชาติ เอก วุฒิ เปรม ชิต ชาญ ธร นพ พงษ์ พล โรจน์ ถนอม คมสัน ธีรวุฒิ บุญฤทธิ์ ส่วนชื่อเพศหญิงมีจำนวนพยางค์ที่มาก เพราะมีลักษณะของการตกแต่งมาก ทำให้ไพเราะ อ่อนหวาน สมกับความเป็นหญิง เช่น สุวิมล วาสนา ศรีสมร กุลสตรี สุพรรณิการ์ อังศุมาลิน พัชราภา วิชุลดา อาภัสสรา สุริวิภา พัชรีพร

2. ลักษณะทางความหมาย

ความหมายของชื่อเพศชายและเพศหญิง มีลักษณะที่แตกต่างกันออกไป โดยพบว่าความหมายของชื่อเพศชายนิยมความหมายเกี่ยวกับอำนาจ ความเจริญ ความยิ่งใหญ่ ความรู้ ความกล้าหาญ ความเข้มแข็ง และชัยชนะ ซึ่งแสดงถึงความเป็นผู้นำในสังคม เช่น นรบดี ณรงค์เดช ชนาธิป วีรยุทธ สุรฤทธิ์ พลรบ ส่วนความหมายของชื่อเพศหญิงนิยมความหมายเกี่ยวกับความงาม ความรัก ความอ่อนหวาน ดอกไม้ ซึ่งแสดงถึงความเป็นหญิง เช่น นุชนารถ สินีนาฏ นงเยาว์ สุดารัตน์ กุลสตรี กานดาวัลย์

3. ลักษณะทางเสียง

เสียงของชื่อเพศชายจะมีความแตกต่างไปจากชื่อเพศหญิง โดยชื่อของเพศชายจะมีการลงท้ายด้วยรูปสระเสียงสั้นหรือรัสสสระ ตามด้วยเสียงที่เกิดจากช่องว่างระหว่างเส้นเสียง / ? / (glottal stop) โดยไม่มีรูปพยัญชนะสะกด ทำให้เสียงดังกล่าวฟังดูหนักแน่นสมกับที่เป็นชื่อเพศชาย เป็นต้นว่า ลงท้ายด้วยรูปสระอะ เช่น มานะ วีระ วิริยะ ชัยชนะ ธีระ พยุหะ ธนะ ปิยะ ลงท้ายด้วยรูปสระอิ เช่น ปิติ กิตติ นิธิ สันติ นุติ ปริติ ปณิธิ และลงท้ายด้วยรูปสระอุ เช่น จิรายุ วายุ ภาณุ ถิรายุส์ วสุ ส่วนชื่อของเพศหญิงจะมีการลงท้ายด้วยรูปสระเสียงยาวหรือทีฆสระ ซึ่งจะทำให้ฟังดูอ่อนโยน ไม่แข็งกระด้าง สมกับที่เป็นชื่อเพศหญิง เป็นต้นว่า ลงท้ายด้วยรูปสระอา เช่น กานดา จิตรลดา นาตยา วีรยา วิรดา อสมา อัจจิมา ศรัณยา หัทยา ปัทมา ลงท้ายด้วยรูปสระอี เช่น สิริมณี วันดี วิชณี สิรี สุพรรณี อารีย์ ลงท้ายด้วยรูปสระอู เช่น ก้ามปู ยอดพธู ลงท้ายด้วยรูปสระออ เช่น ลออ และลงท้ายด้วยรูปสระประสม เช่น บุญเหลือ กลีบบัว

4. ลักษณะทางคำ

การสร้างคำของชื่อเพศชายมีลักษณะที่แตกต่างไปจากเพศหญิง ซึ่งพบว่า ชื่อของเพศชายมีการสร้างคำที่ซับซ้อนน้อยกว่าชื่อของเพศหญิง กล่าวคือ ชื่อของเพศชายมักจะนิยมคำโดด เช่น น้อย บุญ เทียน ส่วนชื่อของเพศหญิงมักจะนิยมคำผสม เช่น น้ำค้าง ซ่อนกลิ่น พรนภา และยังพบคำศัพท์ ที่มักจะประกอบอยู่ในชื่อของเพศชายและเพศหญิงที่ต่างกัน ซึ่งน่าจะเพราะความหมายของศัพท์เป็นสำคัญ

1) กลุ่มคำศัพท์ของชื่อเพศชาย

1.1) กลุ่มของศัพท์ที่มักปรากฏเป็นทั้งหน่วยหน้าและหน่วยหลังของชื่อเพศชาย ตัวอย่างเช่น

- เกียรติ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น เกียรติชัย เกียรติภูมิ เกียรติธร เกียรติวัฒน์ เกียรติพงศ์ สุรเกียรติ สมเกียรติ ศุภเกียรติ พงศ์เกียรติ ธีรเกียรติ วีรเกียรติ เป็นต้น
- พงศ์ หรือ พงษ์ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น พงษ์พัฒน์ พงศ์ศักดิ์ สรพงศ์ ธนพงศ์ พัฒนพงศ์ อิทธิพงศ์ สิทธิพงศ์ พงษ์พัฒน์ พงษ์สิทธ์ นิติพงษ์ เกียรติพงษ์ เป็นต้น
- พล ศัพท์นี้อาจประกอบเป็นชื่อ เช่น พลรบ พลพล พลวิทย์ พลวัฒน์ พลกฤษณ์ พลกฤช ปองพล ชุมพล จุมพล สมพล สุรพล ธีรพล อรรถพล ศักดิพล พัชรพล เป็นต้น

- ยุทธ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น ยุทธพล ยุทธการ ยุทธพิชัย ยุทธนันท์ ยุทธพงศ์ ยงยุทธ ชัยยุทธ สรยุทธ ศรายุทธ พิชัยยุทธ เป็นต้น

- วุฒิ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น วุฒิกร วุฒินันท์ วุฒิพงศ์ วุฒิธร วุฒิศักดิ์ ชาญวุฒิ เรืองวุฒิ กิตติวุฒิ สรวุฒิ ชาติวุฒิ ศุภวุฒิ ณัฐวุฒิ ธีรวุฒิ เป็นต้น

- ศักดิ์ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น ศักดิ์สิทธิ์ ศักดิ์ศรี ศักดิ์ชัย สมศักดิ์ สุรศักดิ์ เรืองศักดิ์ สิทธิศักดิ์ ธีรศักดิ์ ณรงค์ศักดิ์ ยุทธศักดิ์ กิตติศักดิ์ ยิ่งศักดิ์ เป็นต้น

1.2) กลุ่มของศัพท์ที่มักปรากฏเป็นหน่วยหน้าของชื่อเพศชาย ตัวอย่างเช่น

- กิตติ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น กิตติชัย กิตติภูมิ กิตติวัฒน์ กิตติธร กิตตินันท์ กิตติพงศ์ กิตติพล เป็นต้น

- ธีร ศัพท์นี้อาจประกอบเป็นชื่อ เช่น ธีรเกียรติ ธีรชัย ธีรวิทย์ ธีรวุฒิ ธีรพงศ์ ธีรวัฒน์ ธีรพัฒน์ ธีรชาติ เป็นต้น

- สุร ศัพท์นี้อาจประกอบเป็นชื่อ เช่น สุรเกียรติ สุรกานต์ สุรชัย สุรชา สุรนันท์ สุรฤทธิ์ สุรพันธ์ สุรพงศ์ สุรศักดิ์ สุรวัฒน์ สุรชัช สุรนาท สุรภาพ สุรยศ สุรทิน เป็นต้น

- อธิ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น อธิชัย อธิพงศ์ อธิพันธ์ อธิวัฒน์ อธิชาติ อธิโชค เป็นต้น

- อภิ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น อภิชัย อภิชาติ อภินันท์ อภิวัฒน์ อภิรักษ์ อภิศักดิ์ อภิวันท์ เป็นต้น

- เอก ศัพท์นี้อาจประกอบเป็นชื่อ เช่น เอกชัย เอกชาติ เอกพล เอกวัฒน์ เอกพงศ์ เอกพันธ์ เอกภพ เอกราชันย์ เอกบุตร เอกวิทย์ เอกพัฒน์ เอกรินทร์ เอกสิทธิ์ เอกณัฐ เป็นต้น

1.3) กลุ่มของศัพท์ที่มักปรากฏเป็นหน่วยหลังของชื่อเพศชาย ตัวอย่างเช่น

- พัฒน์ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น กิตติพัฒน์ จิรพัฒน์ ธนพัฒน์ วรพัฒน์ สุพัฒน์ สุรพัฒน์ พงศ์พัฒน์ ธนพัฒน์ เป็นต้น

- พันธ์ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น สุรพันธ์ จีรพันธ์ ธีรพันธ์ สุทธิพันธ์ วีรพันธ์ กิตติพันธ์ เป็นต้น

- วิทย์ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น ชาญวิทย์ ศุภวิทย์ ณัฐวิทย์ พัฒนวิทย์ สุรวิทย์ สรวิทย์ เรืองวิทย์ พงศ์วิทย์ เอกวิทย์ ธีรวิทย์ เจริญวิทย์ เป็นต้น

- วัฒน์ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น สุวัฒน์ กิตติวัฒน์ สุรวัฒน์ นันทวัฒน์ อนุวัฒน์ อธิวัฒน์ อภิวัฒน์ วีรวัฒน์ ธีรวัฒน์ ศุภวัฒน์ เป็นต้น

2) กลุ่มคำศัพท์ของชื่อเพศหญิง

2.1) กลุ่มของชื่อที่มักปรากฏเป็นทั้งหน่วยหน้าและหน่วยหลังของชื่อเพศหญิง ตัวอย่างเช่น

- พรรณ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น พรรณนิภา พรรณจิตร พรรณสิริ พรรณรัตน์ แพรวพรรณ พรพรรณ พิมพ์พรรณ อรพรรณ เป็นต้น

- วรรณ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น วรรณภา วรรณษา วรรณสิริ วรรณพร วรรณนิภา พัชรวรรณ โสมวรรณ วรวรรณ จุฑาวรรณ สิริวรรณ เป็นต้น

- ศรี ศัพท์นี้อาจประกอบเป็นชื่อ เช่น ศรีไพร ศรีวรรณ ศรีเมือง ศรีเรือน ศรีสวัสดิ์ สมศรี บุญศรี พรศรี เพ็ญศรี มารศรี พูนศรี เป็นต้น

- สิริ หรือ ศิริ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น สิริพร สิริวรรณ สิริมณี สิริลดา สิริพิชญ์ สิรินุช สิริรัตน์ สิริยาภา สิรินภา สิริธร สิริขวัญ สิรินันท์ สมสิริ รุ่งสิริ วรสิริ เป็นต้น

2.2) กลุ่มของศัพท์ที่มักปรากฏเป็นหน่วยหลังของชื่อเพศหญิง ตัวอย่างเช่น

- ทิพย์ ศัพท์นี้อาจประกอบเป็นชื่อ เช่น กรทิพย์ พรทิพย์ ภรณ์ทิพย์ วันทิพย์ ศิรินทิพย์ ปรางค์ทิพย์ แว่นทิพย์ ธารทิพย์ เนตรทิพย์ อ้อยทิพย์ น้ำทิพย์ เป็นต้น

- วดี ศัพท์นี้อาจประกอบเป็นชื่อ เช่น จิตรวดี ทิพาวดี อายุวดี ขวัญวดี ดาราวดี กุลวดี สิทธิวดี ภัทราวดี นิตยาวดี ภาวดี เป็นต้น

นอกจากชื่อของเพศชายและเพศหญิง ซึ่งมีลักษณะเฉพาะที่แตกต่างกันแล้ว ยังมีชื่อกลาง ๆ หรือชื่อที่สามารถใช้ร่วมกันทั้ง 2 เพศ เช่น แก้ว ตา ทอง วิรัตน์ สุวรรณ สมพร ทองดี บุญมา เป็นต้น

อย่างไรก็ตาม แม้ว่าจะสามารถแยกแยะชื่อเป็น 3 กลุ่ม คือ ชื่อเพศชาย ชื่อเพศหญิง และชื่อกลาง ๆ แต่ก็ยังพบชื่อเพศชายที่น่าจะเป็นชื่อเพศหญิง และชื่อเพศหญิงที่น่าจะเป็นชื่อเพศชาย ยกตัวอย่างชื่อเพศชาย เช่น อารี อุบล อุไร อาภรณ์ อมรรัตน์ สมฤทัย สำอาง เป็นต้น ชื่อเพศหญิง เช่น พัฒนา นิยม เจษฎา มงคล สมบัติ สมศักดิ์ อนุสรณ์ เอกลักษณ์ เป็นต้น

2.1.3 การตัดคำภาษาไทย

การตัดคำ [16] คือการแบ่งตัวอักษรจากข้อความที่ต่อเนื่องกัน เพื่อหาขอบเขตของแต่ละหน่วยคำ (Morpheme) ซึ่งลักษณะการเขียนภาษาไทยจะเขียนติดต่อกันเป็นสายอักขระ โดยไม่มีเครื่องหมายวรรคตอนแสดงการแบ่งคำเหมือนภาษาอังกฤษ ยกเว้นแต่มีการเว้นวรรคเป็นระยะ ๆ เพื่อให้ผู้อ่านได้หยุดพัก และทำความเข้าใจความหมายเป็นตอน ๆ ไปเท่านั้น จึงเป็นอุปสรรคอย่างหนึ่ง โดยแรกเริ่มเป็นการทำเพื่อช่วยงานพิมพ์ภาษาไทยบนคอมพิวเตอร์ ในการปัดข้อความขึ้นบรรทัดใหม่ให้ถูกต้อง ไม่ตัดแยกส่วนคำระหว่างบรรทัด ซึ่งวิธีการต่าง ๆ ที่จะช่วยให้คอมพิวเตอร์รู้จักคำในภาษาไทย โดยในปัจจุบันมีอยู่หลายแนวคิด [17] ตัวอย่างเช่น

1. หลักการตัดคำโดยใช้กฎ (Rule Based Approach)

การตัดคำโดยใช้กฎด้วยการพยายามสร้างหลักไวยกรณ์ (Grammar) ให้กับภาษาไทย โดยกำหนดลักษณะการประสมอักษรลักษณะ การเว้นวรรค และการขึ้นย่อหน้า เพื่อใช้เป็นเกณฑ์ในการกำหนดขอบเขตของคำ ซึ่งวิธีการนี้จะมีข้อจำกัดในการทำงาน คือ ความถูกต้องของการตัดคำในระดับพยางค์สูง แต่ความถูกต้องของการตัดคำค่อนข้างต่ำ โดยข้อดีของวิธีนี้คือ มีความรวดเร็วในการทำงานและการใช้ทรัพยากรน้อย ตัวอย่างเช่น “การพัฒนาระบบถาม-ตอบ” จะสามารถตัดคำได้ว่า “การ|พัฒ|นา|ระ|บบ|ถาม|-|ตอบ”

2. หลักการตัดคำโดยใช้พจนานุกรม (Dictionary Approach)

การตัดคำโดยใช้พจนานุกรม คือ การนำข้อความไปค้นหาและเทียบกับคำในพจนานุกรม เพื่อหาว่าข้อความดังกล่าว ควรตัดคำที่บริเวณใด และประกอบด้วยคำใดบ้าง ซึ่งวิธีการนี้เป็นการตัดคำวิธีที่ง่าย รวดเร็ว และมีความถูกต้องสูง แต่ใช้เวลามากกว่า และเป็นไปไม่ได้ที่จะเก็บ คำทุกคำในพจนานุกรม โดยเฉพาะคำวิสามานยนาม เช่น ชื่อคน ชื่อสถานที่ หรือคำที่เกิดใหม่ และยังสิ้นเปลืองทรัพยากรค่อนข้างมาก ตัวอย่างเช่น “การพัฒนาระบบถาม-ตอบ” จะสามารถตัดคำได้ว่า “การ|พัฒนา|ระบบ|ถาม|-|ตอบ” ซึ่งหลักการตัดคำโดยใช้พจนานุกรมสามารถตัดคำได้ถูกต้องมากกว่าการใช้หลักการตัดคำโดยใช้กฎ และยังมีวิธีการตัดคำอื่น ๆ ร่วมกับหลักการตัดคำโดยใช้พจนานุกรม คือ

1) วิธีการตัดคำแบบยาวที่สุด (Longest Matching) คือ การตัดคำให้ยาวที่สุดก่อน โดยเริ่มจากตัวอักษรซ้ายสุดของข้อความนับไปยังตัวอักษรถัดไป จนกว่าจะพบว่าเป็นคำที่มีอยู่ในพจนานุกรม โดยเลือกคำที่พบที่ยาวที่สุด แล้วค้นหาคำถัดไปจนกว่าจะจบข้อความ ตัวอย่างเช่น คำว่า “กองกลาง” เมื่อนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า “ก”, “กอ” และคำว่า “กอง” ส่วนคำว่า “กองก” ไม่พบอยู่ในพจนานุกรม ดั้งนั้น จึงได้คำว่า “กอง” ซึ่งเป็นคำที่ยาวที่สุดที่หาพบ ส่วนที่เหลือเมื่อนำไปค้นในพจนานุกรมจะได้ว่า “ก”, “กล” และ “กลาง” ดังนั้น จึงเลือกคำว่า “กลาง” ทำให้สามารถตัดคำได้เป็น “กอง|กลอง” แต่ก็อาจจะพบว่ามีปัญหา ตัวอย่างเช่น “ไปหามเหสี” ตัดคำออกมาได้เป็น “ไป|หาม|เห|สี”

2) วิธีการตัดคำแบบสั้นที่สุด (Shortest Matching) คือ การตัดคำให้สั้นที่สุดก่อน ซึ่งพบว่าจะได้จำนวนมากที่สุด แต่ความถูกต้องน้อยกว่าวิธีการตัดคำแบบยาวที่สุด ตัวอย่างเช่น คำว่า “โคลงเรือ” เมื่อนำไปเปรียบเทียบกับคำที่มีอยู่ในพจนานุกรมจะพบคำว่า “โค”, “โคล” และคำว่า “โคลง” ดั้งนั้น จึงได้คำว่า “โค” ซึ่งเป็นคำที่สั้นที่สุดที่หาพบ ส่วนที่เหลือเมื่อนำไปค้นในพจนานุกรม ทำให้สามารถตัดคำได้เป็น “โค|ลง|เรือ”

3) วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) คือ การตัดคำให้ได้จำนวนคำน้อยที่สุด ตัวอย่างเช่น “ไปหามเหสี” การตัดคำที่สามารถเป็นไปได้ คือ “ไป|หาม|เห|สี” และ “ไป|หา|มเหสี” ซึ่งในกรณีนี้จะเลือก “ไป|หา|มเหสี” เพราะมีจำนวนคำที่ได้น้อยกว่า

4) วิธีการย้อนรอยกลับ (Backtracking) คือ เมื่อทำการเปรียบเทียบคำที่นำมาตัดคำโดยใช้พจนานุกรม อาจพบกรณีที่คำที่พบมีมากกว่า 1 คำ แล้วทำการเลือกคำที่ยาวที่สุด ทำให้คำที่เหลือไม่สามารถตัดคำได้ เนื่องจากไม่พบในพจนานุกรม ซึ่งในกรณีนี้จะทำการย้อนไปยังคำที่ไม่ถูกเลือกแล้วทำการตัดคำต่อไป ตัวอย่างเช่น “เมื่อยามนี้” เมื่อใช้วิธีการตัดคำแบบยาวที่สุด จะได้คำว่า “เมื่อย” ส่วนคำที่เหลือคือ “-ามนี้” ซึ่งไม่พบอยู่ในพจนานุกรม ดังนั้น จึงทำการย้อนกลับไปเพื่อเลือกอีกคำหนึ่ง คือ “เมื่อ” ทำให้สามารถตัดคำได้เป็น “เมื่อ|ยาม|นี้” โดยคำว่า “ยาม” เกิดจากการเลือกคำที่ยาวที่สุดระหว่าง “ยา” กับ “ยาม”

ซึ่งวิธีการตัดคำแบบสอดคล้องมากที่สุดให้ผลลัพธ์ที่ดีกว่า เป็นเพราะภาษาไทยมีการสร้างคำจากการประสมคำจำนวนมาก การเลือกผลที่มีจำนวนคำน้อยสุดจึงมีโอกาสถูกต้องมากกว่า แต่ข้อเสียของการตัดคำด้วยพจนานุกรม คือ ถ้าพบรูปคำที่ไม่รู้จัก จะไม่สามารถตัดคำได้ ซึ่งอาจ

มาจากพจนานุกรมที่มีรายการคำครอบคลุมไม่เพียงพอ หรือเป็นคำที่เกิดใหม่ เป็นชื่อต่าง ๆ เป็นคำทับศัพท์ หรืออาจเป็นคำที่สะกดผิด และยังพบปัญหาเรื่องความกำกวม โดยเฉพาะเมื่อตัดแล้วได้จำนวนคำเท่ากัน ตัวอย่างเช่น “ตาก|ลม” กับ “ตา|กลม” และยังเสียเวลาและสิ้นเปลืองทรัพยากรที่มากกว่า

3. หลักการตัดคำโดยใช้คลังข้อมูล (Corpus Based Approach)

การตัดคำโดยใช้วิธีการทางสถิติและคลังข้อมูลทางภาษา (Corpus) เพื่อแก้ปัญหาของคำที่ไม่มีในพจนานุกรม เช่น ชื่อเฉพาะ หรือคำที่มาจากภาษาต่างประเทศ แบ่งออกเป็น 3 วิธี

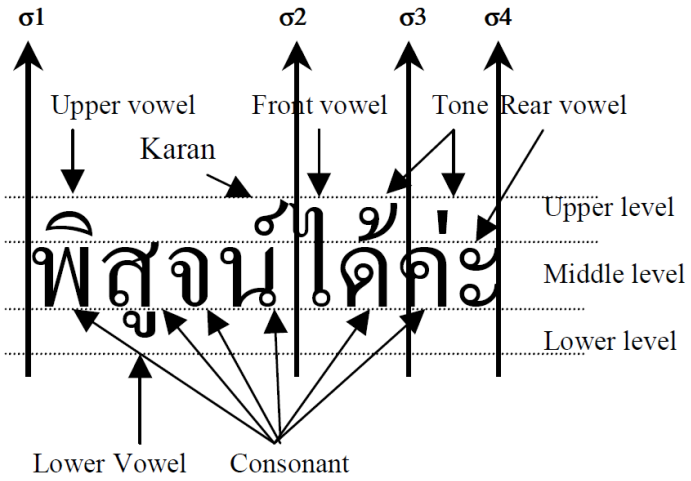
1) วิธีการตัดคำโดยอาศัยความน่าจะเป็น (Probabilistic Word Segmentation) คือ การตัดคำโดยนำค่าสถิติการเกิดของคำ และลำดับของหน้าที่ของคำ (Part of Speech) เข้ามาช่วยคำนวณหาความน่าจะเป็น เพื่อที่จะเลือกแบบที่มีโอกาสเกิดมากที่สุด เป็นวิธีการที่ช่วยแก้ไขปัญหาความกำกวม โดยใช้การวิเคราะห์ความถี่ ตัวอย่างเช่น “ก็อด” สามารถตัดคำได้ว่า “ก็|อด” ซึ่งพบว่ามีความถี่มากกว่าคำว่า “ก็อด” แต่มีข้อจำกัดที่จะต้องมีฐานข้อมูลที่มีการตัดคำที่ถูกต้องและมีการกำหนดหน้าที่ของคำ เพื่อที่จะนำไปใช้ในการสร้างสถิติ

2) วิธีการตัดคำแบบคุณลักษณะ (Feature-based Word Segmentation) คือ การตัดคำโดยพิจารณาจากบริบท การเกิดร่วมกันของคำ หรือหน้าที่ของคำเข้ามาช่วย เพื่อเลือกการตัดคำที่ดีสุด จากทุกแบบที่ได้จากการใช้การตัดคำแบบสอดคล้องมากที่สุดมาก่อน ตัวอย่างเช่น “นั่งตากลมแป๋ว” ในบริบทพบคำว่า “แป๋ว” ทำให้จะสามารถตัดคำได้ว่า “ตา|กลม” ซึ่งวิธีการนี้จำเป็นที่จะต้องมีฐานข้อมูลเป็นจำนวนมาก และจะต้องมีการเรียนรู้การสร้างคำในบริบท หรือการเกิดร่วมกันของคำแต่ละคำ เพื่อให้มีข้อมูลที่จะนำมาใช้ในการตัดคำ

3) เอ็น-แกรม (N-Gram) คือ แบบจำลองที่ใช้คำนวณค่าความน่าจะเป็นของสายอักขระ (Character Sequence) ที่เกิดขึ้นร่วมกันเป็นคำ หรือค่าความน่าจะเป็นของสายคำ (Word Sequence) ที่เกิดขึ้นร่วมกันเป็นประโยค โดยค่าความน่าจะเป็นของสายอักขระหรือคำประมาณได้จากคลังข้อมูลที่สร้างไว้ โดยแกรม คือ หน่วยที่ใช้ในการสร้างแบบจำลอง อาจจะเป็นเสียงอักขระหรือคำก็ได้ และแกรมมีได้หลายขนาดแล้วแต่จะกำหนด ตั้งแต่ 1 จนถึงเอ็น (N) เช่น 2-แกรม (Bigram), 3-แกรม (Trigram), 4-แกรม (Quadigram) เป็นต้น ซึ่งเป็นวิธีที่ไม่ต้องอาศัยหลักความรู้ทางภาษาศาสตร์ และไม่สนใจความหมายของคำ

4. หลักการแยกหน่วยย่อยก่อนการตัดคำ (Thai Character Cluster: TCC)

การตัดคำโดยใช้กลุ่มของตัวอักษรไทยที่ไม่สามารถแยกจากกันได้ตามหลักการเขียนของภาษาไทยโดยไม่อิงพจนานุกรม [18] โดยคำในภาษาไทยมีอักขระหลากหลายประเภท ทั้งพยัญชนะ สระ วรรณยุกต์ และตัวอักษรพิเศษอื่น ๆ และในภาษาไทยยังมีระดับของตัวอักษรอยู่หลายระดับ ได้แก่ ระดับบน ระดับกลาง และระดับล่าง ดังภาพที่ 1



ภาพที่ 1 ตัวอย่างอักขระภาษาไทย

อักขระภาษาไทย ประกอบไปด้วยอักขระ 7 ประเภท ได้แก่ พยัญชนะ สระด้านบน สระด้านล่าง สระด้านหน้า สระด้านหลัง วรรณยุกต์ และการันต์ โดยการตัดคำแบบแยกหน่วยย่อยก่อนการตัดคำนี้ จะกำหนดโดยใช้กฎ ยกตัวอย่างเช่น สระด้านหน้าและพยัญชนะตัวถัดไปต้องจัดเป็นกลุ่มเดียวกัน เช่น “เ-” จะไม่ตัดหลังสระเอ เครื่องหมายวรรณยุกต์จะต้องอยู่ด้านบนพยัญชนะเสมอ เช่น “ก่” จะไม่ตัดก่อนไม้เอก และไม่สามารถแยกออกจากพยัญชนะได้ สระด้านหลังและพยัญชนะก่อนหน้าต้องจัดเป็นกลุ่มเดียวกัน เช่น “-า” จะไม่ตัดหน้าสระอา โดยมักใช้วิธีนี้รวมกับวิธีเรียนรู้จากคลังข้อมูลอื่น ๆ

2.1.4 ชนิดของคำภาษาไทย

ชนิดของคำ (Part of Speech) หมายถึง ประเภทของคำในทางภาษาศาสตร์ โดยเกณฑ์การจำแนกชนิดของคำภาษาไทยมีหลายเกณฑ์ จำนวนชนิดของคำก็แตกต่างกันออกไป จากตำราหลักภาษาไทยของพระยาอุปกิตศิลปะสาร [19] แบ่งคำในภาษาไทยออกเป็น 7 ชนิด คือ คำนาม

คำสรรพนาม คำกริยา คำวิเศษณ์ คำบุพบท คำสันธาน และคำอุทาน ส่วนหนังสืออุเทศภาษาไทย ชุด บรรทัดฐานภาษาไทย เล่ม 3 [20] แบ่งคำในภาษาไทยออกเป็น 12 ชนิด คือ คำนาม คำสรรพนาม คำกริยา คำช่วยกริยา คำวิเศษณ์ คำเกี่ยวกับจำนวน คำบอกกำหนด คำบุพบท คำเชื่อม คำลงท้าย คำอุทาน และคำปฏิเสธ โดยใช้เกณฑ์หน้าที่ ตำแหน่งที่คำปรากฏและสัมพันธ์กับคำอื่นและความหมายประกอบกัน ซึ่งคำแต่ละชนิดจะมีลักษณะและหน้าที่แตกต่างกันออกไป ได้แก่

1. คำนาม คือ คำที่แสดงความหมายถึงบุคคล สัตว์ วัตถุ สิ่งของ สภาพ อาการ สถานที่ ทั้งสิ่งมีชีวิตและไม่มีชีวิต ทั้งที่เป็นรูปธรรมและนามธรรม

คำนามแบ่งออกเป็น 4 ชนิด ดังนี้

1) คำนามสามัญ คือ คำนามที่เรียกสิ่งต่าง ๆ โดยทั่วไป เป็นคำเรียกสิ่งต่าง ๆ โดยไม่ชี้เฉพาะเจาะจง เช่น ปลา คน สุนัข วัด ต้นไม้ หนังสือ ปากกา

2) คำนามวิสามัญ คือ คำนามที่ใช้เรียกชื่อเฉพาะที่สมมติตั้งขึ้น เรียก คน สัตว์ สิ่งของ หรือสถานที่ เช่น พระพุทธชินราช เด็กชายวิทวัส จังหวัดพิจิตร เดือนมกราคม

3) คำลักษณนาม คือ คำนามที่บอกลักษณะของคำนาม เพื่อแสดงรูปลักษณะ ขนาด ปริมาณ ของคำนามนั้นให้ชัดเจน เช่น บ้าน 1 หลัง โต๊ะ 5 ตัว ลูกคนนี้ ขอกอดที

4) คำอาการนาม คือ คำนามที่เป็นชื่อกริยาอาการ เป็นสิ่งที่เป็นนามธรรม มักมีคำว่า "การ" หรือ "ความ" นำหน้า เช่น การกิน การเดิน การพูด ความรัก ความดี ความจำ

2. คำสรรพนาม คือ คำที่ใช้แทนนาม เพื่อจะได้ไม่ต้องกล่าวคำนามนั้นซ้ำอีก

คำสรรพนามแบ่งออกเป็น 5 ชนิด ดังนี้

1) คำบุรุษสรรพนาม คือ คำสรรพนามที่ใช้ระบุแทนบุคคล เพื่อบอกว่าเป็นผู้พูด แบ่งเป็นชนิดย่อยได้ 3 ชนิด คือ

1.1) สรรพนามบุรุษที่ 1 ใช้แทนผู้พูด เช่น ผม ฉัน เรา กระผม ข้าพเจ้า อาตมา

1.2) สรรพนามบุรุษที่ 2 ใช้แทนผู้ฟัง เช่น คุณ เธอ ท่าน ฝ่าบาท ใต้ฝ่าพระบาท

1.3) สรรพนามบุรุษที่ 3 ใช้แทนผู้ที่กล่าวถึง เช่น เขา มัน เธอ ท่าน พระองค์

2) คำสรรพนามถาม คือ คำสรรพนามที่ใช้แทนนามที่เป็นคำถาม เช่น อะไร ใคร ไหน ตัวอย่างเช่น น้องกำลังทำอะไรอยู่ ใครทำแก้วแตก เขาไปที่ไหน

3) คำสรรพนามชี้เฉพาะ คือ คำสรรพนามที่ใช้แทนนามชี้เฉพาะเจาะจงหรือบอกกำหนดความให้ผู้พูดกับผู้ฟังเข้าใจกัน เช่น นี่ นั่น โน่น ตัวอย่างเช่น นั่งนี่ไหม ร่มอยู่นั่นไง ดูโน่นซิ

4) คำสรรพนามไม่ชี้เฉพาะ คือ คำสรรพนามที่ไม่ระบุหรือกำหนดแน่นอนว่าหมายถึงผู้ใด สิ่งใด เช่น อะไร ใคร ไหน บางครั้งก็เป็นคำซ้ำ ๆ เช่น ใคร ๆ อะไร ๆ ไหน ๆ

5) คำสรรพนามแยกฝ่าย คือ คำสรรพนามที่ใช้แทนนาม ซึ่งแสดงให้เห็นว่านามนั้นมีหลายส่วน เช่น ต่าง บ้าง กัน ตัวอย่างเช่น นักเรียนบ้างเรียนบ้างเล่น นักเรียนต่างก็อ่านหนังสือ

3. คำกริยา คือ คำที่แสดงอาการ การกระทำ บอกสภาพของนาม หรือสรรพนาม โดยคำกริยามี 2 ประเภท คือ คำกริยาที่มีหน่วยกรรม และคำกริยาที่ไม่มีหน่วยกรรม

1) คำกริยาที่มีหน่วยกรรม แบ่งเป็น 2 ชนิด คือ

1.1) คำกริยาสกรรม คือ คำกริยาที่มีนามตามหลังซึ่งทำหน้าที่เป็นกรรม เช่น กิน ฟัง เห็น อ่าน ตัวอย่างเช่น ฉันกินข้าว พ่อฟังข่าว เขาเห็นนก

1.2) คำกริยาทวิกรรม คือ คำกริยาที่มีนาม 2 นามตามหลัง นามแรกทำหน้าที่กรรมตรง นามที่สองทำหน้าที่กรรมรอง เช่น สอน ป้อน ให้ แจก อบรม ตัวอย่างเช่น เขาสอนหนังสือพี่ป้อนข้าวน้อง แม่ให้เงินลูก

2) คำกริยาที่ไม่มีหน่วยกรรม แบ่งเป็น 2 ชนิด คือ

2.1) คำกริยาอกรรม คือ คำกริยาที่ไม่ต้องมีนามทำหน้าที่เป็นกรรมมาเติมเต็ม เช่น ยืน นอน เสียใจ ตัวอย่างเช่น เขายืนอยู่ น้องนอน เราเสียใจ

2.2) คำกริยาคุณศัพท์ คือ คำกริยาที่ไม่ต้องมีนามที่เป็นกรรมมาเติมเต็ม เช่น ดี สวย สูง ตัวอย่างเช่น เด็กคนนี้ดี บ้านแถวนี้สวยทุกหลัง เขาสูงขึ้นมาก

2.3) คำกริยาต้องเติมเต็ม คือ คำกริยาที่ต้องมีนามมาเติมเต็มเสมอ เช่น เป็น เหมือน คล้าย คือ ตัวอย่างเช่น เขาเป็นนักเรียน เธอเหมือนพ่อมาก เขาคือครูของฉันเอง

2.4) คำกริยาที่นำ คือ คำกริยาที่อยู่หน้าคำกริยาอื่นเสมอ เช่น ชอบ หัด อยาก ตัวอย่างเช่น เขาชอบเป็นหวัด เด็กหัดขี่จักรยาน ฉันอยากพักผ่อนมาก

2.5) คำกริยาตาม คือ คำกริยาที่อยู่ตามหลังกริยาอื่นเสมอ เช่น ไป มา ขึ้น ลง ตัวอย่างเช่น เขาส่งพัสดุไปแล้ว ลูกโป่งลอยขึ้น น้ำลดลง

4. คำช่วยกริยา คือ คำที่ไม่ใช่คำกริยาและไม่อยู่ตามลำพัง แต่จะอยู่ร่วมกับคำกริยา เช่น จะ กำลัง โดน ตัวอย่างเช่น ฉันจะไปหาเขา ผมกำลังอาบน้ำอยู่ สุนัขโดนขัง

5. คำวิเศษณ์ คือ คำที่ใช้ขยายคำกริยาอื่นเพื่อให้ได้ใจความชัดเจนยิ่งขึ้น ซึ่งมักอยู่ตามหลังคำกริยา

คำวิเศษณ์แบ่งออกเป็น 4 ชนิด ดังนี้

1) คำวิเศษณ์สามัญ คือ คำวิเศษณ์ที่ขยายคำกริยาโดยทั่วไป เช่น จัง แล้ว เอง ตัวอย่างเช่น วันนี้รถติดจัง น้องนอนแล้ว แม่ทำกับข้าวเอง

2) คำวิเศษณ์ขยายเฉพาะ คือ คำวิเศษณ์ที่ขยายคำกริยาคำใดคำหนึ่งโดยเฉพาะ เช่น แจ๋ ดิ๊ก กริบ ตัวอย่างเช่น แดงแจ๋ กลมดิ๊ก คมกริบ

3) คำวิเศษณ์แสดงคำถาม คือ คำวิเศษณ์ที่ใช้แสดงคำถามเกี่ยวกับการกระทำ เช่น ทำไม เมื่อไร ทำไม ตัวอย่างเช่น ทำไมสัตว์ป่าจึงสูญพันธ์ เครื่องบินลงเมื่อไร ทำไมปิดวิทยุ

4) คำวิเศษณ์บอกเวลา คือ คำวิเศษณ์ที่ใช้บ่งบอกเวลาที่เกิดเหตุการณ์ บอกจำนวนหรือปริมาณ เช่น พรุ่งนี้ สาย ตะกี้ ตัวอย่างเช่น พรุ่งนี้เป็นวันเกิดของคุณแม่ เขามาโรงเรียนสาย ใครโทรศัพท์มาตะกี้

6. คำที่เกี่ยวกับจำนวน

คำที่เกี่ยวกับจำนวนแบ่งออกเป็น 4 ชนิด ดังนี้

1) คำบอกจำนวน คือ คำที่มีความหมายถึงจำนวน เช่น 1 เจ็ด ครึ่ง ตัวอย่างเช่น เขามีลูก 1 คน เขาเที่ยวทั่วแล้วทั้งเจ็ดทวีป ไก่ย่างครึ่งตัว

2) คำบอกลำดับ คือ คำที่แสดงลำดับที่ เช่น ที่หนึ่ง ที่โหล่ คนเดียว ตัวอย่างเช่น ฉันได้รางวัลที่หนึ่ง เขาสอบได้ที่โหล่ เขามีลูกคนเดียว

3) คำหน้าจำนวน คือ คำที่อยู่หน้าคำบอกจำนวน เช่น อีก ตั้ง ราว ๆ ตัวอย่างเช่น ฉันจะซื้อเสื้ออีก 2 ตัว เขามีลูกตั้ง 5 คน มีคนประชุมราว ๆ 200 คน

4) คำหลังจำนวน คือ คำที่อยู่หลังคำบอกจำนวน เช่น กว่า ถ้วน พอดี ตัวอย่างเช่น มีคนมาชุมนุมพันกว่าคน เขาต้องจ่ายค่าไฟ 900 บาทถ้วน ไก่ตัวนี้หนัก 2 กิโลพอดี

7. คำบอกกำหนด คือ คำขยายนามที่อยู่ท้ายสุดในนาม

คำบอกกำหนดแบ่งออกเป็น 4 ชนิด ดังนี้

1) คำบอกกำหนดชี้เฉพาะ คือ คำที่ขยายนามเพื่อชี้เฉพาะว่านามนี้อยู่ตำแหน่งไหน เช่น นี่ นั่น โน่น นี้ นั้น โน้น ตัวอย่างเช่น เสื้อนั่นไม่ได้ซัก บ้านหลังนั้นไม่มีคนอยู่ โต๊ะตัวนู้นราคาแพง

2) คำบอกกำหนดไม่ชี้เฉพาะ คือ คำที่ขยายนามที่ไม่ชี้เฉพาะว่านามนี้อยู่ตำแหน่งไหน เช่น อื่น ต่าง ๆ ไหน ตัวอย่างเช่น รถอื่นเข้าอู่หมด ธัญพืชต่าง ๆ มีประโยชน์ บ้านไหนถูกโจรขึ้น

8. คำบุพบท คือ คำที่อยู่หน้านามเพื่อบอกตำแหน่ง หน้าที่ ความเกี่ยวข้อง ความเป็นเจ้าของเพื่อบอกความสัมพันธ์ระหว่างนามหรือกริยากับนาม เช่น ของ ใน บน ตัวอย่างเช่น ปากกาของฉัน ต้นไม้ในสวน เขานอนบนโซฟา

9. คำเชื่อม คือ คำที่ใช้เชื่อมคำ วลี หรือประโยคเข้าด้วยกัน

คำเชื่อมแบ่งออกเป็น 4 ชนิด ดังนี้

1) คำเชื่อมสมภาค คือ คำเชื่อมที่ใช้เชื่อมหน่วยภาษาเดียวกัน เช่น กับ หรือ แต่ ตัวอย่างเช่น คุณกับผม จะดูหนังหรือฟังเพลง น้องชอบดูฟุตบอลแต่พี่ชอบดูเทนนิส

2) คำเชื่อมอนุประโยค คือ คำที่นำหน้าอนุประโยคในประโยคซ้อน เช่น ว่า ที่ เพราะ ตัวอย่างเช่น เขาพูดว่าเขาจะมา เด็กที่ได้รับรางวัล เขาไม่มาทำงานเพราะฝนตก

3) คำเชื่อมเสริม คือ คำเชื่อมที่เพิ่มขึ้นเพื่อให้มีความสัมพันธ์ชัดเจนขึ้น เช่น จึง เลย ถึง ก็ ตัวอย่างเช่น เหตุใดจึงมานั่งเศร้าอยู่คนเดียว ทำไมถึงยังไม่กินข้าว ก็ใช่

4) คำเชื่อมสัมพันธสาร คือ คำเชื่อมประโยคตั้งแต่ 2 ประโยคขึ้นไปให้รวมกัน เช่น กล่าวคือ อย่างไรก็ตาม ในที่สุด

10. คำลงท้าย คือ คำที่อยู่ในตำแหน่งท้ายสุดของประโยค ไม่มีความหมายเด่นชัด และไม่สัมพันธ์กับส่วนใดส่วนหนึ่งของประโยค

คำลงท้ายแบ่งออกเป็น 2 ชนิด ดังนี้

1) คำลงท้ายแสดงทัศนภาวะ คือ คำลงท้ายที่แสดงเจตนาหรือความรู้สึกด้วยการออกเสียง เช่น ล่ะ ซิ่ นะ ตัวอย่างเช่น เงินทอนล่ะ เข้ามาซิ่ พรุ่งนี้วันหยุดนะ

2) คำลงท้ายแสดงมารยาท คือ คำลงท้ายที่แสดงความสุภาพ และอาจแสดงความสัมพันธ์ระหว่างผู้พูดด้วย เช่น ครับ ค่ะ ขา ตัวอย่างเช่น มาแล้วครับ ฉันไม่สบายค่ะ คุณพ่อขา

11. คำอุทาน คือ คำที่เปล่งออกมาเพื่อแสดงอารมณ์ความรู้สึกต่าง ๆ ได้แก่ สะเทือนใจ ตกใจ ดีใจ เห็นใจ เจ็บปวด มักมีเครื่องหมายอัศเจรีย์ (!) กำกับหลัง เช่น ว้าย ไชโย ตายจริง ตัวอย่างเช่น ว้าย! ช่วยด้วย ไชโย! ฉันสอบได้ที่หนึ่ง ตายจริง! ลืมของอีกแล้ว

12. คำปฏิเสธ คือ คำที่ใช้บอกปัดหรือไม่ยอมรับ

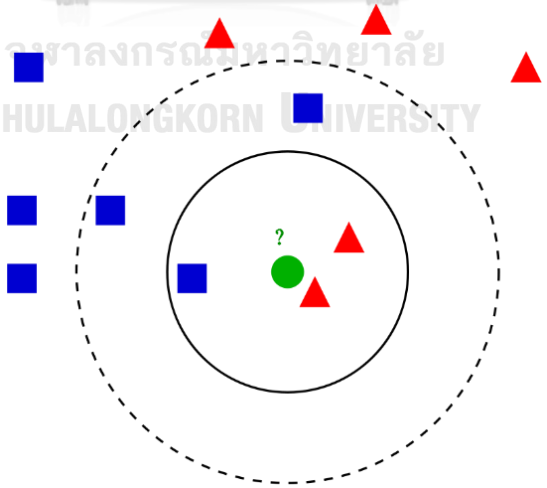
คำปฏิเสธแบ่งออกเป็น 2 ชนิด ดังนี้

1) คำปฏิเสธกริยา เช่น มิ ไม่ หาไม่ หา...ไม่ ตัวอย่างเช่น ผมมิใช่คนดี เขาเดินไม่ระวัง เขาจะสำนึกผิดก็หาไม่

2) คำปฏิเสธข้อความ เช่น หามิได้ มิได้ เปล่า

2.1.5 เพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor)

เพื่อนบ้านที่ใกล้ที่สุด (K-Nearest Neighbor) เป็นการเรียนรู้ของเครื่อง [21] โดยการจัดข้อมูลที่อยู่ใกล้กันให้เป็นกลุ่มเดียวกัน โดยใช้การหาระยะห่างระหว่างแต่ละตัวแปร (Attribute) ซึ่ง k แทนจำนวนของข้อมูลในชุดข้อมูล ดังภาพที่ 2



ภาพที่ 2 การจำแนกข้อมูลด้วยวิธีเพื่อนบ้านที่ใกล้ที่สุด

ขั้นตอนวิธีการเพื่อนบ้านใกล้ที่สุดมีวิธีการที่ไม่ซับซ้อน เข้าใจง่าย แต่จะใช้ระยะเวลาในการคำนวณนาน ถ้าตัวแปรมีจำนวนมาก

ขั้นตอนการจำแนกข้อมูลแบบเพื่อนบ้านที่ใกล้ที่สุด มีดังนี้

1. กำหนดขนาดของ k

2. คำนวณระยะระหว่างข้อมูล โดยการคำนวณระยะทางสำหรับข้อมูลที่มีค่าต่อเนื่องมักใช้ Euclidean distance ดังสมการ 2.1

$$distance(X,Y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2} \quad (2.1)$$

ส่วนข้อมูลที่มีค่าไม่ต่อเนื่อง เช่น การจำแนกประเภทข้อความ มักใช้ Hamming distance ดังสมการ 2.2

$$distance(X,Y) = \sum_{i=1}^{n}|x_i - y_i| \quad (2.2)$$

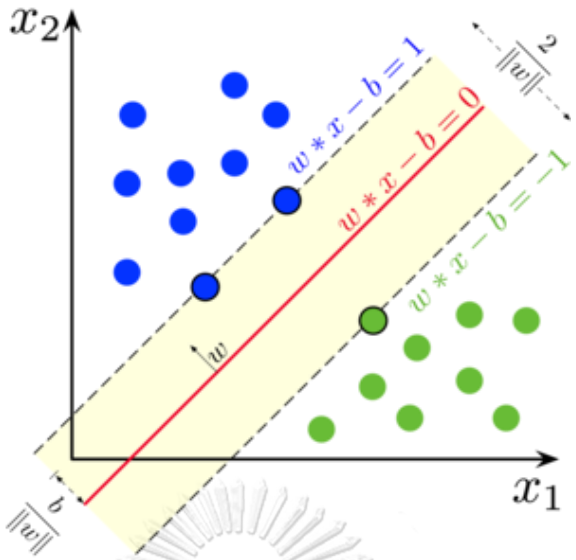
3. จัดเรียงลำดับของระยะห่าง และเลือกพิจารณาชุดข้อมูลที่ใกล้จุดที่ต้องการพิจารณาตามจำนวน k ที่กำหนดไว้

4. พิจารณาข้อมูลจำนวน k ชุด และสังเกตว่ากลุ่มไหนที่ใกล้จุดที่พิจารณาเป็นจำนวนมากที่สุด

5. กำหนดกลุ่มให้กับจุดที่พิจารณา

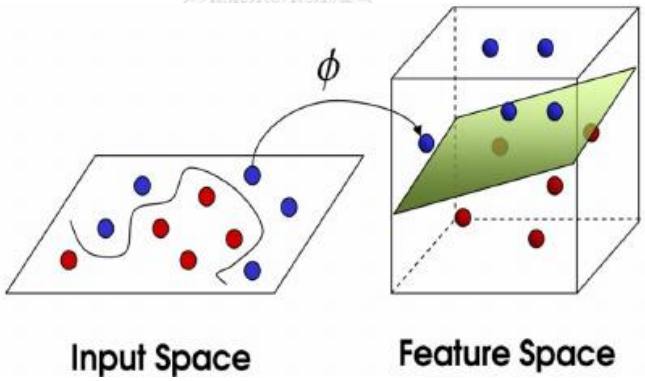
2.1.6 ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)

ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine) [22] เป็นโมเดลที่ใช้การคำนวณทางคณิตศาสตร์จัดอยู่ในกลุ่มเดียวกับโครงข่ายประสาทเทียม (Neural Network) เทคนิคนี้ถูกเสนอโดย Vapnik ในปี 1999 ซึ่งหลักการทำงาน คือ สร้างแบบจำลอง (Model) ที่ได้มาจากกลุ่มข้อมูลเรียนรู้ (Training Data) และนำแบบจำลองนี้ไปพยากรณ์ข้อมูลในอนาคต ซัพพอร์ตเวกเตอร์แมชชีนจะสร้างไฮเปอร์เพลน (Hyperplan) ที่เหมาะสม โดยแบ่งข้อมูลออกเป็นสองส่วน โดยพยายามสร้างเส้นแบ่งตรงกึ่งกลางระหว่างกลุ่มให้มีระยะห่างระหว่างขอบเขตของทั้งสองกลุ่มมากที่สุด หรือมีค่าขอบ (Margin) กว้างที่สุด ดังภาพที่ 3



ภาพที่ 3 การจำแนกประเภทด้วยวิธีซัพพอร์ตเวกเตอร์แมชชีน

ซัพพอร์ตเวกเตอร์แมชชีนจะใช้ฟังก์ชั่นเพื่อย้ายข้อมูลจากพื้นที่ข้อมูลนำเข้า (Input Space) ไปยังพื้นที่คุณลักษณะ (Feature Space) และสร้างฟังก์ชั่นวัดความคล้ายที่เรียกว่าเคอร์เนลฟังก์ชั่น (Kernel Function) เพื่อทำให้ข้อมูลเรียงตัวในมิติที่สูงขึ้น (Higher Dimensional Space) ดังภาพที่ 4



ภาพที่ 4 การจัดข้อมูลจากพื้นที่ข้อมูลนำเข้าไปยังพื้นที่คุณลักษณะที่เรียงตัวในมิติสูงขึ้น

กำหนดให้ $(x_1 y_1), \ldots, (x_n y_n)$ คือ ตัวอย่างที่ใช้สำหรับการสอน

โดย

n คือ จำนวนข้อมูลตัวอย่าง

m คือ จำนวนมิติข้อมูลเข้า

y คือ ผลลัพธ์มีค่า +1 หรือ -1

ดังสมการ 2.3

$$(x_1 y_1), \ldots, (x_n y_n) \text{ เมื่อ } x \in R^m y \in \{+1, -1\} \quad (2.3)$$

สำหรับปัญหาเชิงเส้น มิติข้อมูลขนาดสูงได้ถูกแบ่งเป็น 2 กลุ่ม โดยระนาบตัดสินใจซึ่งคำนวณได้ ดังสมการ 2.4

$$(w * x) + b = 0 \quad (2.4)$$

โดย

w คือ ค่าน้ำหนัก

b คือ ค่าเอนเอียง (bias)

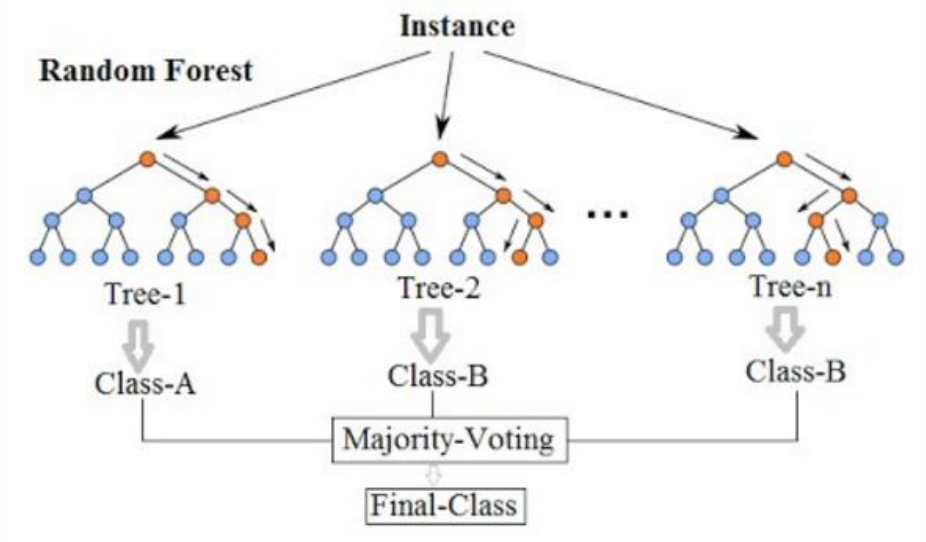
และจำแนกประเภทของข้อมูล ดังสมการ 2.5 และ 2.6

$$(w * x) + b > 0 \text{ ถ้า } y_i = +1 \quad (2.5)$$

$$(w * x) + b < 0 \text{ ถ้า } y_i = -1 \quad (2.6)$$

2.1.7 ป่าแบบสุ่ม (Random Forest)

ป่าแบบสุ่ม (Random Forest) เป็นการเรียนรู้ของเครื่อง ซึ่งเป็นการจำแนกประเภทแบบไม่ตัดแต่งกิ่ง (Unpruned) หรือต้นไม้ถดถอย (Regression Trees) ถูกเสนอโดย Tin Kam Ho ในปี 1995 และถูกขยายและจัดทำให้เป็นรูปแบบทั่วไปโดย Breiman [23] โดยเทคนิคป่าแบบสุ่มจะนำข้อมูลฝึกสอนไปสุ่มเลือกตัวอย่างข้อมูล และคุณลักษณะข้อมูล แล้วนำมาสร้างเป็นต้นไม้ตัดสินใจ (Decision Tree) หลาย ๆ ต้น และทำการสุ่มเลือกคุณลักษณะ (Feature) ต่าง ๆ เป็นหลาย ๆ ชุด ไม่เหมือนกัน ซึ่งมีตัวอย่างส่วนหนึ่งที่ไม่ถูกเลือก (Out-of-Bag: OOB) จะถูกนำมาใช้ในการทดสอบต้นไม้ตัดสินใจ เรียกว่า แบ็กกิ้ง (Bagging) แล้วนำไปสร้างโมเดล โดยการทำนายคำตอบจะเลือกจากผลโหวตจากต้นไม้ตัดสินใจที่มากที่สุด (Majority Vote) ดังภาพที่ 5



ภาพที่ 5 แนวคิดหาคำตอบของวิธีการป่าแบบสุ่ม

2.1.8 นาอีฟเบย์ส (Naïve Bayes)

นาอีฟเบย์ส (Naïve Bayes) เป็นการเรียนรู้ของเครื่อง [24] ซึ่งเป็นตัวจำแนกประเภทของแบบจำลองเชิงทำนาย (Supervised Modeling) โดยนาอีฟเบย์สมีพื้นฐานมาจากทฤษฎีของเบย์ส (Bayes Theorem) ซึ่งเป็นวิธีการเรียนรู้ที่ใช้หลักการของความน่าจะเป็น เพื่อหาว่าสมมติฐานใดน่าจะถูกต้องที่สุด โดยใช้ความรู้ก่อนหน้า ดังสมการ 2.7

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \tag{2.7}$$

โดย

$P(h)$ คือ ความน่าจะเป็นก่อนหน้าของสมมติฐาน h

$P(D)$ คือ ความน่าจะเป็นก่อนหน้าของชุดข้อมูลตัวอย่าง D

$P(h|D)$ คือ ความน่าจะเป็นของ h เมื่อรู้ D

$P(D|h)$ คือ ความน่าจะเป็นของ D เมื่อรู้ h

นาอีฟเบย์สเป็นวิธีการจำแนกข้อมูลที่ง่าย รวดเร็ว และมีประสิทธิภาพวิธีหนึ่ง เหมาะกับกรณีของเซตตัวอย่างที่มีจำนวนมาก และคุณสมบัติไม่ขึ้นต่อกัน สามารถประยุกต์ใช้งานการจำแนกประเภทข้อความ (Text Classification) และการวินิจฉัย (Diagnosis) [25]

สมมุติให้ $A_1, A_1, \ldots, A_1$ เป็นคุณสมบัติของตัวอย่าง จะได้ค่าที่น่าจะเป็นที่สุดของตัวอย่าง x ดังสมการ 2.8, 2,9 และ 2.10

$$v_{MAP} = \underset{v_j \in V}{argmax}\, P\left(v_j \middle| a_1, a_2, \ldots, a_n\right) \tag{2.8}$$

$$v_{MAP} = \underset{v_j \in V}{argmax}\, \frac{P(a_1, a_2, \ldots, a_n | v_j) P(v_j)}{P(a_1, a_2, \ldots, a_n)} \tag{2.9}$$

$$v_{MAP} = \underset{v_j \in V}{argmax}\, P\left(a_1, a_2, \ldots, a_n \middle| v_j\right) P(v_j) \tag{2.10}$$

โดย

a_i คือ ค่าคุณสมบัติของ A_i

V คือ เซตของประเภทหรือค่าที่เป็นไปได้ของ x

แต่พบว่าสมการนี้ไม่สามารถใช้งานได้อย่างมีประสิทธิภาพ เนื่องจากว่าการคำนวณค่าของ $P\left(a_1, a_2, \ldots, a_n \middle| v_j\right)$ ทำได้ยากลำบากมากเพื่อให้ได้ค่าที่น่าเชื่อถือในเชิงสถิติ โดยสมมติฐานของตัวจำแนกประเภทนาอีฟเบย์ส คือ กำหนดให้คุณสมบัติแต่ละตัวเป็นอิสระกับคุณสมบัติอื่น ๆ ซึ่งทำให้สามารถเขียนแทน $P\left(a_1, a_2, \ldots, a_n \middle| v_j\right)$ ด้วยผลคูณของค่าความน่าจะเป็น ดังสมการ 2.11

$$P\left(a_1, a_2, \ldots, a_n \middle| v_j\right) = \prod_{i=1}^{n} P(a_i | v_j) \tag{2.11}$$

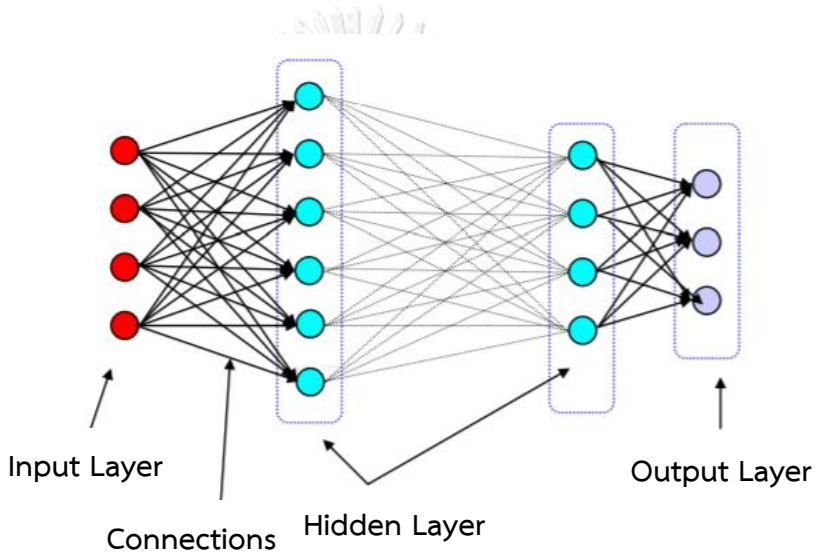
โดยที่ $\prod$ หมายถึงการนำค่า $P(a_i|v_j)$ ทั้งหมดมาคูณกัน ซึ่งสมการทางด้านซ้ายจะเท่ากับทางด้านขวาก็ต่อเมื่อคุณสมบัติ $a_1, a_2, \ldots, a_n$ ไม่ขึ้นต่อกัน ตามสมมติฐานความไม่ขึ้นต่อกัน (Conditional Independence Assumption) ดังนั้น จะได้ตัวแปรจำแนกประเภทเบย์สอย่างง่าย ดังสมการ 2.12

$$v_{NB} = \underset{v_j \in V}{argmax}\, P(v_j) \prod_{i=1}^{n} P(a_i | v_j) \tag{2.12}$$

2.1.9 โครงข่ายประสาทเทียม (Neural Network)

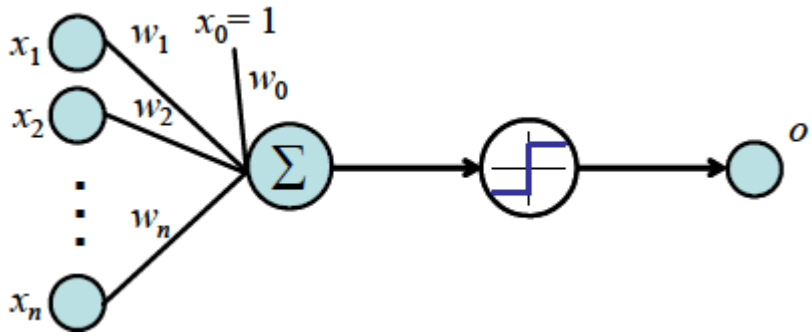
โครงข่ายประสาทเทียม (Neural Network) เป็นการเรียนรู้ของเครื่อง [26] ที่จำลองการทำงานของเซลล์ประสาท (Neuron) ของสมองมนุษย์ เพื่อให้มีความสามารถในการเรียนรู้ การจดจำแบบรูป (Pattern Recognition) และการอุปมานความรู้ (Knowledge Deduction)

พื้นฐานของโครงข่ายประสาทเทียมประกอบไปด้วย 3 ส่วน ได้แก่ ชั้นนำเข้า (Input Layer) ที่ถูกเชื่อมต่อกับชั้นซ่อน (Hidden Layer) ซึ่งเชื่อมต่อกับชั้นส่งออก (Output Layer) ดังภาพที่ 6



ภาพที่ 6 สถาปัตยกรรมของโครงข่ายประสาทเทียม

โดยชั้นนำเข้าจะทำหน้าที่แทนส่วนของข้อมูลดิบ ที่จะถูกป้อนเข้าสู่เครือข่าย ชั้นซ่อนจะถูกกำหนด โดยการทำงานของชั้นนำเข้า และค่าน้ำหนักบนความสัมพันธ์ระหว่างชั้นนำเข้าและชั้นซ่อน ซึ่งชั้นซ่อนสามารถมีได้มากกว่า 1 ชั้น หากชั้นซ่อนมีมากกว่า 1 ชั้น การคำนวณข้อมูลจะใช้ข้อมูลของชั้นซ่อนก่อนหน้าและส่งต่อไปเรื่อย ๆ ส่วนชั้นส่งออกจะขึ้นอยู่กับการทำงานของชั้นซ่อนและค่าน้ำหนักระหว่างชั้นซ่อนและชั้นส่งออก โดยโครงข่ายประสาทเทียมมีหน่วยพื้นฐานที่เล็กที่สุดเรียกว่า เพอร์เซ็ปตรอน (Perceptron) มีโครงสร้างดังภาพที่ 7



ภาพที่ 7 แสดงโครงสร้างของเพอร์เซ็ปตรอน

เพอร์เซ็ปตรอนทำหน้าที่แยกข้อมูลส่งออกแต่ละกลุ่มออกจากกัน และมีฟังก์ชั่นการทำงาน แสดงดังสมการ 2.13

$$f(x) = \sum_{i=1}^{n} w_i x_i + b \tag{2.13}$$

โดย

w คือ ค่าน้ำหนัก

b คือ ค่าเอนเอียง (Bias)

n คือ จำนวนของข้อมูลนำเข้า

โดยข้อมูลส่งออกจะถูกนำไปคำนวณค่าผิดพลาด (Error) เพื่อนำมาปรับน้ำหนักของข้อมูลนำเข้าต่อไป แสดงดังสมการ 2.14

$$E = \frac{1}{2}\sum_{d \in D}(t_d - o_d)^2 \tag{2.14}$$

โดย

D คือ เซตตัวอย่างสอน

t_d คือ ข้อมูลเป้าหมายของตัวอย่าง d

o_d คือ ข้อมูลส่งออกของเพอร์เซ็ปตรอนของตัวอย่าง d

เมื่อคำนวณค่าความผิดพลาดแล้ว แบบจำลองจะทำการปรับค่าน้ำหนัก แสดงดังสมการ 2.15

$$\Delta w = \eta \sum_{d \in D} (t_d - o_d) x_{id} \quad (2.15)$$

โดย

η คือ ค่าเรียนรู้ (Learning Rate)

x_{id} คือ สมาชิก x_i ของตัวอย่าง d

2.1.10 การวัดประสิทธิภาพการจำแนกของแบบจำลอง

การวัดประสิทธิภาพการจำแนกของโมเดลและแสดงผลด้วยเมตริกซ์ เป็นส่วนสำคัญในขั้นตอนสุดท้ายของการทำเหมืองข้อมูล เนื่องจากการวัดประสิทธิภาพของการจำแนกข้อมูล (Classifier) จะบอกถึงความน่าเชื่อถือของแบบจำลอง

เมทริกซ์ความสับสน (Confusion Matrix) [27] คือ ตารางที่มีจำนวนแถวเท่ากับจำนวนคอลัมน์ และเท่ากับจำนวนคลาส ตัวอย่างเช่น มีคลาสคำตอบอยู่ 2 ค่า คือ คลาสบวก และคลาสลบ ฉะนั้น จะสร้างได้เป็นตารางขนาด 2x2 โดยข้อมูลด้านคอลัมน์ คือ ผลการจำแนกจากแบบจำลอง และข้อมูลในแนวแถว คือ ค่าที่แท้จริง ได้ดังตารางที่ 2

ตารางที่ 2 ตารางเมทริกซ์ความสับสน

		ผลการจำแนก	
		คลาสบวก	คลาสลบ
ค่าที่แท้จริง	คลาสบวก	TP	FN
	คลาสลบ	FP	TN

โดย

TP (True Positive) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นคลาสบวก

TN (True Negative) คือ จำนวนข้อมูลที่จำแนกถูกว่าเป็นคลาสลบ

FP (False Positive) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นคลาสบวก

FN (False Negative) คือ จำนวนข้อมูลที่จำแนกผิดว่าเป็นคลาสลบ

ค่า TP, TN, FP และ FN สามารถนำมาใช้คำนวณมาตรวัดต่าง ๆ เพื่อวัดประสิทธิภาพการจำแนกข้อมูลของแบบจำลอง โดยมาตรวัดที่นิยมใช้โดยทั่วไปประกอบด้วยค่าความแม่นยำ ค่าความครบถ้วน ค่าความถูกต้อง โดยแต่ละมาตรวัดมีการคำนวณได้ดังนี้

การวัดค่าความแม่นยำ (Precision) เป็นการวัดความถูกต้อง โดยสนใจเฉพาะส่วนของการจำแนก คำนวณได้จากการหาอัตราส่วนของจำนวนข้อมูลที่จำแนกถูกว่าเป็นคลาสบวก เทียบกับจำนวนข้อมูลที่จำแนกว่าเป็นคลาสบวกทั้งหมด ดังสมการ 2.16

$$Precision = \frac{TP}{TP+FP} \tag{2.16}$$

การวัดค่าความครบถ้วน (Recall) เป็นการวัดความถูกต้อง โดยสนใจเฉพาะส่วนของค่าที่แท้จริง คำนวณได้จากการหาอัตราส่วนของจำนวนข้อมูลที่จำแนกถูกว่าเป็นคลาสบวก เทียบกับจำนวนข้อมูลที่แท้จริงของคลาสบวกทั้งหมด ดังสมการ 2.17

$$Recall = \frac{TP}{TP+FN} \tag{2.17}$$

การวัดค่าความถูกต้อง (Accuracy) เป็นการวัดความถูกต้อง โดยพิจารณารวมทุกคลาส จะมีค่าอยู่ระหว่าง 0-1 โดยถ้าค่ายิ่งเข้าใกล้ 1 แปลว่า แบบจำลองสามารถทำนายผลได้ดี ดังสมการ 2.18

$$Accuracy = \frac{TP+TN}{TP+FP+TN+FN} \tag{2.18}$$

2.2 งานวิจัยที่เกี่ยวข้อง

การจำแนกลักษณะส่วนบุคคลของผู้ใช้งานบนสื่อสังคมออนไลน์โดยการใช้แบบจำลองการเรียนรู้ของเครื่องร่วมกับการประมวลผลภาษาธรรมชาติ มีด้วยกันหลากหลายวิธีจากหลายงานวิจัย โดยมีการใช้ภาษาต่าง ๆ กัน ข้อมูลต่าง ๆ กัน ทั้งข้อความ ชื่อ ภาพ สี หรือตำแหน่ง เพื่อจำแนกเพศ อายุ บุคลิกภาพ การศึกษา สถานภาพการสมรส ประเทศ ภาษา ถิ่นกำเนิด เชื้อชาติ หรือชาติพันธ์ สรุปงานวิจัยที่เกี่ยวข้องกับการจำแนกลักษณะส่วนบุคคลจากสื่อสังคมออนไลน์ได้ดังตารางที่ 3

ตารางที่ 3 ตารางงานวิจัยที่เกี่ยวข้องกับการจำแนกลักษณะส่วนบุคคลจากสื่อสังคมออนไลน์

ผู้วิจัย	ที่มาข้อมูล	ข้อมูล					ภาษา	จำแนก
		ข้อความ	ชื่อ	ภาพ	อื่นๆ	ตำแหน่ง		
Schwartz, H.A. และคณะ (2013)	เฟซบุ๊ก	✓					อังกฤษ	บุคลิกภาพ, เพศ, อายุ
Alowibdi, J.S. และคณะ (2013)	ทวิตเตอร์		✓		✓		อังกฤษ	เพศ
Bergsma, S. และคณะ (2013)	ทวิตเตอร์		✓			✓	อังกฤษ	ประเทศ, ภาษา, เพศ, ถิ่นกำเนิด, เชื้อชาติ, ชาติพันธุ์
Akbar, R. (2016)			✓				อินโดนีเซีย	เพศ
Septiandri, A.A. (2017)			✓				อินโดนีเซีย	เพศ
Briediene, M. และคณะ (2018)	เฟซบุ๊ก	✓					ลิทัวเนีย	เพศ, อายุ, การศึกษา, บุคลิกภาพ, สถานภาพการสมรส
Hirt, R. และคณะ (2019)	ทวิตเตอร์	✓	✓	✓			เยอรมัน	เพศ
Vicente, M. และคณะ (2019)	ทวิตเตอร์	✓	✓	✓			อังกฤษ โปรตุเกส	เพศ

งานวิจัยของ Schwartz, H.A., Eichstaedt, J.C., Kern, M.L., Dziurzynski, L., Ramones, S.M., Agrawal, M และคณะ [5] ศึกษาเกี่ยวกับคำหรือวลีต่าง ๆ ที่ใช้บนสื่อสังคมออนไลน์เพื่อสำรวจการใช้รูปแบบภาษาของผู้ใช้ที่มีบุคลิกภาพ เพศ และอายุที่ต่างกัน โดยใช้เทคนิค คำศัพท์แบบเปิด (Open-Vocabulary) ของการวิเคราะห์ภาษาที่ต่างกัน (Differential Language Analysis:

DLA) โดยจำแนกกลุ่มของบุคลิกภาพตามโมเดลห้าปัจจัย (Five Factor Model) หรือ บิ๊ก5 (Big 5) ได้แก่ ความสนใจต่อสิ่งภายนอก (Extraversion) ความยินยอมเห็นใจ (Agreeableness) ความพิถีพิถัน (Conscientiousness) ความไม่เสถียรทางอารมณ์ (Neuroticism) และความเปิดรับ (Openness) และแบ่งช่วงอายุเป็น 4 ช่วง คือ ช่วงอายุ 13-18 ปี อายุ 19-22 ปี อายุ 23-29 ปี และอายุ 30-65 ปี โดยใช้คุณลักษณะการสืบสวนทางภาษาและการนับคำ (Linguistic Inquiry and Word Count: LIWC) หัวข้อ คำหรือวลี จากผลการศึกษาพบความถูกต้องอยู่ที่ 91.9% โดยคุณลักษณะทั้งหัวข้อ คำ และวลี ให้ความถูกต้องมากที่สุด โดยการศึกษารูปแบบภาษาของเพศที่ต่างกัน พบว่า เพศชายใช้สัญลักษณ์แทนอารมณ์มากกว่าเพศหญิง ส่วนการใช้คำพบว่าเพศชายมักใช้คำว่า “ฟุตบอล”, “ภรรยา”, “แฟนสาว” เพศหญิงมักใช้คำว่า “ชอปปิ้ง”, “สามี”, “แฟนหนุ่ม” โดยช่วงอายุ 13-18 ปี มักพบการใช้คำ “โรงเรียน”, “การบ้าน”, “การ์ตูน” ช่วงอายุ 19-22 ปี มักพบการใช้คำ “มหาวิทยาลัย”, “วิทยานิพนธ์” ช่วงอายุ 23-29 ปี มักพบการใช้คำ “ทำงาน”, “สังสรรค์”, “ชอปปิ้ง” ช่วงอายุ 30-65 ปี มักพบการใช้คำ “ครอบครัว”, “ลูก”, “การเมือง” และการศึกษารูปแบบภาษาของบุคลิกภาพที่ต่างกัน พบว่า บุคลิกภาพด้านความสนใจต่อสิ่งภายนอก มักพบการใช้คำ “สังสรรค์”, “รักคุณ”, “ผู้หญิง” บุคลิกด้านการเก็บตัว มักพบการใช้คำ “คอมพิวเตอร์”, “อินเตอร์เน็ต”, “อ่านหนังสือ” บุคลิกภาพด้านความวิตกกังวล มักพบการใช้คำ “เครียด”, “กดดัน”, “ร้องไห้” และบุคลิกด้านอารมณ์ที่มั่นคง มักพบการใช้คำ “ประสบความสำเร็จ”, “วันที่สวยงาม”, “มีโชค”

งานวิจัยของ Alowibdi, J.S., U.A. Buy และ P. Yu [6] ศึกษาเกี่ยวกับการจำแนกเพศจากลักษณะของข้อมูลส่วนตัวจากทวิตเตอร์ โดยทดลองจาก 3 ลักษณะ ได้แก่ ชื่อจริง, ชื่อผู้ใช้งาน และสีของข้อมูลส่วนตัว (สีของพื้นหลัง, สีของตัวอักษร, สีของลิงค์, สีของแถบด้านข้าง และสีของขอบแถบด้านข้าง) โดยใช้ข้อมูลส่วนตัวของผู้ใช้งานทวิตเตอร์จำนวน 194,293 คน โดยเปรียบเทียบแบบจำลองระหว่าง นาอีฟเบย์ส (Naive Bayes), ต้นไม้ตัดสินใจ (Decision Tree) และผสมระหว่างเบส์อย่างง่ายกับต้นไม้ตัดสินใจ และประเมินค่าความแม่นยำด้วยการเลือกสุ่มข้อมูลแบบความเที่ยงตรงโดยการแบ่งชุดข้อมูลเป็น 10 ส่วน (10-Fold Cross Validation) โดยได้แบ่งการทดลองออกเป็น 3 การทดลอง คือ 1) การทดลองใช้ข้อมูลสีของข้อมูลส่วนตัว ได้ผลลัพธ์ความถูกต้องที่มากที่สุดที่ 74% จากการใช้ข้อมูลทั้ง 5 สี พร้อมด้วยการใช้การปรับสีและการจัดเรียงสี กับขั้นตอนวิธีตัวจำแนก

ประเภทผสมระหว่างนาอีฟเบย์สกับต้นไม้ตัดสินใจ ซึ่งสีที่มีการใช้มากที่สุด 5 อันดับแรกของเพศหญิง ได้แก่ สีชมพู สีเหลือง สีเขียว สีแดง และสีฟ้า ส่วนของเพศชาย ได้แก่ สีดำ สีน้ำตาล สีส้ม สีเทา และสีน้ำเงิน 2) การทดลองใช้ข้อมูลชื่อจริง ได้ผลลัพธ์ความถูกต้องมากที่สุดที่ 82.5% ด้วยการตัดคำแบบ 3-แกรม ร่วมกับคุณลักษณะของหน่วยเสียง กับตัวจำแนกประเภทต้นไม้ตัดสินใจ 3) การทดลองใช้ข้อมูลชื่อผู้ใช้งาน ได้ผลลัพธ์ความถูกต้องมากที่สุดที่ 75.2% ด้วยการตัดคำแบบ 3-แกรมร่วมกับคุณลักษณะของหน่วยเสียง กับตัวจำแนกประเภทต้นไม้ตัดสินใจ จากผลการทดลองสรุปได้ว่า การจำแนกเพศจากชื่อจริงให้ความถูกต้องมากที่สุด

งานวิจัยของ Bergsma, S., Dredze, M., Van D.B., Wilson, T., และ Yarowsky, D. [7] ศึกษาเกี่ยวกับการจำแนกประเทศ ภาษา เพศ ถิ่นกำเนิด เชื้อชาติ (Ethnicity) และชาติพันธุ์ (Race) โดยใช้ข้อมูลชื่อจริง นามสกุล และตำแหน่งจากทวิตเตอร์ โดยแบ่งการทดลองออกเป็นการใช้วิธีการตัดคำอย่างเดียว, เอ็น-แกรมอย่างเดียว, จัดกลุ่มอย่างเดียว, ตัดคำร่วมกับเอ็น-แกรม และร่วมกันทั้งการตัดคำ, เอ็น-แกรม และจัดกลุ่ม โดยใช้แบบจำลองซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM) ซึ่งจากการจำแนกทั้ง 6 ประเภท มีการใช้ข้อมูลแตกต่างกัน คือ 1) จำแนกประเทศ ใช้ข้อมูลชื่อจริง, นามสกุล และตำแหน่ง 2) จำแนกภาษา ใช้ข้อมูลชื่อจริง นามสกุล และตำแหน่ง 3) จำแนกเพศ ใช้ข้อมูลชื่อจริง และนามสกุล 4) จำแนกถิ่นกำเนิด ใช้ข้อมูลชื่อจริง และนามสกุล 5) จำแนกเชื้อชาติ ใช้ข้อมูลชื่อจริง และนามสกุล 6) จำแนกชาติพันธุ์ ใช้ข้อมูลชื่อจริง และนามสกุล ซึ่งจากผลการทดลองพบว่าการใช้ร่วมกันทั้งการตัดคำ, เอ็น-แกรม และจัดกลุ่มให้ความถูกต้องมากที่สุดคิดเป็น 86.7%, 82.7%, 90.2%, 88.0%, 81.3% และ 84.6% สำหรับประเทศ, ภาษา, เพศ, ถิ่นกำเนิด, เชื้อชาติ และชาติพันธุ์ ตามลำดับ

งานวิจัยของ Akbar, R. [8] ศึกษาเกี่ยวกับการจำแนกเพศจากชื่อชาวอินโดนีเซีย โดยเปรียบเทียบแบบจำลองระหว่างนาอีฟเบย์สแบบอเนกนาม (Multinomial Naive Bayes) และป่าแบบสุ่ม (Random Forrest) โดยใช้ชื่อชาวอินโดนีเซีย 50,000 ชื่อ ซึ่งมีการทดลอง 2 รูปแบบ คือ ใช้ข้อมูลเฉพาะชื่อจริงในคำแรก และใช้ข้อมูลชื่อจากทุกคำ โดยใช้คุณลักษณะจำนวนของคำ ความถี่ของจำนวนตัวอักษร และตัวอักษรตัวสุดท้าย ซึ่งผลการศึกษาพบว่า แบบจำลองป่าแบบสุ่มให้ความถูกต้องที่มากกว่า ที่ 83% และนาอีฟเบย์สแบบอเนกนามให้ความถูกต้องที่ 70% และพบว่าการใช้ข้อมูลชื่อจากทุกคำให้ความถูกต้องที่มากกว่าการใช้ข้อมูลชื่อเพียงอย่างเดียว

งานวิจัยของ Septiandri, A.A. [9] ศึกษาเกี่ยวกับการจำแนกเพศจากชื่อชาวอินโดนีเซีย โดยใช้เทคนิคหน่วยความจำระยะสั้นแบบยาวระดับอักขระ (Character-Level Long-Short Term Memory: Char-LSTM) เปรียบเทียบแบบจำลองระหว่าง นาอีฟเบย์ส (Naive Bayes), ถดถอยโลจิสติกส์ (Logistic Regression) และเอ็กจีบูส (Extreme Gradient Boost: XGBoost) ร่วมกับเอ็น-แกรม (N-gram) โดยใช้ข้อมูลชื่อชาวอินโดนิเซีย 6,881 ชื่อ เป็นชื่อเพศชาย 4,580 ชื่อ และชื่อเพศหญิง 2,301 ชื่อ โดยแบ่งออกเป็น 2 การทดลอง คือ ใช้ทั้งข้อมูลชื่อและนามสกุล และใช้ข้อมูลชื่อเพียงอย่างเดียว ซึ่งผลการศึกษาพบว่าประสิทธิภาพที่ดีที่สุดของแบบจำลองนาอีฟเบย์ส และถดถอยโลจิสติกส์ ด้วยการใช้ 3-แกรม ส่วนประสิทธิภาพที่ดีที่สุดของแบบจำลองเอ็กจีบูส ด้วยการใช้ 2-แกรม และเมื่อใช้เทคนิคหน่วยความจำระยะสั้นแบบยาวระดับอักขระสามารถทำนายเพศได้ถูกต้องเพิ่มขึ้นกว่าการใช้แบบจำลองถดถอยโลจิสติกส์ จาก 85.28% เป็น 92.25% ในกรณีที่ใช้ทั้งชื่อและนามสกุล ส่วนถ้าใช้เพียงชื่อเท่านั้น ให้ความถูกต้องที่ 90.65%

งานวิจัยของ Briediene, M. และ Kapociute-Dzikiene, J. [10] ศึกษาเกี่ยวกับการหาประวัติของผู้เขียนจากข้อความแบบอัตโนมัติจากเฟซบุ๊ก ซึ่งเป็นข้อความสั้น ๆ ภาษาลิทัวเนียแบบไม่เป็นทางการ โดยภาษาลิทัวเนียมีการแปลงคำที่สูง งานวิจัยต้องการหาเพศ, อายุ, การศึกษา, สถานภาพการสมรส และบุคลิกภาพ โดยได้แบ่งการทดลองออกเป็น 2 ส่วน คือ การทดลองแบบจำลองระหว่างซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine: SVM), นาอีฟเบย์ส (Naive Bayes: NB), นาอีฟเบย์สแบบอเนกนาม (Naive Bayes Multinomial: NBM), ไอบีเค (IBK), เคสตาร์ (Kstar) และการทดลองลักษณะสำคัญระหว่างการแบ่งแบบคำและแบ่งแบบตัวอักษร ซึ่งผลการศึกษา พบว่าตัวจำแนกนาอีฟเบย์สแบบอเนกนามร่วมกับการแบ่งแบบตัวอักษรให้ความถูกต้องมากที่สุด คิดเป็น 84.3%, 52.7%, 79.6%, 76.6% และ 79.1% สำหรับเพศ, อายุ, การศึกษา, สถานภาพการสมรส และบุคลิกภาพ ตามลำดับ

งานวิจัยของ Hirt, R., N. Kühl และ G. Satzger [11] ศึกษาเกี่ยวกับการคาดการณ์เพศจากทวิตเตอร์โดยอัตโนมัติ โดยใช้ข้อมูลข้อความทวิต, ชื่อ และรูปประจำตัว จาก 2,916 ผู้ใช้งานทวิตเตอร์ที่ใช้ภาษาเยอรมัน โดยใช้ตัวจำแนกนาอีฟเบย์สในการเปรียบเทียบการทดลอง ได้แบ่งการทดลองออกเป็น 5 การทดลอง คือ 1) ข้อมูลข้อความทวิตเท่านั้น ได้ค่าคะแนนเอฟ1 (F1-score) 69.79% 2) ข้อมูลชื่อเท่านั้น ได้ค่าคะแนนเอฟ1 69.06% 3) ข้อมูลรูปประจำตัวเท่านั้น ได้ค่าคะแนน

เอฟ1 25.37% 4) ข้อมูลข้อความทวิตและชื่อ ได้ค่าคะแนนเอฟ1 79.63% 5) ข้อมูลข้อความทวิต, ชื่อ และรูปประจำตัว ได้ค่าคะแนนเอฟ1 81.46% ซึ่งผลการศึกษาพบว่า ผลลัพธ์เปรียบเทียบด้วยคะแนนเอฟ1 การใช้ข้อมูลข้อความทวิต, ชื่อ และรูปประจำตัว ให้ค่ามากที่สุด

งานวิจัยของ Vicente, M., F. Batista และ J.P. Carvalho [12] ศึกษาเกี่ยวกับการจำแนกเพศจากชื่อผู้ใช้งาน, คำอธิบายตัวตน, ข้อความทวิต, รูปประจำตัว และกิจกรรมส่วนตัว จากผู้ใช้งานทวิตเตอร์ โดยใช้ข้อมูลทวิตเตอร์ภาษาอังกฤษ 6.5 ล้านทวิต 65,000 บัญชี และภาษาโปรตุเกส 5.8 ล้านทวิต 58,000 บัญชี ซึ่งได้เสนอวิธีการรวมแบบจำลองจากหลายแบบจำลอง โดยสร้างแบบจำลองที่ใช้ข้อมูลที่แตกต่างกัน 4 ข้อมูล ได้แก่ ชื่อผู้ใช้งาน, คำอธิบายตัวตน, ข้อความทวิต, และรูปประจำตัว โดยไม่ได้ใช้จำนวนผู้ติดตามและจำนวนที่ติดตาม เนื่องจากไม่บ่งบอกถึงเพศ โดยเปรียบเทียบประสิทธิภาพระหว่างตัวจำแนกถดถอยโลจิสติกส์ (Logistic Regression), นาอีฟเบย์สแบบอเนกนาม (Multinomial Naïve Bayes), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machines) และต้นไม้ตัดสินใจ (Decision Tree) ซึ่งแบบจำลองสุดท้าย ทำการรวมทั้ง 4 แบบจำลองที่ใช้ข้อมูลต่างกันไว้ด้วยกัน ซึ่งประสิทธิภาพที่ดีที่สุดได้ 93.2% ในภาษาอังกฤษ และ 96.9% ในภาษาโปรตุเกส

บทที่ 3
แนวคิดและวิธีการดำเนินงาน

ในบทนี้จะกล่าวถึงแนวคิดและวิธีการดำเนินงาน โดยนำทฤษฎีและงานวิจัยที่เกี่ยวข้องในบทที่ 2 มาประยุกต์ใช้กับการจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊ก ได้แก่ สภาพแวดล้อมและเครื่องมือ การรวบรวมชุดข้อมูล การสร้างคุณลักษณะ การออกแบบการทดลอง และการแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพการจำแนกของแบบจำลอง

3.1 สภาพแวดล้อมและเครื่องมือ

สภาพแวดล้อม

- ระบบปฏิบัติการวินโดว์ 10 แบบ 64 บิต
- หน่วยประมวลผล Intel Core i5 2.5 กิกะเฮิรตซ์
- หน่วยความจำ 8 กิกะไบต์

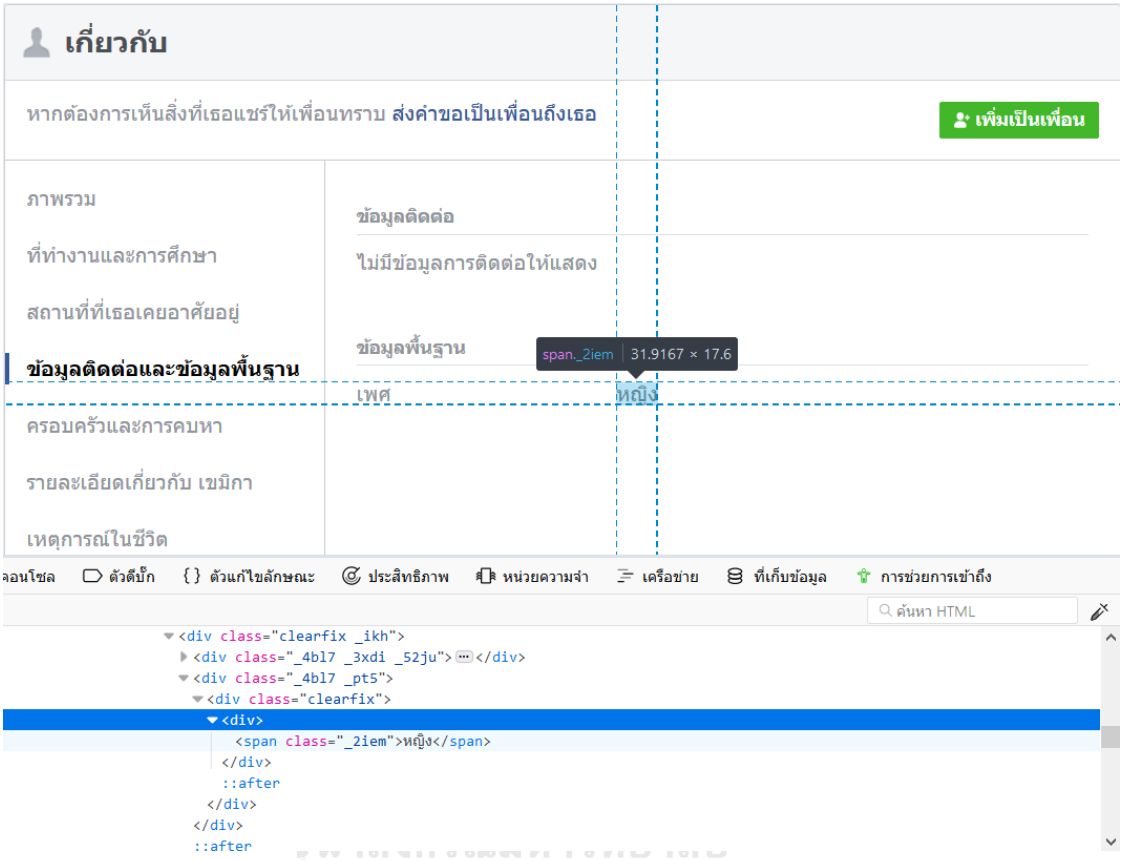
เครื่องมือ

- เครื่องมือกูเกิล โคแล็บ (Google Colab)
- เครื่องมืออนาคอนด้า 3 (Anaconda3)
- โปรแกรมไซคิตเลิร์น (Scikit-learn) 0.20.0
- ไลบรารีซิลิเนียม (Selenium)
- ไลบรารีไพไทยเอ็นแอลพี (PythaiNLP)
- ภาษาไพทอน (Python)

3.2 การรวบรวมชุดข้อมูล

การเลือกชุดข้อมูลสำหรับการทำวิจัยนี้ นำข้อมูลมาจากเฟซบุ๊กโดยเลือกเฉพาะชื่อผู้ใช้งานที่เป็นภาษาไทย โดยเลือกผู้ใช้งานแบบสุ่ม แล้วทำการไปดึงข้อมูลเพศที่แต่ละผู้ใช้งานเปิดเผยไว้ แล้วทำการระบุประเภทของชื่อว่าเป็นชื่อจริง หรือชื่อแฝง

การดึงข้อมูลจากเฟซบุ๊กสามารถทำได้โดยการใช้กราฟเอพีไอ (Graph API) แต่ปัจจุบันไม่เปิดให้บริการแล้ว ดังนั้น จึงทำการเก็บข้อมูลโดยการเขียนโปรแกรมภาษาไพธอน (Python) โดยใช้เครื่องมือซิลิเนียม (Selenium) ช่วยในการดึงข้อมูลที่อยู่ในองค์ประกอบ (Element) ที่ต้องการบนหน้าเว็บ ดังตัวอย่างภาพที่ 8



ภาพที่ 8 ตัวอย่างการดึงข้อมูลเพศของผู้ใช้งานเฟซบุ๊ก

หลังจากได้ข้อมูลชื่อผู้ใช้งานและเพศของผู้ใช้งานเฟซบุ๊กแล้ว จะนำชื่อผู้ใช้งานมาแยกประเภทชื่อแบ่งออกเป็นชื่อจริง และชื่อแฝง โดยในงานวิจัยนี้ทำการเก็บรวบรวมเฉพาะข้อมูลที่สนใจ คือ ชื่อผู้ใช้งาน เพศ และประเภทชื่อ ลงในไฟล์รูปแบบ .csv ตัวอย่างข้อมูลชื่อ เพศ และประเภทชื่อของผู้ใช้งานเฟซบุ๊ก ตามตารางที่ 4

ตารางที่ 4 ตารางตัวอย่างข้อมูลชื่อ เพศ และประเภทชื่อ ของผู้ใช้งานเฟซบุ๊ก

ชื่อผู้ใช้งาน	เพศ	ประเภทชื่อ
เทพบุตร ซาตาน	เพศชาย	ชื่อแฝง
หทัยรัตน์ โพธิ์ทอง	เพศหญิง	ชื่อจริง
เจ้าหญิงสายลม	เพศหญิง	ชื่อแฝง
ผู้ชายคนหนึ่ง ที่แสนธรรมดา	เพศชาย	ชื่อแฝง
สุรพงษ์ หลีเจริญ	เพศชาย	ชื่อจริง
แม่หญิงแสนดี ที่มีหัวใจ	เพศหญิง	ชื่อแฝง
เอกพล จิตเจริญสมุทร	เพศชาย	ชื่อจริง
หมูหวาน ชิลชิล	เพศชาย	ชื่อแฝง
กอ กุ๊ก' กุ๊กไก่	เพศหญิง	ชื่อแฝง
เบญจพร สำเริงกิจ	เพศหญิง	ชื่อจริง
พี่ เสือ	เพศชาย	ชื่อแฝง
ธีรยุทธ พลวงศ์ษา	เพศชาย	ชื่อจริง
นางมารร้าย ชอบเอาแต่ใจ	เพศหญิง	ชื่อแฝง
กานต์สินี เชิดชู	เพศหญิง	ชื่อจริง
ไม่หล่อ แต่ ดูไม่เบื่อ	เพศชาย	ชื่อแฝง

3.3 การสร้างคุณลักษณะ

คุณลักษณะสำหรับการจำแนกเพศชื่อผู้ใช้งานภาษาไทย ได้แก่ การตัดคำภาษาไทย การจำแนกชนิดของคำภาษาไทย การนับความถี่ตัวอักษร และการตัดตัวอักษรภาษาไทย

3.3.1 การตัดคำภาษาไทย

การตั้งชื่อของคนไทยระหว่างเพศชายและเพศหญิง มีรูปแบบการใช้คำที่นำมาตั้งชื่อแตกต่างกัน งานวิจัยนี้จึงได้ทำการสร้างคุณลักษณะการตัดคำภาษาไทย ซึ่งการตัดคำภาษาไทยมีไลบรารีหลายตัวที่สามารถตัดคำภาษาไทยได้ เช่น PythaiNLP, LibThai, PyICU เป็นต้น งานวิจัยนี้

เลือกใช้ไลบรารี PyThaiNLP โดยใช้วิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) ตัวอย่างข้อมูลชื่อและการตัดคำภาษาไทย ดังภาพที่ 10

ชื่อผู้ใช้งาน	การตัดคำ
แม่เพียร เขียนหาเห็ด	['แม่', 'เพียร', ' ', 'เขียน', 'หา', 'เห็ด']
แม่น้องพระเจ้า กับน้องพระคุณ	['แม่', 'น้อง', 'พระเจ้า', ' ', 'กับ', 'น้อง', 'พระคุณ']
เท้าเหยี่ยบดิน ตามองฟ้ามือคว้าดาว	['เท้า', 'เหยี่ยบ', 'ดิน', ' ', 'ตา', 'มอง', 'ฟ้า', 'มือ', 'คว้า', 'ดาว']
เราก็แค่ผู้ชาย หัวใจสองล้อ	['เรา', 'ก็', 'แค่', 'ผู้ชาย', ' ', 'หัวใจ', 'สอง', 'ล้อ']
แดน วิศรุต คนหล่อคับ	['แดน', ' ', 'วิศรุต', ' ', 'คน', 'หล่อ', 'คับ']
กิ๊ฟ ไม่กินผัก แม่บอกให้กินผัก	['กิ๊ฟ', ' ', 'ไม่', 'กิน', 'ผัก', ' ', 'แม่', 'บอก', 'ให้', 'กิน', 'ผัก']
แมวโหด กระโดดข่วน	['แมว', 'โหด', ' ', 'กระโดด', 'ข่วน']
โชกุน เด็กหลังเขา	['โชกุน', ' ', 'เด็ก', 'หลัง', 'เขา']
เวลาผมโกรธ จะโหดและน่ากลัวมาก	['เวลา', 'ผม', 'โกรธ', ' ', 'จะ', 'โหด', 'และ', 'น่ากลัว', 'มาก']
เอ็มแอนด์เอ็ม ของสีแดง	['เอ็ม', 'แอนด์', 'เอ็ม', ' ', 'ของ', 'สีแดง']
โจน้ำตื่ม คุ้มประชา	['โจ', 'น้ำตื่ม', ' ', 'คุ้ม', 'ประชา']
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']
ไก่จ๋า ได้ยินไหมว่าเสียงใคร	['ไก่', 'จ๋า', ' ', 'ได้ยิน', 'ไหม', 'ว่า', 'เสียง', 'ใคร']

ภาพที่ 9 ตัวอย่างข้อมูลชื่อและการตัดคำภาษาไทย

ภาพที่ 11 แสดงคำที่มีการใช้ในชื่อแฝงเพศชายและเพศหญิงจากการตัดคำ โดยคำที่พบในชื่อเพศชายแสดงในภาพด้านบน ส่วนของเพศหญิงแสดงในภาพด้านล่าง โดยขนาดของคำบ่งบอกถึงความถี่ กล่าวคือคำที่ใหญ่กว่าหมายถึงคำที่มีการใช้มากกว่า คำที่เล็กกว่าหมายยถึงคำที่มีการใช้งานน้อยกว่า โดยผลของการวิเคราะห์การตัดคำแสดงให้เห็นได้ว่าแต่ละเพศจะมีการใช้คำที่สื่อถึงตนเอง ตัวอย่างเช่น เพศชายมักใช้คำว่า “พี่”, “หนุ่ม”, “เสือ” ส่วนเพศหญิงมักใช้คำว่า “น้อง”, “แม่”, “หญิง”

ภาพที่ 10 เวิร์ดคลาวด์ของการตัดคำในชื่อแฝง

3.3.2 การจำแนกชนิดของคำภาษาไทย

คำในภาษาไทยมีหน้าที่แตกต่างกันออกไป โดยปัจจุบันมีงานวิจัยทางภาษาศาสตร์ของ Virach S. และคณะ [28] ทำการจำแนกชนิดของคำในภาษาไทยออกเป็นชนิดต่าง ๆ โดยจำแนกออกเป็น 14 กลุ่ม ได้แก่ คำนาม, คำสรรพนาม, คำกริยา, คำกริยาช่วย, คำลักษณนาม, คำวิเศษณ์,

คำบอกกำหนด, คำเชื่อม, คำบุพบท, คำอุทาน, คำขึ้นต้น, คำลงท้าย, คำปฏิเสธ และเครื่องหมายวรรคตอน ซึ่งแบ่งออกเป็น 47 ชนิด ตามตารางที่ 5

ตารางที่ 5 ตารางชนิดของคำในภาษาไทย

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
1	NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โค้ก, พระอาทิตย์
2	NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 3
3	NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2, ที่3
4	NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b
5	NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
6	NTTL	Title noun	ดร., พลเอก
7	PPRS	Personal pronoun	คุณ, เขา, ฉัน
8	PDMN	Demonstrative pronoun	นี่, นั่น, ที่นั่น, ที่นี่
9	PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
10	PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
11	VACT	Active verb	ทำงาน, ร้องเพลง, กิน
12	VSTA	Stative verb	เห็น, รู้, คือ
13	VATT	Attributive verb	อ้วน, ดี, สวย
14	XVBM	Pre-verb auxiliary, before negator “ไม่”	เกิด, เกือบ, กำลัง
15	XVAM	Pre-verb auxiliary, after negator “ไม่”	ค่อย, น่า, ได้
16	XVMM	Pre-verb, before or after negator “ไม่”	ควร, เคย, ต้อง
17	XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม
18	XVAE	Post-verb auxiliary	ไป, มา, ขึ้น

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
19	DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
20	DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น
21	DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
22	DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
23	DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่าง ๆ
24	DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ
25	DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
26	DCNM	Determiner, cardinal number expression	หนึ่งคน, เสือ 2 ตัว

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
27	DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
28	ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
29	ADVI	Adverb with iterative form	เร็ว ๆ, เสมอ ๆ, ช้า ๆ
30	ADVP	Adverb with prefixed	โดยเร็ว
31	ADVS	Sentential adverb	โดยปกติ, ธรรมดา
32	CNIT	Unit classifier	ตัว, คน, เล่ม
33	CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง, ด้าน, แบบ, รุ่น
34	CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง
35	CFQC	Frequency classifier	ครั้ง, เที่ยว
36	CVBL	Verbal classifier	ม้วน, มัด
37	JCRG	Coordinating conjunction	และ, หรือ, แต่
38	JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
39	JSBR	Subordinating conjunction	เพราะว่า, เนื่องจาก, ที่, แม้ว่า, ถ้า
40	RPRE	Preposition	จาก, ละ, ของ, ใต้, บน
41	INT	Interjection	โอ้ย, โอ้, เออ, เอ๋, อ๋อ
42	FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
43	FIXV	Adverbial prefix	อย่างเร็ว
44	EAFF	Ending for affirmative sentence	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
45	EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
46	NEG	Negator	ไม่, มิได้, ไม่ได้, มิ

ลำดับ	ชนิดของคำ	คำอธิบาย	ตัวอย่าง
47	PUNC	Punctuation	(,), “, „, ;

ชื่อจริงและชื่อแฝงภาษาไทยมักมีการประกอบไปด้วยชนิดของคำที่แตกต่างกัน โดยชื่อแฝงจะมีการใช้ชนิดของคำที่หลากหลาย เพื่อประกอบกันเป็นประโยค เช่น คำนาม คำสรรพนาม คำกริยา หรือคำวิเศษณ์ แต่ในชื่อจริงมักจะประกอบคำนามเพียงอย่างเดียว งานวิจัยนี้จึงได้ทำการสร้างคุณลักษณะการจำแนกชนิดของคำภาษาไทย โดยงานวิจัยนี้เลือกใช้ไลบรารี PyThaiNLP โดยใช้การตัดคำก่อน แล้วทำการจำแนกชนิดของคำ ตัวอย่างข้อมูลชื่อและการจำแนกชนิดของคำภาษาไทย ดังภาพที่ 12

ชื่อผู้ใช้งาน	ชนิดของคำ
เราก็แค่ผู้ชาย หัวใจสองล้อ	[('เรา', 'PPRS'), ('ก็', 'JSBR'), ('แค่', 'XVBM'), ('ผู้ชาย', 'NCMN'), ('หัวใจ', 'NCMN'), ('สอง', 'DCNM'), ('ล้อ', 'CNIT')]
สุนทร ดวงฤทัย	[('สุนทร', 'NPRP'), ('ดวง', 'NCMN'), ('ฤทัย', 'NCMN')]
ฉันหรือเธอ ที่เปลี่ยนไป	[('ฉัน', 'PPRS'), ('หรือ', 'JCRG'), ('เธอ', 'PPRS'), ('ที่', 'PREL'), ('เปลี่ยนไป', 'VSTA')]
เนตรชนก วงษ์สีโส	[('เนตร', 'NCMN'), ('ชนก', 'NCMN'), ('วงษ์', 'NCMN'), ('สี', 'NCMN'), ('โส', 'VATT')]
เหนื่อยก็พัก เขาไม่รักก็พอ	[('เหนื่อย', 'NCMN'), ('ก็', 'JSBR'), ('พัก', 'VACT'), ('เขา', 'PPRS'), ('ไม่', 'NEG'), ('รัก', 'VSTA'), ('ก็', 'JSBR'), ('พอ', 'ADVN')]
หัวใจ ไม่ใช่ก้อนดิน	[('หัวใจ', 'NCMN'), ('ไม่ใช่', 'NEG'), ('ก้อนดิน', 'NCMN')]
นายสมศักดิ์ ทองสุวรรณ	[('นาย', 'NTTL'), ('สม', 'NPRP'), ('ศักดิ์', 'NPRP'), ('ทอง', 'DDBQ'), ('สุวรรณ', 'NPRP')]
เวลาผมโกรธ จะโหดและน่ากลัวมาก	[('เวลา', 'NCMN'), ('ผม', 'PPRS'), ('โกรธ', 'ADVP'), ('จะ', 'XVBM'), ('โหด', 'VACT'), ('และ', 'JCRG'), ('น่ากลัว', 'VSTA'), ('มาก', 'ADVN')]
เค สอง.หนึ่ง.สี่	[('เค', 'NCMN'), ('สอง', 'DCNM'), ('.', 'PUNC'), ('หนึ่ง', 'DCNM'), ('.', 'PUNC'), ('สี่', 'NPRP')]
เมธ ไงจะใครล่ะ	[('เมธ', 'NCMN'), ('ไง', 'NCMN'), ('จะ', 'XVBM'), ('ใคร', 'PNTR'), ('ล่ะ', 'VACT')]
อุ่มอิ่ม วิจัยถึงไหนแล้ว	[('อุ่มอิ่ม', 'NCMN'), ('วิจัย', 'VACT'), ('ถึง', 'RPRE'), ('ไหน', 'PNTR'), ('แล้ว', 'XVAE')]
พิรุณ อุดมเดช	[('พิรุณ', 'NCMN'), ('อุดม', 'NPRP'), ('เดช', 'NCMN')]
เจ็บเพราะเขา เหงาเพราะเธอ	[('เจ็บ', 'NCMN'), ('เพราะ', 'JSBR'), ('เขา', 'PPRS'), ('เหงา', 'VACT'), ('เพราะ', 'JSBR'), ('เธอ', 'PPRS')]
เฉลิมชัย เรืององอาจ	[('เฉลิม', 'NCMN'), ('ชัย', 'NCMN'), ('เรือง', 'VSTA'), ('องอาจ', 'ADVP')]
ผู้ชายคนหนึ่ง ที่แสนธรรมดา	[('ผู้ชาย', 'NCMN'), ('คน', 'CNIT'), ('หนึ่ง', 'DCNM'), ('ที่', 'PREL'), ('แสน', 'VACT'), ('ธรรมดา', 'VATT')]
เอกพล จิตเจริญสมุทร	[('เอก', 'NCMN'), ('พล', 'NCMN'), ('จิต', 'NCMN'), ('เจริญ', 'NCMN'), ('สมุทร', 'NCMN')]
อิทธิณัฏฐ์ ธรรมมหามงคล	[('อิทธิ', 'NCMN'), ('ณัฏฐ์', 'NCMN'), ('ธรรม', 'NCMN'), ('มหา', 'NCMN'), ('มงคล', 'NPRP')]

ภาพที่ 11 ตัวอย่างข้อมูลชื่อและการจำแนกชนิดของคำภาษาไทย

การจำแนกชนิดของคำในชื่อจริงและชื่อแฝง พบว่า ในชื่อจริงมักจะพบคำนามเป็นส่วนใหญ่ และไม่ค่อยพบชนิดของคำบางประเภท ตัวอย่างเช่น คำเชื่อม เช่นคำว่า “และ”, “หรือ”, “แต่” คำลงท้าย เช่นคำว่า “ค่ะ”, “ครับ”, “ไหม" คำสรรพนาม เช่นคำว่า “ฉัน”, “นี้”, “ซึ่ง” ส่วนชนิดของคำในชื่อแฝงมักจะพบคำนามเป็นส่วนใหญ่เช่นเดียวกัน แต่มักจะพบชนิดของคำอื่น ๆ มากกว่าชื่อจริง ตัวอย่างเช่น คำกริยา เช่นคำว่า “รัก”, “สวย”, “เหงา” คำวิเศษณ์ เช่นคำว่า “เร็ว”, “สุด”,

“อีก” เครื่องหมายวรรคตอน เช่นคำว่า “(”, “’”,";” และไม่พบคำอุทาน เช่นคำว่า “เออ”, “อ๋อ”, “โอ้ย” ทั้งในชื่อจริงและชื่อแฝง ค่าสถิติการพบชนิดของของคำในชื่อจริงและชื่อแฝง ตามตารางที่ 6

ตารางที่ 6 ตารางค่าสถิติการพบชนิดของคำในชื่อจริงและชื่อแฝง

ชนิดของคำ	ชื่อจริง	ชื่อแฝง
คำนาม	95.40%	75.16%
คำสรรพนาม	0.04%	1.54%
คำกริยา	2.43%	7.87%
คำกริยาช่วย	0.48%	1.64%
คำลักษณนาม	0.06%	0.47%
คำวิเศษณ์	0.19%	1.36%
คำบอกกำหนด	0.88%	1.18%
คำเชื่อม	0.02%	1.01%
คำบุพบท	0.30%	1.12%
คำอุทาน	0.00%	0.00%
คำขึ้นต้น	0.03%	0.10%
คำลงท้าย	0.00%	0.41%
คำปฏิเสธ	0.14%	0.98%
เครื่องหมายวรรคตอน	0.03%	7.16%

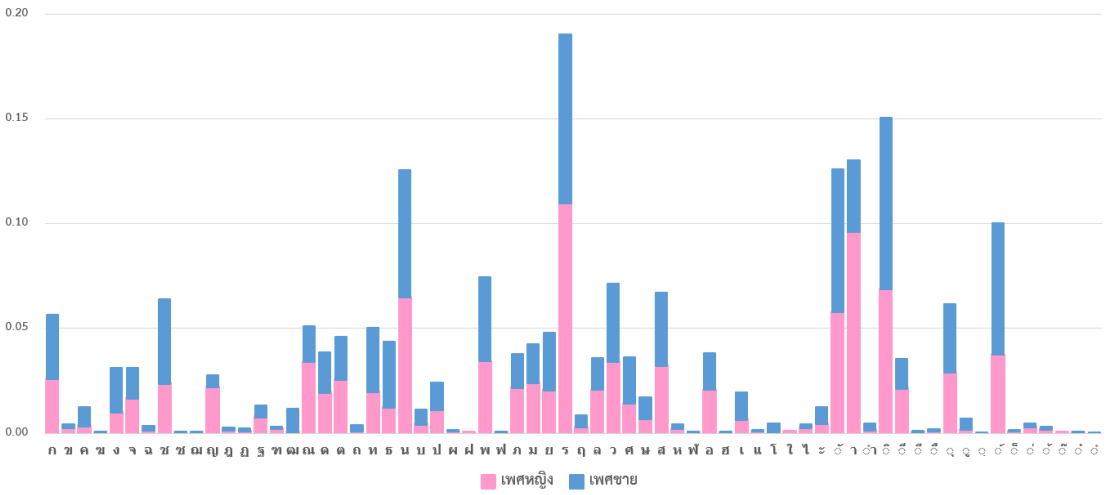
3.3.3 การนับความถี่ตัวอักษรภาษาไทย

ชื่อจริงของคนไทยมีการใช้ตัวอักษรภาษาไทยในชื่อของเพศชายและเพศหญิงที่ต่างกัน เช่น ตัวอักษร ‘ช’ มักจะพบในชื่อจริงของเพศชายมากกว่าเพศหญิง หรือตัวอักษร ‘ญ’ มักจะพบในชื่อจริงของเพศหญิงมากกว่าเพศชาย งานวิจัยนี้จึงได้ทำการสร้างคุณลักษณะการนับความถี่ตัวอักษรภาษาไทย โดยใช้การแยกตัวอักษรทีละตัว ก่อนนำมานับความถี่ของแต่ละตัวอักษร ตัวอย่างข้อมูลชื่อและการนับความถี่ตัวอักษรภาษาไทย ดังภาพที่ 13

ชื่อผู้ใช้งาน	การตัดตัวอักษร
เนตรชนก	['เ','น','ต','ร','ช','น','ก']
เบญจพรรณ	['เ','บ','ญ','จ','พ','ร','ร','ณ']
เสกสรร	['เ','ส','ก','ส','ร','ร']
แจ่มจันทร์	['แ','จ','่','ม','จ','ั','น','ท','ร','์']
สมศักดิ์	['ส','ม','ศ','ั','ก','ด','ิ','์']
กมลลักษณ์	['ก','ม','ล','ล','ั','ก','ษ','ณ','์']
ชัยชาญ	['ช','ั','ย','ช','า','ญ']
วรางคณา	['ว','ร','า','ง','ค','ณ','า']
ชลวิทย์	['ช','ล','ว','ิ','ท','ย','์']
สุนันธิณี	['ส','ุ','น','ั','น','ธ','ิ','ณ','ี']
ชัญญาวีร์	['ช','ั','ญ','ญ','า','ว','ี','ร','์']
กิตติพงษ์	['ก','ิ','ต','ต','ิ','พ','ง','ษ','์']
อภินันท์	['อ','ภ','ิ','น','ั','น','ท','์']
พัชรินทร์	['พ','ั','ช','ร','ิ','น','ท','ร','์']
สมชาย	['ส','ม','ช','า','ย']
อุบลรัตน์	['อ','ุ','บ','ล','ร','ั','ต','น','์']
อรรถพล	['อ','ร','ร','ถ','พ','ล']

ภาพที่ 12 ตัวอย่างข้อมูลชื่อและการนับความถี่ตัวอักษรภาษาไทย

สถิติการนับความถี่ของแต่ละตัวอักษรที่พบในชื่อจริงของเพศชายและเพศหญิงสามารถสรุปได้ดังแผนภาพที่ 14 ซึ่งแสดงให้เห็นว่า ‘ธ’ มักพบบ่อยในชื่อเพศชายมากกว่าเพศหญิง เช่นเดียวกับ ‘ช’, ‘ง’, ‘ิ’ และ ‘์’ ในทางกลับกัน ‘ร’, ‘ณ’, ‘ญ’ และ ‘า’ จะพบในชื่อเพศหญิงบ่อยมากกว่าเพศชาย



ภาพที่ 13 การนับความถี่ตัวอักษรในชื่อจริงของเพศชายและเพศหญิง

ในการวัดว่าโอกาสการพบของแต่ละตัวอักษรในชื่อจริงของเพศชายและเพศหญิง ใช้วิธีคำนวณเอนโทรปี (Entropy) โดยการรวมความน่าจะเป็นการพบตัวอักษรในชื่อจริงของเพศชายและเพศหญิง คูณด้วยล็อก (Log) ของความน่าจะเป็นของแต่ละเพศ ตามสมการที่ 3.1

$$Entropy = (-P_{male} \log_2 P_{female}) + (- P_{female} \log_2 P_{male}) \quad (3.1)$$

โดยถ้าค่าเอนโทรปีสูงแสดงถึงมีการพบตัวอักษรนั้น ๆ ในชื่อจริงทั้ง 2 เพศในทางกลับกัน ถ้าค่าเอนโทรปีต่ำแสดงถึงมีการพบตัวอักษรนั้น ๆ ในชื่อจริงของเพศใดเพศหนึ่งมากกว่า ซึ่งตัวอักษรที่พบในชื่อจริง 10 อันดับแรกที่มีค่าเอนโทรปีที่ต่ำที่สุด ได้แก่ 'ใ', 'ฝ', 'ฒ', 'แ', 'ฌ', 'โ', '์', 'ญ', 'ฑ' และ 'ผ' ตามลำดับ สื่อได้ว่า ได้แก่ 'ฒ' และ 'โ' พบในชื่อจริงของเพศชายมากกว่าเพศหญิง และในทางกลับกัน 'ใ', 'ฝ', 'แ', 'ฌ', '์', 'ญ', 'ฑ' และ 'ผ' พบในชื่อจริงของเพศหญิงมากกว่าเพศชาย แสดงค่าเอนโทรปีที่มีค่าต่ำที่สุดของ 10 ตัวอักษรแรกที่พบในชื่อจริง ตามตารางที่ 7

ตารางที่ 7 ตารางค่าเอนโทรปีที่มีค่าต่ำที่สุดของ 10 ตัวอักษรแรกที่พบในชื่อจริง

ตัวอักษร	จำนวนชื่อเพศชาย	จำนวนชื่อเพศหญิง	เอนโทรปี
ใ	0	12	0.000
ฝ	0	4	0.000
ฒ	174	10	0.305
แ	2	15	0.523
ฌ	1	5	0.65
โ	61	13	0.671
์	4	16	0.722
ญ	75	287	0.739
ฑ	11	38	0.768
ผ	5	17	0.773

3.3.4 การตัดตัวอักษรภาษาไทย

ชื่อจริงของคนไทยมีลักษณะของชื่อเพศชายและเพศหญิงที่ต่างกัน เช่น ชื่อที่ขึ้นต้นด้วย “พล-” มักจะเป็นชื่อเพศชาย ส่วนชื่อที่ขึ้นต้นด้วย “พร-” มักจะเป็นชื่อเพศหญิง หรือชื่อที่ลงท้ายด้วย “-วัฒน์” มักจะเป็นชื่อเพศชาย ส่วนชื่อที่ลงท้ายด้วย “-วรรณ” มักจะเป็นชื่อเพศหญิง งานวิจัยนี้จึงได้ทำการสร้างคุณลักษณะโดยตัดตัวอักษรภาษาไทยด้วยตัวอักษรแรกและตัวอักษรสุดท้าย โดยแบ่งออกเป็น 6 ประเภท ได้แก่ 2 ตัวอักษรแรก, 3 ตัวอักษรแรก, 4 ตัวอักษรแรก, 2 ตัวอักษรสุดท้าย, 3 ตัวอักษรสุดท้าย และ 4 ตัวอักษรสุดท้าย ตัวอย่างการตัดตัวอักษรภาษาไทย ดังภาพที่ 15

ชื่อผู้ใช้งาน	การตัดตัวอักษร
เมตรชนก	{'first2-char': 'เม', 'first3-char': 'เมต', 'first4-char': 'เมตร', 'last2-chars': 'นก', 'last3-chars': 'ชนก', 'last4-chars': 'รชนก'}
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'}
อุบลรัตน์	{'first2-char': 'อุ', 'first3-char': 'อุบ', 'first4-char': 'อุบล', 'last2-chars': 'น์', 'last3-chars': 'ตน์', 'last4-chars': 'ัตน์'}
อรรถพล	{'first2-char': 'อร', 'first3-char': 'อรร', 'first4-char': 'อรรถ', 'last2-chars': 'พล', 'last3-chars': 'ถพล', 'last4-chars': 'รถพล'}

ภาพที่ 14 ตัวอย่างข้อมูลชื่อและการตัดตัวอักษร

ภาพที่ 16 แสดงการตัดตัวอักษรที่พบบ่อยที่สุด 10 อันดับแรกที่พบในชื่อจริงของเพศชายและเพศหญิง โดยชื่อของเพศชายแสดงทางด้านซ้ายและชื่อของเพศหญิงแสดงทางด้านขวา ซึ่งพบว่าชื่อของเพศชายและเพศหญิงมีลักษณะที่แตกต่างกัน โดย 10 อันดับแรกของชื่อเพศชาย ได้แก่ 3 ตัวอักษรสุดท้าย: “ชัย”, 2 ตัวอักษรแรก: “สุ”, 4 ตัวอักษรสุดท้าย: “กดิ์”, 2 ตัวอักษรแรก: “ปร”, 2 ตัวอักษรสุดท้าย: “พล”, 2 ตัวอักษรแรก: “ธน”, 3 ตัวอักษรสุดท้าย: “ัฒน์”, 2 ตัวอักษรแรก: “สม”, 2 ตัวอักษรแรก: “วิ” และ 4 ตัวอักษรสุดท้าย: “พงษ์” ตามลำดับ ส่วน 10 อันดับแรกของชื่อเพศหญิง ได้แก่ 2 ตัวอักษรแรก: “สุ”, 2 ตัวอักษรสุดท้าย: “ดา”,

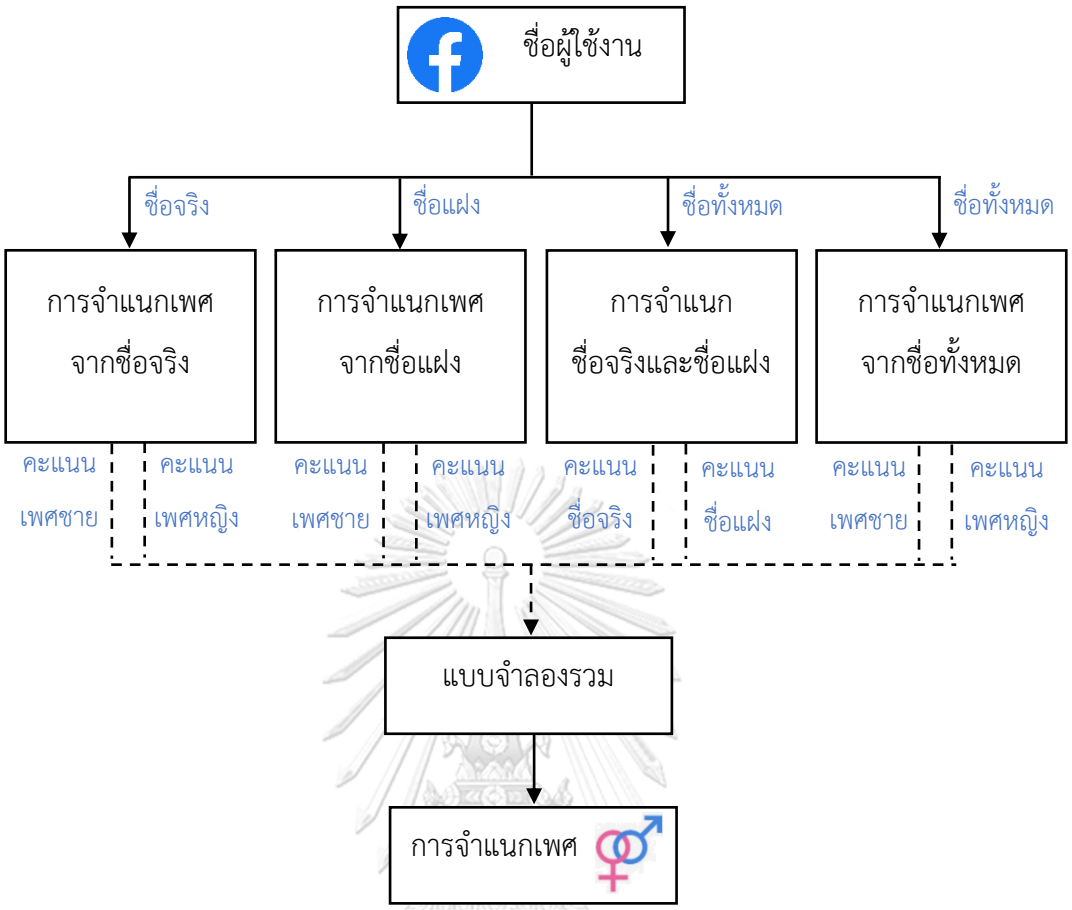
2 ตัวอักษรสุดท้าย: “พร”, 2 ตัวอักษรสุดท้าย: “ณ์”, 4 ตัวอักษรสุดท้าย: “ัตน์”, 3 ตัวอักษรสุดท้าย: “รรณ”, 2 ตัวอักษรสุดท้าย: “รา”, 2 ตัวอักษรแรก: “วร”, 2 ตัวอักษรสุดท้าย: “ภา” และ 2 ตัวอักษรแรก: “ณั” ตามลำดับ

1	3 ตัวอักษรสุดท้าย: “ชัย”	1	2 ตัวอักษรแรก: “สุ”
2	2 ตัวอักษรแรก: “สุ”	2	2 ตัวอักษรสุดท้าย: “ดา”
3	4 ตัวอักษรสุดท้าย: “กติ์”	3	2 ตัวอักษรสุดท้าย: “พร”
4	2 ตัวอักษรแรก: “ปร”	4	2 ตัวอักษรสุดท้าย: “ณ์”
5	2 ตัวอักษรสุดท้าย: “พล”	5	4 ตัวอักษรสุดท้าย: “ัตน์”
6	2 ตัวอักษรแรก: “ธน”	6	3 ตัวอักษรสุดท้าย: “รรณ”
7	4 ตัวอักษรสุดท้าย: “ัฒน์”	7	2 ตัวอักษรสุดท้าย: “รา”
8	2 ตัวอักษรแรก: “สม”	8	2 ตัวอักษรแรก: “วร”
9	2 ตัวอักษรแรก: “วิ”	9	2 ตัวอักษรสุดท้าย: “ภา”
10	4 ตัวอักษรสุดท้าย: “พงษ์”	10	2 ตัวอักษรแรก: “ณั”

ภาพที่ 15 การตัดตัวอักษรที่พบมากที่สุด 10 อันดับแรกของชื่อจริงของเพศชายและเพศหญิง

3.4 การออกแบบการทดลอง

การจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊ก พบว่า ชื่อในเฟซบุ๊กมีการใช้ทั้งชื่อจริงรวมทั้งนามสกุล และยังมีการใช้ชื่อแฝง ซึ่งชื่อทั้ง 2 ประเภท มีลักษณะที่ต่างกัน งานวิจัยนี้จึงทำการสร้างแบบจำลองของชื่อจริงและชื่อแฝงที่ต่างกัน โดยชื่อแต่ละประเภทจะมีรูปแบบที่สามารถบ่งบอกเพศได้ที่แตกต่างกัน จึงมีการเลือกใช้คุณลักษณะที่ต่างกัน แผนภาพกระบวนการทดลองการจำแนกเพศจากชื่อผู้ใช้งาน ดังภาพที่ 17



ภาพที่ 16 แผนภาพกระบวนการทดลองการจำแนกเพศจากชื่อผู้ใช้งาน

ในการทดลอง มีการใช้คุณลักษณะ 4 ประเภท ได้แก่

1. การตัดคำภาษาไทย
2. การจำแนกชนิดของคำภาษาไทย
3. การนับความถี่ตัวอักษรภาษาไทย
4. การตัดตัวอักษรภาษาไทย

โดยได้มีการแบ่งแบบจำลองออกเป็น 4 แบบจำลอง ที่มีข้อมูลนำเข้า และข้อมูลส่งออกที่ต่างกัน โดยมีการใช้คุณลักษณะที่ต่างกันให้เหมาะสมกับแต่ละแบบจำลอง ได้แก่

1. แบบจำลองการจำแนกเพศจากชื่อจริง

ข้อมูลนำเข้าเป็นชื่อผู้ใช้งานที่เป็นชื่อจริง โดยเลือกใช้เฉพาะชื่อแรก ไม่รวมนามสกุล เนื่องจากนามสกุลไม่สามารถนำมาจำแนกเพศได้ โดยมีการใช้คุณลักษณะการนับความถี่ตัวอักษรภาษาไทย และการตัดตัวอักษรภาษาไทย โดยข้อมูลส่งออกเป็นคะแนนชื่อเพศชาย และคะแนนชื่อเพศหญิง

2. แบบจำลองการจำแนกเพศจากชื่อแฝง

ข้อมูลนำเข้าเป็นชื่อผู้ใช้งานที่เป็นชื่อแฝง โดยมีการใช้คุณลักษณะการตัดตัวอักษรภาษาไทย ซึ่งแบบจำลองนี้จะแบ่งกลุ่มของการจำแนกออกเป็น 3 กลุ่ม ได้แก่ เพศชาย เพศหญิง และไม่สามารถระบุเพศได้ โดยเพศชายและเพศหญิง หมายถึง มีค่าความน่าจะเป็นมากกว่า 60% ส่วนชื่อที่ไม่สามารถระบุเพศได้ หมายถึง มีค่าความน่าจะเป็นน้อยกว่า 60% โดยข้อมูลส่งออกเป็นคะแนนชื่อเพศชาย และคะแนนชื่อเพศหญิง

3. แบบจำลองการจำแนกชื่อจริงและชื่อแฝง

ข้อมูลนำเข้าเป็นชื่อผู้ใช้งานทั้งหมด โดยมีการใช้คุณลักษณะการตัดคำภาษาไทย และการจำแนกชนิดของคำภาษาไทย โดยข้อมูลส่งออกเป็นคะแนนชื่อจริง และคะแนนชื่อแฝง เพื่อนำไปใช้ร่วมกับแบบจำลองการจำแนกเพศจากชื่อจริง และแบบจำลองการจำแนกเพศจากชื่อแฝง เนื่องจากชื่อจริงและชื่อแฝงมีลักษณะเฉพาะที่ต่างกัน

4. แบบจำลองการจำแนกเพศจากชื่อผู้ใช้งานทั้งหมด

ข้อมูลนำเข้าเป็นชื่อผู้ใช้งานทั้งหมดทั้งชื่อจริงและชื่อแฝง โดยมีการใช้คุณลักษณะการตัดคำภาษาไทย โดยข้อมูลส่งออกเป็นคะแนนชื่อเพศชาย และคะแนนชื่อเพศหญิง

โดยทั้ง 4 แบบจำลองจะมีการเปรียบเทียบตัวจำแนกประเภทที่ต่างกัน 5 ตัวจำแนก ได้แก่

1. เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor)
2. ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)
3. ป่าแบบสุ่ม (Random Forest)
4. นาอีฟเบย์สแบบเอกนาม (Multinomial Naïve Bayes)
5. โครงข่ายประสาทเทียม (Neural Network)

หลังจากมีการแบ่งแบบจำลองเป็นประเภทต่าง ๆ แล้ว จะมีการนำข้อมูลส่งออกมาเป็นข้อมูลนำเข้าของแบบจำลองรวม โดยแบ่งการทดลองออกเป็น 3 ส่วน

1. การรวมกัน 2 แบบจำลอง ได้แก่ แบบจำลองการจำแนกเพศจากชื่อจริง และแบบจำลองการจำแนกเพศจากชื่อแฝง

2. การรวมกัน 3 แบบจำลอง ได้แก่ แบบจำลองการจำแนกเพศจากชื่อจริง แบบจำลองการจำแนกเพศจากชื่อแฝง และแบบจำลองการจำแนกชื่อจริงและชื่อแฝง

3. การรวมกัน 4 แบบจำลอง ได้แก่ แบบจำลองการจำแนกเพศจากชื่อจริง แบบจำลองการจำแนกเพศจากชื่อแฝง แบบจำลองการจำแนกชื่อจริงและชื่อแฝง และแบบจำลองการจำแนกเพศจากชื่อผู้ใช้งานทั้งหมด

โดยแบบจำลองรวมใช้ตัวจำแนกประเภทโครงข่ายประสาทเทียม (Neural Network) เป็นตัวจำแนก

3.5 การวัดประสิทธิภาพการจำแนกของแบบจำลอง

การแบ่งข้อมูลเพื่อใช้ในการวัดประสิทธิภาพการจำแนกของแบบจำลอง ใช้วิธีการตรวจสอบไขว้ (Crossvalidation Test) และการวัดความแม่นยำ (Accuracy) โดยใช้ 10-Fold Crossvalidation คือ ทำการแบ่งข้อมูลออกเป็น 10 ส่วน ซึ่งแต่ละส่วนจะมีจำนวนข้อมูลที่เท่ากัน โดยข้อมูล 1 ส่วน จะเป็นข้อมูลสำหรับการทดสอบ (Testing set) และอีก 9 ส่วนที่เหลือจะเป็นข้อมูลสำหรับการเรียนรู้ (Training set) โดยทำการแบ่งข้อมูล เพื่อทดสอบประสิทธิภาพการจำแนกของแบบจำลอง 10 รอบ ดังนี้

รอบที่ 1 : ใช้ข้อมูลส่วนที่ 2, 3, 4, 5, 6, 7, 8, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 1 สำหรับการทดสอบ

รอบที่ 2 : ใช้ข้อมูลส่วนที่ 1, 3, 4, 5, 6, 7, 8, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 2 สำหรับการทดสอบ

รอบที่ 3 : ใช้ข้อมูลส่วนที่ 1, 2, 4, 5, 6, 7, 8, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 3 สำหรับการทดสอบ

รอบที่ 4 : ใช้ข้อมูลส่วนที่ 1, 2, 3, 5, 6, 7, 8, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 4 สำหรับการทดสอบ

รอบที่ 5 : ใช้ข้อมูลส่วนที่ 1, 2, 3, 4, 6, 7, 8, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 5 สำหรับการทดสอบ

รอบที่ 6 : ใช้ข้อมูลส่วนที่ 1, 2, 3, 4, 5, 6, 8, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 6 สำหรับการทดสอบ

รอบที่ 7 : ใช้ข้อมูลส่วนที่ 1, 2, 3, 4, 5, 6, 8, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 7 สำหรับการทดสอบ

รอบที่ 8 : ใช้ข้อมูลส่วนที่ 1, 2, 3, 4, 5, 6, 7, 9 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 8 สำหรับการทดสอบ

รอบที่ 9 : ใช้ข้อมูลส่วนที่ 1, 2, 3, 4, 5, 6, 7, 8 และ 10 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 9 สำหรับการทดสอบ

รอบที่ 10 : ใช้ข้อมูลส่วนที่ 1, 2, 3, 4, 5, 6, 7, 8 และ 9 สำหรับการเรียนรู้
ใช้ข้อมูลส่วนที่ 10 สำหรับการทดสอบ

บทที่ 4

ผลการทดลอง

ในบทนี้จะกล่าวถึงผลการทดลองและชุดข้อมูลที่ใช้ในการทดลอง โดยได้นำแนวคิดและวิธีการดำเนินงานในบทที่ 3 มาทดลอง ได้แก่ ชุดข้อมูล ผลการจำแนกเพศจากชื่อโดยใช้คน และผลการทดลองจำแนกเพศโดยใช้แบบจำลอง ซึ่งได้แบ่งผลการทดลองออกเป็นการจำแนกเพศจากชื่อจริง การจำแนกเพศจากชื่อแฝง การจำแนกชื่อจริงและชื่อแฝง การจำแนกเพศจากชื่อทั้งหมด การจำแนกเพศด้วยแบบจำลองรวม และการเปรียบเทียบการจำแนกเพศจากชื่อระหว่างแบบจำลองและคน

4.1 ชุดข้อมูล

ชุดข้อมูลชื่อผู้ใช้งานเฟซบุ๊กที่ดึงข้อมูลมามีทั้งหมด 28,778 ชื่อ จากนั้นเลือกเฉพาะชื่อภาษาไทยได้ 13,429 ชื่อ ซึ่งสามารถหาข้อมูลเพศได้ 11,572 ชื่อ แบ่งเป็นชื่อเพศชาย 6,111 ชื่อ และเพศหญิง 5,461 ชื่อ และแบ่งเป็นชื่อจริง 5,210 ชื่อ และชื่อแฝง 6,362 ชื่อ โดยใช้ข้อมูลในช่วงเวลาเดือนมกราคมถึงเดือนมีนาคม 2562 สรุปจำนวนชุดข้อมูลได้ดังตารางที่ 8

ตารางที่ 8 ตารางจำนวนชุดข้อมูลที่ใช้ในงานวิจัย

	เพศชาย	**เพศหญิง**	**รวม**
ชื่อจริง	2,751	2,459	5,210
ชื่อแฝง	3,360	3,002	6,362
รวม	6,111	5,461	11,572

4.2 การจำแนกเพศจากชื่อโดยใช้คน

ในการทดสอบประสิทธิภาพของแบบจำลอง จะทำการเปรียบเทียบกับบรรทัดฐาน โดยใช้คน 3 คน ในการจำแนกเพศจากชื่อ โดยได้ทำการสุ่มชื่อประมาณ 10% จากชื่อทั้งหมด คิดเป็น 1,200 ชื่อ

โดยแบ่งออกเป็นชื่อจริงเพศชาย 300 ชื่อ ชื่อจริงเพศหญิง 300 ชื่อ ชื่อแฝงเพศชาย 300 ชื่อ และชื่อแฝงเพศหญิง 300 ชื่อ ซึ่งผลการจำแนกเพศจากชื่อโดยคน 3 คน ได้ค่าเฉลี่ยความถูกต้องที่ 77.03% โดยค่าเฉลี่ยความถูกต้องของชื่อจริงที่ 83.61% และค่าเฉลี่ยความถูกต้องของชื่อแฝงที่ 70.44% สรุปผลการจำแนกเพศจากชื่อโดยใช้คน ตามตารางที่ 9

ตารางที่ 9 ตารางผลการจำแนกเพศจากชื่อโดยใช้คน

ประเภทชื่อ	คนที่ 1	คนที่ 2	คนที่ 3	ค่าเฉลี่ย
ชื่อจริง	86.83%	84.17%	79.83%	83.61%
ชื่อแฝง	71.83%	66.83%	72.67%	70.44%
รวม	79.33%	75.50%	76.25%	77.03%

4.3 การจำแนกเพศจากชื่อโดยใช้แบบจำลอง

แบบจำลองที่ทำการทดลองจะเปรียบเทียบตัวจำแนกระหว่างเพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor), ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine), ป่าแบบสุ่ม (Random Forest), นาอีฟเบย์สแบบเอกนาม (Multinomial Naïve Bayes) และโครงข่ายประสาทเทียม (Neural Network) โดยทดสอบประสิทธิภาพด้วยวิธีการตรวจสอบไขว้ (10-Fold Crossvalidation) และเปรียบเทียบด้วยค่าความถูกต้อง โดยได้แบ่งการทดลองออกเป็น 5 การทดลอง ได้แก่

4.3.1 แบบจำลองการจำแนกเพศจากชื่อจริง

แบบจำลองการจำแนกเพศจากชื่อจริงใช้ข้อมูลทั้งหมด 5,210 ชื่อ แบ่งเป็นชื่อเพศชาย 2,751 ชื่อ และชื่อเพศหญิง 2,459 ชื่อ โดยผลการทดลองการจำแนกเพศจากชื่อจริง พบว่าตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนได้ประสิทธิภาพที่ดีที่สุด คือ ได้ความถูกต้องที่ 89.02% รองลงมาเป็นนาอีฟเบย์สแบบเอกนาม (88.59%), โครงข่ายประสาทเทียม (86.71%), เพื่อนบ้านใกล้ที่สุด (85.79%) และป่าแบบสุ่ม (84.31%) ตามลำดับ สรุปผลการทดลองการจำแนกเพศจากชื่อจริงตามตารางที่ 10

ตารางที่ 10 ตารางผลการทดลองการจำแนกเพศจากชื่อจริง

แบบจำลอง	ค่าความถูกต้อง
เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor)	85.79%
ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)	89.02%
ป่าแบบสุ่ม (Random Forest)	84.31%
นาอีฟเบย์สแบบเอกนาม (Multinomial Naïve Bayes)	88.59%
โครงข่ายประสาทเทียม (Neural Network)	86.71%

แบบจำลองการจำแนกเพศจากชื่อจริงของตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนที่ให้ความถูกต้องมากที่สุด พบว่า มีการจำแนกเพศจากชื่อจริงของเพศหญิงถูกต้องมากกว่าเพศชาย โดยการจำแนกเพศชายมีความถูกต้องที่ 88.70% และการจำแนกเพศหญิงมีความถูกต้องที่ 89.39% เมทริกซ์ความสับสนของการจำแนกเพศจากชื่อจริง ตามตารางที่ 11

ตารางที่ 11 ตารางเมทริกซ์ความสับสนของการจำแนกเพศจากชื่อจริง

		ผลการจำแนก	
		เพศชาย	เพศหญิง
ค่าที่แท้จริง	เพศชาย	2,440	311
	เพศหญิง	261	2,198

ตัวอย่างผลการจำแนกเพศจากชื่อจริงที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง ตามตางรางที่ 12 และตัวอย่างผลการจำแนกเพศจากชื่อจริงที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชาย ตามตางรางที่ 13

ตารางที่ 12 ตารางตัวอย่างผลการจำแนกเพศจากชื่อจริงที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
สุรีย์ อาจภักดี	เพศชาย	เพศหญิง
นิทรารัตน์ แพทย์วงศ์	เพศชาย	เพศหญิง
วัชรชนก วงษ์สุวรรณ	เพศชาย	เพศหญิง
ศรีวรรณ สุวรรณมณี	เพศชาย	เพศหญิง
พรพิเศษ ทิมเทศ	เพศชาย	เพศหญิง
สถาพร อินทชาติ	เพศชาย	เพศหญิง
ปารดล กมลหรรษา	เพศชาย	เพศหญิง
วนรัตน์ ซื่อสัตย์	เพศชาย	เพศหญิง
ปาณัท ฉ่ำสูงเนิน	เพศชาย	เพศหญิง
สมพร โสดากุล	เพศชาย	เพศหญิง

ตารางที่ 13 ตารางตัวอย่างผลการจำแนกเพศจากชื่อจริงที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชาย)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
อรุณ ชุบทอง	เพศหญิง	เพศชาย
สหัชญาณ คำมูล	เพศหญิง	เพศชาย
คีตภัทร หอมเขียว	เพศหญิง	เพศชาย
ศรัณรัฏน์ ก๊กผล	เพศหญิง	เพศชาย
ประกายเพ็ชร เงางาม	เพศหญิง	เพศชาย
ดิษยะพรรษ์ ภิญโญพันธุ์	เพศหญิง	เพศชาย
นัยเนตร หมั่นกู้	เพศหญิง	เพศชาย
สุขุมาล สิทธา	เพศหญิง	เพศชาย
สุรพีร์ สมรูป	เพศหญิง	เพศชาย
ณัฐพัชร์ ทองขาว	เพศหญิง	เพศชาย

4.3.2 แบบจำลองการจำแนกเพศจากชื่อแฝง

แบบจำลองการจำแนกเพศจากชื่อแฝงใช้ข้อมูลทั้งหมด 6,362 ชื่อ แบ่งเป็นชื่อเพศชาย 3,360 ชื่อ และชื่อเพศหญิง 3,002 ชื่อ โดยผลการทดลองการจำแนกเพศจากชื่อแฝง พบว่า ตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนได้ประสิทธิภาพที่ดีที่สุด คือ ได้ความถูกต้องที่ 73.41% รองลงมาเป็นนาอีฟเบย์สแบบเอกนาม (70.63%), โครงข่ายประสาทเทียม (69.02%), ป่าแบบสุ่ม (67.21%) และเพื่อนบ้านใกล้ที่สุด (64.32%) ตามลำดับ สรุปผลการทดลองการจำแนกเพศจากชื่อแฝง ตามตารางที่ 14

ตารางที่ 14 ตารางผลการทดลองการจำแนกเพศจากชื่อแฝง

แบบจำลอง	ค่าความถูกต้อง
เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor)	64.32%
ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)	73.41%
ป่าแบบสุ่ม (Random Forest)	67.21%
นาอีฟเบย์สแบบเอกนาม (Multinomial Naïve Bayes)	70.63%
โครงข่ายประสาทเทียม (Neural Network)	69.02%

แบบจำลองการจำแนกเพศจากชื่อแฝงของตัวจำแนกซัพพอร์ตเวกเตอร์แมชชีนที่ให้ความถูกต้องมากที่สุด พบว่า มีการจำแนกเพศจากชื่อแฝงของเพศชายถูกต้องมากกว่าเพศหญิง โดยการจำแนกเพศชายมีความถูกต้องที่ 74.43% และการจำแนกเพศหญิงมีความถูกต้องที่ 72.25% เมทริกซ์ความสับสนของการจำแนกเพศจากชื่อจริง ตามตารางที่ 15

ตารางที่ 15 ตารางเมทริกซ์ความสับสนของการจำแนกเพศจากชื่อแฝง

		ผลการจำแนก	
		เพศชาย	เพศหญิง
ค่าที่แท้จริง	เพศชาย	2,501	859
	เพศหญิง	833	2,169

ตัวอย่างผลการจำแนกเพศจากชื่อแฝงที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง ตามตางรางที่ 16 และตัวอย่างผลการจำแนกเพศจากชื่อแฝงที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชาย ตามตางรางที่ 17

ตารางที่ 16 ตารางตัวอย่างผลการจำแนกเพศจากชื่อแฝงที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
ไก่กา อาลาเล่	เพศชาย	เพศหญิง
รุ่งน้อย เด็กขนมหวาน	เพศชาย	เพศหญิง
กานต์ คนดี มีแค่คนเดียว	เพศชาย	เพศหญิง
หมาน้อย ยูมิ	เพศชาย	เพศหญิง
เด่นนรา ฟ้าแว่บแว๊บ	เพศชาย	เพศหญิง
อยากเห็นคนไทย บินได้	เพศชาย	เพศหญิง
น้องบอก ให้อ้ายฮัก	เพศชาย	เพศหญิง
ดุ๊กดุ่ย ดุ๊กดิ๊ก	เพศชาย	เพศหญิง
น้องปอนนอนน้อย แต่นอนนะ	เพศชาย	เพศหญิง
สายฝน กลางลมหนาว	เพศชาย	เพศหญิง

ตารางที่ 17 ตารางตัวอย่างผลการจำแนกเพศจากชื่อแฝงที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชาย)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
นางฟ้าในคราบซาตาน ไม่ต้องเผือก	เพศหญิง	เพศชาย
ดรีม สุจัดปลัดบอก	เพศหญิง	เพศชาย
กลมดิ๊ก ตัวแสบ	เพศหญิง	เพศชาย
ชื่อหงษ์ พ่อตั้งให้	เพศหญิง	เพศชาย
น้องยุ่งรักปอนด์ คนเดียว	เพศหญิง	เพศชาย
เจ็บเพราะเขา เหงาเพราะเธอ	เพศหญิง	เพศชาย
อวตาร มารน้อย	เพศหญิง	เพศชาย
จุ๊บจิ๊บ คนในจัยเทอร์	เพศหญิง	เพศชาย
ป๋องแป้ง ตาแป๋ว	เพศหญิง	เพศชาย
หากตะวัน ยังเคียงคู่ฟ้า	เพศหญิง	เพศชาย

4.3.3 แบบจำลองการจำแนกชื่อจริงและชื่อแฝง

แบบจำลองการจำแนกชื่อจริงและชื่อแฝงใช้ข้อมูลทั้งหมด 11,572 ชื่อ แบ่งเป็นชื่อจริง 5,210 ชื่อ และชื่อแฝง 6,362 ชื่อ โดยผลการทดลองการจำแนกชื่อจริงและชื่อแฝง พบว่า นาอีฟเบย์สแบบเอกนามได้ประสิทธิภาพที่ดีที่สุด คือ ได้ความถูกต้องที่ 89.24% รองลงมาเป็นซัพพอร์ตเวกเตอร์แมชชีน (86.11%), โครงข่ายประสาทเทียม (85.48%), ป่าแบบสุ่ม (82.85%) และเพื่อนบ้านใกล้ที่สุด (81.49%) ตามลำดับ สรุปผลการทดลองการจำแนกชื่อจริงและชื่อแฝงตามตารางที่ 18

ตารางที่ 18 ตารางผลการทดลองการจำแนกชื่อจริงและชื่อแฝง

แบบจำลอง	ค่าความถูกต้อง
เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor)	81.49%
ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)	86.11%
ป่าแบบสุ่ม (Random Forest)	82.85%
นาอีฟเบย์สแบบเอกนาม (Multinomial Naïve Bayes)	89.24%
โครงข่ายประสาทเทียม (Neural Network)	85.48%

แบบจำลองการจำแนกชื่อจริงและชื่อแฝงของตัวจำแนกนาอีฟเบย์สแบบเอกนามที่ให้ความถูกต้องมากที่สุด พบว่า มีการจำแนกชื่อจริงถูกต้องมากกว่าชื่อแฝง โดยการจำแนกชื่อจริงมีความถูกต้องที่ 92.15% และการจำแนกชื่อแฝงมีความถูกต้องที่ 86.86% เมทริกซ์ความสับสนของการจำแนกชื่อจริงและชื่อแฝง ตามตารางที่ 19

ตารางที่ 19 ตารางเมทริกซ์ความสับสนของการจำแนกชื่อจริงและชื่อแฝง

		ผลการจำแนก	
		ชื่อจริง	ชื่อแฝง
ค่าที่แท้จริง	ชื่อจริง	4,801	409
	ชื่อแฝง	836	5,526

ตัวอย่างผลการจำแนกชื่อจริงและชื่อแฝงที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นชื่อจริง แต่ผลการจำแนกเป็นชื่อแฝง ตามตางรางที่ 20 และตัวอย่างผลการจำแนกชื่อจริงและชื่อแฝงที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นชื่อแฝง แต่ผลการจำแนกเป็นชื่อจริง ตามตางรางที่ 21

ตารางที่ 20 ตารางตัวอย่างผลการจำแนกชื่อจริงและชื่อแฝงที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นชื่อจริง แต่ผลการจำแนกเป็นชื่อแฝง)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
นลิน หิ่งห้อยดี	ชื่อจริง	ชื่อแฝง
ถนอม คงยิ้มละมัย	ชื่อจริง	ชื่อแฝง
ทองหล่อ เที่ยงธรรม	ชื่อจริง	ชื่อแฝง
นงนุช หล่อเลี่ยม	ชื่อจริง	ชื่อแฝง
กีรติ พรรื่นเริง	ชื่อจริง	ชื่อแฝง
ชนิดา เซ็งขุนทด	ชื่อจริง	ชื่อแฝง
ดนตรี สุทธิพิบูลย์	ชื่อจริง	ชื่อแฝง
ธวัช เยาวยัง	ชื่อจริง	ชื่อแฝง
นภาพร อิ่มสุข	ชื่อจริง	ชื่อแฝง
วสัน ไพเราะ	ชื่อจริง	ชื่อแฝง

ตารางที่ 21 ตารางตัวอย่างผลการจำแนกชื่อจริงและชื่อแฝงที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นชื่อแฝง แต่ผลการจำแนกเป็นชื่อจริง)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
จันทร์ เจ้าเอย	ชื่อแฝง	ชื่อจริง
แก้วตา ดวงใจ	ชื่อแฝง	ชื่อจริง
ชีวิต ติดธรรมชาติ	ชื่อแฝง	ชื่อจริง
กานต์นิพนธ์ คนช่างฝัน	ชื่อแฝง	ชื่อจริง
เกษร คนดนตรี	ชื่อแฝง	ชื่อจริง
ดาวรุ่ง พุ่งพรวด.	ชื่อแฝง	ชื่อจริง
ณัชชา เซมากูเตะ	ชื่อแฝง	ชื่อจริง
มานะมานี ปิติชูใจ	ชื่อแฝง	ชื่อจริง
นักรบ เขาตะนาวศรี	ชื่อแฝง	ชื่อจริง
สำคัญ บางเวลา	ชื่อแฝง	ชื่อจริง

4.3.4 แบบจำลองการจำแนกเพศจากชื่อทั้งหมด

แบบจำลองการจำแนกเพศจากชื่อทั้งหมดใช้ข้อมูลทั้งหมด 11,572 ชื่อ แบ่งเป็นชื่อเพศชาย 6,111 ชื่อ และชื่อเพศหญิง 5,461 ชื่อ โดยผลการทดลองการจำแนกเพศจากชื่อทั้งหมด พบว่า นาอีฟเบย์สแบบเอกนามได้ประสิทธิภาพที่ดีที่สุด คือ ได้ความถูกต้องที่ 69.18% รองลงมาเป็น ซัพพอร์ตเวกเตอร์แมชชีน (66.53%), ป่าแบบสุ่ม (64.99%), โครงข่ายประสาทเทียม (64.71%) และเพื่อนบ้านใกล้ที่สุด (59.40%) ตามลำดับ สรุปผลการทดลองการจำแนกเพศจากชื่อทั้งหมด ตามตารางที่ 22

ตารางที่ 22 ตารางผลการทดลองการจำแนกเพศจากชื่อทั้งหมด

แบบจำลอง	ค่าความถูกต้อง
เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor)	59.40%
ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)	66.53%
ป่าแบบสุ่ม (Random Forest)	64.99%
นาอีฟเบย์สแบบเอกนาม (Multinomial Naïve Bayes)	69.18%
โครงข่ายประสาทเทียม (Neural Network)	64.71%

แบบจำลองการจำแนกเพศจากชื่อทั้งหมดของตัวจำแนกนาอีฟเบย์สแบบเอกนามที่ให้ความถูกต้องมากที่สุด พบว่ามีการจำแนกเพศจากชื่อทั้งหมดของเพศหญิงถูกต้องมากกว่าเพศชาย โดยการจำแนกเพศชายมีความถูกต้องที่ 68.55% และการจำแนกเพศหญิงมีความถูกต้องที่ 69.90% เมทริกซ์ความสับสนของการจำแนกเพศจากชื่อทั้งหมด ตามตารางที่ 23

ตารางที่ 23 ตารางเมทริกซ์ความสับสนของการจำแนกเพศจากชื่อทั้งหมด

		ผลการจำแนก	
		เพศชาย	เพศหญิง
ค่าที่แท้จริง	เพศชาย	4,189	1,922
	เพศหญิง	1,644	3,817

ตัวอย่างผลการจำแนกเพศจากชื่อทั้งหมดที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง ตามตางรางที่ 24 และตัวอย่างผลการจำแนกเพศจากชื่อทั้งหมดที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชาย ตามตางรางที่ 25

ตารางที่ 24 ตารางตัวอย่างผลการจำแนกเพศจากชื่อทั้งหมดที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
สุขใจ เสมอ เมื่อมีเธอ อยู่เคียงข้าง	เพศชาย	เพศหญิง
ณัฐกานต์ ละอองนวล	เพศชาย	เพศหญิง
สมหมาย สุวรรณมณี	เพศชาย	เพศหญิง
อ้าก สปุ๊กนิก ปาปิยองกุ๊กุ๊	เพศชาย	เพศหญิง
ซูกัส ผูกโบว์	เพศชาย	เพศหญิง
ปามม เหมือนเดิม.	เพศชาย	เพศหญิง
สมพร หอมรสระรื่น	เพศชาย	เพศหญิง
ปุณณัตถ์กิจ หอมทน	เพศชาย	เพศหญิง
ยืนงง ในดงหญิง	เพศชาย	เพศหญิง
โพธิ์ พระกานดา	เพศชาย	เพศหญิง

ตารางที่ 25 ตารางตัวอย่างผลการจำแนกเพศจากชื่อทั้งหมดที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชาย)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
สมพักตร์ หอมหวล	เพศหญิง	เพศชาย
ทับทิมสีแดง กับ ความสำเร็จ	เพศหญิง	เพศชาย
สมถวิล ขี่พายุทะลุอวกาศ	เพศหญิง	เพศชาย
สุนทรี แซ่หล่อ	เพศหญิง	เพศชาย
จุ๊กกรู๊ชู บีดูวับ	เพศหญิง	เพศชาย
นันทภพ ธรภาทองใส	เพศหญิง	เพศชาย
ธนวรรณ ลีลานุตร	เพศหญิง	เพศชาย
หลงรักคนหล่อ	เพศหญิง	เพศชาย
คุณชาย เชอรี่	เพศหญิง	เพศชาย
บุญส่ง มาตรวิจิตร	เพศหญิง	เพศชาย

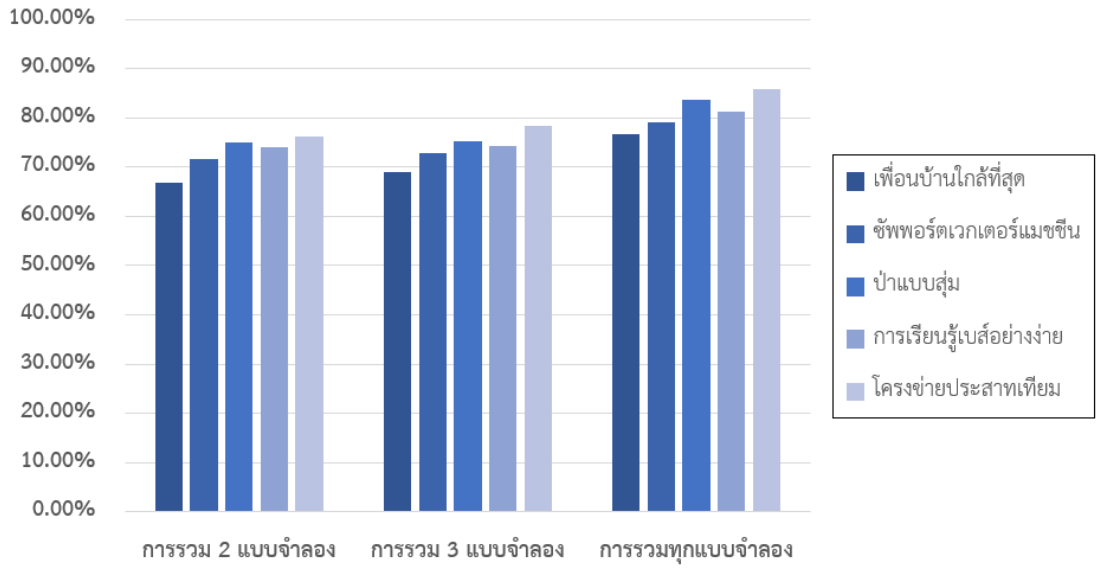
4.3.5 แบบจำลองการจำแนกเพศด้วยการรวมแบบจำลอง

แบบจำลองการจำแนกเพศด้วยการรวมแบบจำลองจากชื่อผู้ใช้งานจะทำการเปรียบเทียบระหว่าง 5 ตัวจำแนก ได้แก่ เพื่อนบ้านใกล้ที่สุด, ซัพพอร์ตเวกเตอร์แมชชีน, ป่าแบบสุ่ม, นาอีฟเบย์สแบบเอกนาม และโครงข่ายประสาทเทียม และยังได้ทำการเปรียบเทียบการรวมกันของ 3 ชุดข้อมูลนำเข้าที่ต่างกัน ได้แก่ การรวมทุกแบบจำลอง การรวมกัน 3 แบบจำลอง และการรวมกัน 2 แบบจำลอง ซึ่งประสิทธิภาพของการรวมกันทุกแบบจำลองได้ความถูกต้องที่มากที่สุดที่ 85.85% ตามด้วยการรวมกัน 3 แบบจำลอง และการรวมกัน 2 แบบจำลอง ตามลำดับ โดยใช้ตัวจำแนกโครงข่ายประสาทเทียมเช่นเดียวกัน สรุปผลการทดลองการจำแนกเพศด้วยการรวมแบบจำลอง ตามตารางที่ 26

ตารางที่ 26 ตารางผลการทดลองการจำแนกเพศด้วยการรวมแบบจำลอง

แบบจำลอง	ค่าความถูกต้องการรวม 2 แบบจำลอง	ค่าความถูกต้องการรวม 3 แบบจำลอง	ค่าความถูกต้องการรวมทุกแบบจำลอง
เพื่อนบ้านใกล้ที่สุด (K-Nearest Neighbor)	66.88%	69.05%	76.64%
ซัพพอร์ตเวกเตอร์แมชชีน (Support Vector Machine)	71.64%	72.80%	79.19%
ป่าแบบสุ่ม (Random Forest)	75.09%	75.31%	83.59%
นาอีฟเบย์สแบบเอกนาม (Multinomial Naïve Bayes)	73.99%	74.19%	81.24%
โครงข่ายประสาทเทียม (Neural Network)	76.14%	78.40%	85.85%

ผลการทดลองความถูกต้องของการรวมกันของแบบจำลอง โดยการรวมกัน 2 แบบจำลอง ได้ความถูกต้องมากที่สุดที่ 76.14% โดยใช้ตัวจำแนกโครงข่ายประสาทเทียม ในการรวมกัน 3 แบบจำลอง ได้ความถูกต้องมากที่สุดที่ 78.40 % โดยใช้ตัวจำแนกโครงข่ายประสาทเทียม และในการรวมกันทุกแบบจำลอง ได้ความถูกต้องมากที่สุดที่ 85.85% โดยใช้ตัวจำแนกโครงข่ายประสาทเทียม โดยความถูกต้องสามารถเรียงลำดับแบบจำลองได้ดังนี้ โครงข่ายประสาทเทียม, ป่าแบบสุ่ม, นาอีฟเบย์สแบบเอกนาม, ซัพพอร์ตเวกเตอร์แมชชีน และ เพื่อนบ้านใกล้ที่สุด ตามลำดับทั้งการรวมกับของแบบจำลองทั้ง 3 แบบ สรุปผลได้ดังภาพที่ 18



ภาพที่ 17 แผนภาพผลการทดลองการจำแนกเพศจากชื่อด้วยการรวมแบบจำลอง

แบบจำลองการจำแนกเพศจากชื่อด้วยการรวมแบบจำลองทุกแบบจำลองของตัวจำแนกโครงข่ายประสาทเทียมที่ให้ความถูกต้องมากที่สุด พบว่า มีการจำแนกเพศจากชื่อทั้งหมดของเพศหญิงถูกต้องมากกว่าเพศชาย โดยการจำแนกเพศชายมีความถูกต้องที่ 85.04% และการจำแนกเพศหญิงมีความถูกต้องที่ 86.76% เมทริกซ์ความสับสนของการจำแนกเพศจากชื่อทั้งหมดตามตารางที่ 27

ตารางที่ 27 ตารางเมทริกซ์ความสับสนของการจำแนกเพศด้วยการรวมแบบจำลอง

		ผลการจำแนก	
		เพศชาย	เพศหญิง
ค่าที่แท้จริง	เพศชาย	5,197	914
	เพศหญิง	723	4,738

ตัวอย่างผลการจำแนกเพศด้วยการรวมแบบจำลองที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง ตามตางรางที่ 28 และตัวอย่างผลการจำแนกเพศด้วยการรวมแบบจำลองที่ไม่ถูกต้อง โดยค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชายตามตางรางที่ 29

ตารางที่ 28 ตารางตัวอย่างผลการจำแนกเพศด้วยการรวมแบบจำลองที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศชาย แต่ผลการจำแนกเป็นเพศหญิง)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
ผู้ชายคนนี้ ไม่ได้มีไว้ให้เธอหลอก	เพศชาย	เพศหญิง
พรศักดิ์ ดิษยนันทน์	เพศชาย	เพศหญิง
รณวรรณ ยอดแก้ว	เพศชาย	เพศหญิง
นนท์นี่ ผู้ฆ่าหมีด้วยมือเปล่า	เพศชาย	เพศหญิง
วรินทร รักษากุล	เพศชาย	เพศหญิง
น้องโมไม่เคยแคร์ใคร ไม่มีรอยสักชอบกินแต่ผัก	เพศชาย	เพศหญิง
อ๊อฟ แม่บอกให้ กลับบ้าน	เพศชาย	เพศหญิง
ฉัตริน อวริสาร	เพศชาย	เพศหญิง
หมูบอล โลกสวย	เพศชาย	เพศหญิง
หยกฟ้า วรรณแก้ว	เพศชาย	เพศหญิง

ตารางที่ 29 ตารางตัวอย่างผลการจำแนกเพศด้วยการรวมแบบจำลองที่ไม่ถูกต้อง (ค่าที่แท้จริงเป็นเพศหญิง แต่ผลการจำแนกเป็นเพศชาย)

ชื่อผู้ใช้งาน	ค่าที่แท้จริง	ผลการจำแนก
แจนยูตุ๊ดตุ่ย ตะลุยอวกาศ	เพศหญิง	เพศชาย
ศรณ์ธรา เกไกรษร	เพศหญิง	เพศชาย
พชรมน โยธา	เพศหญิง	เพศชาย
ตูนมันซื่อ ไม่ดื้อหรอกพี่	เพศหญิง	เพศชาย
ชยานันต์ สุขุมลจันทร์	เพศหญิง	เพศชาย
พักตร์ษร สิรบุณยภัค	เพศหญิง	เพศชาย
เพชรคุง ไงจะใครละ	เพศหญิง	เพศชาย
หมาน้อย ธรรมดา	เพศหญิง	เพศชาย
ต้อม กะล่อน	เพศหญิง	เพศชาย
อณัฐภิชา มูลการณ์	เพศหญิง	เพศชาย

4.4 การเปรียบเทียบการจำแนกเพศจากชื่อระหว่างแบบจำลองและคน

การเปรียบเทียบการจำแนกเพศจากชื่อระหว่างแบบจำลองและคน โดยทำการเปรียบเทียบระหว่างชื่อจริง ชื่อแฝง และรวมทั้งชื่อจริงและชื่อแฝง โดยผลความถูกต้องของการจำแนก พบว่าการจำแนกด้วยแบบจำลองมีผลความถูกต้องรวมที่ 85.85% ส่วนการจำแนกด้วยคนมีความถูกต้องเฉลี่ยที่ 77.03% ซึ่งผลการจำแนกด้วยแบบจำลองมีความถูกต้องมากกว่าการจำแนกด้วยคนในทุกประเภทของชื่อ โดยผลการเปรียบเทียบความถูกต้อง ตามตารางที่ 30

ตารางที่ 30 ตารางผลการเปรียบเทียบการจำแนกเพศจากชื่อระหว่างแบบจำลองและคน

ประเภทชื่อ	คน	แบบจำลอง
ชื่อจริง	83.61%	89.02%
ชื่อแฝง	70.44%	73.41%
รวม	77.03%	85.85%

บทที่ 5
สรุปผลและอภิปรายผลการทดลอง

ในบทนี้จะกล่าวถึงการสรุปผลจากผลการทดลองการจำแนกเพศจากชื่อผู้ใช้งานเฟซบุ๊กด้วยแบบจำลองแบบต่าง ๆ ในบทที่ 4 แล้วนำมาอภิปรายผลการทดลอง รวมทั้งวิเคราะห์ปัญหาและอุปสรรค ข้อเสนอแนะ และผลงานวิจัยที่ได้รับการตีพิมพ์

5.1 สรุปและอภิปรายผลการทดลอง

ข้อมูลบนเฟซบุ๊กสามารถนำไปวิเคราะห์และศึกษาได้มากมาย แต่ข้อมูลมักจะไม่มีรายละเอียดข้อมูลประชากรของผู้ใช้งาน ส่งผลให้การนำข้อมูลของผู้ใช้งานไปใช้วิเคราะห์หรือสื่อสารเกิดความคลาดเคลื่อน หรือไม่ตรงกับกลุ่มเป้าหมาย โดยงานวิจัยนี้มีวัตถุประสงค์เพื่อจำแนกเพศจากชื่อผู้ใช้งานเฟซบุ๊ก เพื่อช่วยแก้ปัญหาการนำข้อมูลไปใช้ไม่ถูกต้อง และเพื่อเพิ่มให้การนำข้อมูลไปใช้ให้เกิดประโยชน์มากที่สุด

งานวิจัยนี้แสดงให้เห็นว่าชื่อของคนไทยมีรูปแบบที่สามารถจำแนกเพศได้ โดยได้ทำการเสนอวิธีการจำแนกเพศจากชื่อผู้ใช้งานเฟซบุ๊กด้วยวิธีการรวมแบบจำลอง แทนการใช้แบบจำลองเดียวกันในทุกคุณลักษณะ โดยมีการจัดกลุ่มคุณลักษณะที่เกี่ยวข้องกันในแบบจำลองเดียวกัน เนื่องจากชื่อจริงและชื่อแฝงจะมีลักษณะที่แตกต่างกัน โดยแบ่งออกเป็น 4 แบบจำลอง ได้แก่ แบบจำลองการจำแนกเพศจากชื่อจริง แบบจำลองการจำแนกเพศจากชื่อแฝง แบบจำลองการจำแนกชื่อจริงและชื่อแฝง และแบบจำลองการจำแนกเพศจากชื่อทั้งหมด ซึ่งผลลัพธ์ของแต่ละแบบจำลองจะถูกนำมาเป็นข้อมูลนำเข้าของแบบจำลองสุดท้าย โดยได้แบ่งการรวมกันของแบบจำลองเป็น 3 รูปแบบ ได้แก่ การรวมกัน 2 แบบจำลอง การรวมกัน 3 แบบจำลอง และการรวมกันทุกแบบจำลอง ซึ่งคุณลักษณะที่ใช้ ประกอบไปด้วย การตัดคำภาษาไทย การจำแนกชนิดของคำภาษาไทย การนับความถี่ตัวอักษรภาษาไทย และการตัดตัวอักษรภาษาไทย

ผลการทดลองแสดงให้เห็นว่าการจำแนกเพศจากชื่อผู้ใช้งานภาษาไทยบนเฟซบุ๊กโดยแบบจำลองการจำแนกเพศของชื่อจริง มีความถูกต้องอยู่ที่ 89.02% ส่วนแบบจำลองการจำแนกเพศของชื่อแฝงมีความถูกต้องอยู่ที่ 73.41% ซึ่งชื่อแฝงมีความถูกต้องที่น้อยกว่าชื่อจริง เนื่องจากชื่อจริง

มีรูปแบบตายตัวที่มากกว่า และชื่อแฝงมีความซับซ้อนที่มากกว่า และมีอิสระในการตั้งชื่อที่มากกว่า โดยเมื่อทำการรวมแบบจำลอง ได้ผลลัพธ์ที่มีความถูกต้องมากที่สุดที่ 85.85% ซึ่งเมื่อนำไปเปรียบเทียบกับการจำแนกเพศโดยใช้คน ได้ความถูกต้องที่ 77.03% ซึ่งแบบจำลองสามารถจำแนกเพศได้ความแม่นยำที่มากกว่าการจำแนกโดยใช้คน

5.2 ปัญหาและอุปสรรค

ข้อมูลที่ใช้ในการวิจัยนี้เก็บรวบรวมจากเฟซบุ๊ก ซึ่งได้ทำการเก็บข้อมูลผ่านเครื่องมือซิลิเนียม ซึ่งเฟซบุ๊กได้มีการเปลี่ยนองค์ประกอบ (Element) ที่แสดงผลบนหน้าเว็บอยู่ตลอด ทำให้ต้องมีการพัฒนาโปรแกรมหลายครั้งให้รองรับกับการเปลี่ยนแปลง

5.3 ข้อเสนอแนะ

จากการเก็บข้อมูลชื่อผู้ใช้งานเฟซบุ๊กของคนไทย พบว่า นิยมตั้งชื่อด้วยตัวอักษรภาษาอังกฤษเป็นจำนวนมาก โดยงานวิจัยนี้เลือกศึกษาเฉพาะชื่อที่มีตัวอักษรภาษาไทย ดังนั้น แนวทางในการทำงานวิจัยต่อไปควรเพิ่มให้สามารถจำแนกเพศจากชื่อผู้ใช้งานคนไทยที่เป็นภาษาอังกฤษได้ ซึ่งสามารถนำวิธีการของงานวิจัยนี้ไปประยุกต์ใช้กับตัวอักษรภาษาอังกฤษได้ โดยเพิ่มการปรับเปลี่ยนคุณลักษณะให้เหมาะสม และยังพบว่าชื่อผู้ใช้งานเฟซบุ๊กของคนรุ่นใหม่ มีรูปแบบที่เปลี่ยนแปลงไปจากคนรุ่นก่อน ซึ่งควรจะต้องมีการเพิ่มคุณลักษณะให้รองรับเพิ่มขึ้น

5.4 ผลงานวิจัยที่ได้รับการตีพิมพ์

งานวิจัยนี้ได้รับการคัดเลือกและตีพิมพ์เป็นบทความวิชาการเรื่อง “Gender Classification of Thai Facebook Usernames” โดย สุพิชชา ยืนยงค์ และ สุกรี สิญธุภิญโญ และได้ไปนำเสนอผลงานในงานประชุมวิชาการ “2019 3rd Asia Conference on Machine Learning and Computing (ACMLC 2019)” ซึ่งจัดขึ้นที่เขตปกครองพิเศษฮ่องกง ระหว่างวันที่ 7-9 ธันวาคม 2562

รายการอ้างอิง

[1] กองราช, ภ., "การศึกษาพฤติกรรมการใช้เครือข่ายสังคมออนไลน์", วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาการบริหารเทคโนโลยี วิทยาลัยนวัตกรรม มหาวิทยาลัยธรรมศาสตร์, 2011.

[2] วิจิตรบุญยรักษ์, พ., "สื่อสังคมออนไลน์ สื่อแห่งอนาคต", Excusive Journal, 2011.

[3] We Are Social Ltd., "*Global and Thailand digital report 2019",* 2019. [ออนไลน์]. https://wearesocial.com/global-digital-report-2019

[4] Cesare, N., et al., "*Demographics in Social Media Data for Public Health Research: Does it matter?",* arXiv preprint arXiv:1710.11048, 2017.

[5] Schwartz, H.A., et al., "*Personality, gender, and age in the language of social media: The open-vocabulary approach",* PloS one, 2013. 8(9): p. e73791.

[6] Alowibdi, J.S., U.A. Buy, and P. Yu, "*Empirical Evaluation of Profile Characteristics for Gender Classification on Twitter*", in *Proceedings of the 2013 12th International Conference on Machine Learning and Applications - Volume 01*. 2013, IEEE Computer Society. p. 365-369.

[7] Bergsma, S., et al. "*Broadly improving user classification via communication-based name and location clustering on twitter*", in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2013.

[8] Akbar, R., "*Gender Classification of Indonesian Names Using Multinomial Naive Bayes and Random Forrest Classifiers".* 2016.

[9] Septiandri, A.A., "*Predicting the gender of Indonesian names",* arXiv preprint arXiv:1707.07129, 2017.

[10] Briedienė, M. and J. Kapočiutė-Dzikienė, "*An Automatic Author Profiling from Non-Normative Lithuanian Texts",* 2018.

[11] Hirt, R., N. Kühl, and G. Satzger, "*Cognitive computing for customer profiling: meta classification for gender prediction",* Electronic Markets, 2019. 29(1): p. 93-106.

[12] Vicente, M., F. Batista, and J.P. Carvalho, "*Gender detection of Twitter users based on multiple information sources*", in *Interactions Between Computational Intelligence and Mathematics Part 2*. 2019, Springer. p. 39-54.

[13] สำนักงานราชบัณฑิตยสภา, เฟซบุ๊ก, 2011, [ออนไลน์]. http://www.royin.go.th/?knowledges=เฟซบุ๊ก-๒๒-กรกฎาคม-๒๕๕๔

[14] สว่างตระกูล, ว., "การศึกษาภาษาที่ใช้ในการตั้งชื่อของคนไทยในกรุงเทพมหานคร", วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาภาษาศาสตร์ คณะศิลปศาสตร์ มหาวิทยาลัยธรรมศาสตร์, 1997.

[15] ศิริวัฒนะนาวิน, ว., "การศึกษาการตั้งชื่อของคนไทย", วิทยานิพนธ์ปริญญามหาบัณฑิต สาขาวิชาภาษาไทย คณะอักษรศาสตร์ มหาวิทยาลัยศิลปากร, 2001.

[16] ศรเลิศล้ำวาณิช, ว., "การตัดคำในระบบแปลภาษา (*Word Segmentation for Thai in Machine Translation System)*", NECTEC, 1993.

[17] อุราธรรมกุล, ป., "การตัดคำภาษาไทยด้วยกฎปรับปรุงและพจนานุกรมแบบใหม่", วิทยานิพนธ์ปริญญามหาบัณฑิต วิศวกรรมคอมพิวเตอร์ มหาวิทยาลัยขอนแก่น, 2007.

[18] Theeramunkong, T., et al. "*Character cluster based Thai information retrieval",* in *Proceedings of the fifth international workshop on on Information retrieval with Asian languages*. 2000. ACM.

[19] พระยาอุปกิตศิลปสาร, "หลักภาษาไทย : (อักขรวิธี วจีวิภาค วากยสัมพันธ์ ฉันทลักษณ์)", 1995.

[20] ภาณุพงศ์, ว., "หนังสืออุเทศภาษาไทย ชุด บรรทัดฐานภาษาไทย เล่ม *3",* สถาบันภาษาไทย สำนักวิชาการและมาตรฐานการศึกษา สำนักงานคณะกรรมการการศึกษาขั้นพื้นฐาน กระทรวงศึกษาธิการ, 2009.

[21] Peterson, L.E., "*K-nearest neighbor",* Scholarpedia, 2009. 4(2): p. 1883.

[22] Suykens, J.A. and J. Vandewalle, "*Least squares support vector machine classifiers",* Neural processing letters, 1999. 9(3): p. 293-300.

[23] Breiman, L., "*Random forests",* Machine learning, 2001. 45(1): p. 5-32.

[24] Mitchell, T., "*Machine Learning",* McGraw Hill, 2017.

[25] กิจศิริกุล, บ., "ปัญญาประดิษฐ์ *Artificial Intelligence",* เอกสารคำสอนวิชา 2110654 ภาควิชาวิศวกรรมคอมพิวเตอร์ จุฬาลงกรณ์มหาวิทยาลัย, 2005

[26] Haykin, S., "*Neural networks: a comprehensive foundation",* 1994: Prentice Hall PTR.

[27] Visa, S., et al., "*Confusion Matrix-based Feature Selection",* MAICS, 2011. 710: p. 120-127.

[28] Sornlertlamvanich, V., T. Charoenporn, and H. Isahara, "*ORCHID: Thai part-of-speech tagged corpus",* National Electronics and Computer Technology Center Technical Report, 1997: p. 5-19.

บรรณานุกรม

ประวัติผู้เขียน

ชื่อ-สกุล	สุพิชชา ยืนยงค์
วัน เดือน ปี เกิด	11 พฤษภาคม 2533
สถานที่เกิด	สุพรรณบุรี
วุฒิการศึกษา	ปริญญาตรี วิศวกรรมศาสตรบัณฑิต (เกียรตินิยมอันดับ 1) ภาควิชาวิศวกรรมคอมพิวเตอร์ คณะวิศวกรรมศาสตร์ มหาวิทยาลัยเทคโนโลยีพระจอมเกล้าธนบุรี
ที่อยู่ปัจจุบัน	358/560 ประชาราษฎร์ 1 บางซื่อ กรุงเทพมหานคร 10800
ผลงานตีพิมพ์	“Gender Classification of Thai Facebook Usernames” โดย สุพิชชา ยืนยงค์ และ สุกรี สิญธุภิญโญ งานประชุมวิชาการ “2019 3rd Asia Conference on Machine Learning and Computing (ACMLC 2019)” จัดขึ้นที่เขตปกครองพิเศษฮ่องกง ระหว่างวันที่ 7-9 ธันวาคม 2562