การจำแนกเพศจากข้อความบนโซเชียลเน็ตเวิร์ก

GENDER CLASSIFICATION FROM TEXT DATA IN SOCIAL NETWORK

เอกภพ พูลสวัสดิ์

บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ

2563

การจำแนกเพศจากข้อความบนโซเชียลเน็ตเวิร์ก

เอกภพ พูลสวัสดิ์

สารนิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตร
วิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ
คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒ
ปีการศึกษา 2563

GENDER CLASSIFICATION FROM TEXT DATA IN SOCIAL NETWORK

EKKAPOB POONSAWAT

A Master's Project Submitted in Partial Fulfillment of the Requirements

for the Degree of MASTER OF SCIENCE

(Information Technology)

Faculty of Science, Srinakharinwirot University

2020

สารนิพนธ์

เรื่อง

การจำแนกเพศจากข้อความบนโซเชียลเน็ตเวิร์ก

ของ

เอกภพ พูลสวัสดิ์

ได้รับอนุมัติจากบัณฑิตวิทยาลัยให้นับเป็นส่วนหนึ่งของการศึกษาตามหลักสูตร

ปริญญาวิทยาศาสตรมหาบัณฑิต สาขาวิชาเทคโนโลยีสารสนเทศ

ของมหาวิทยาลัยศรีนครินทรวิโรฒ

(รองศาสตราจารย์ นายแพทย์ฉัตรชัย เอกปัญญาสกุล)

คณบดีบัณฑิตวิทยาลัย

คณะกรรมการสอบปากเปล่าสารนิพนธ์

..... ที่ปรึกษาหลัก

(ผู้ช่วยศาสตราจารย์ ดร.วีรยุทธ เจริญเรืองกิจ)

..... ประธาน

(อาจารย์ ดร.ธรรมศักดิ์ เธียรนิเวศน์)

..... กรรมการ

(อาจารย์ ดร.โสภณ มงคลลักษมี)

ชื่อเรื่อง	การจำแนกเพศจากข้อความบนโซเชียลเน็ตเวิร์ก
ผู้วิจัย	เอกภพ พูลสวัสดิ์
ปริญญา	วิทยาศาสตรมหาบัณฑิต
ปีการศึกษา	2563
อาจารย์ที่ปรึกษา	ผู้ช่วยศาสตราจารย์ ดร. วีรยุทธ เจริญเรืองกิจ

ในปัจจุบันการเก็บข้อมูลจากช่องทางต่างๆ บนอินเตอร์เน็ต โดยเฉพาะโซเชียลเน็ตเวิร์กนั้นกำลังเป็นที่แพร่หลาย โดยเฉพาะอย่างยิ่งการเก็บข้อมูลเพื่อใช้สำหรับวิจัยทางด้านการตลาด ซึ่งจะทำให้สามารถเข้าใจผู้บริโภคได้มากยิ่งขึ้น บริษัทใดที่สามารถเก็บข้อมูลผู้บริโภคได้มากจะทำให้มีความสามารถในการแข่งขันในตลาดได้มากกว่าบริษัทที่เก็บข้อมูลได้น้อย ซึ่งข้อมูลผู้บริโภคนั้นก็มีหลากหลายปัจจัยที่ส่งผลให้เกิดพฤติกรรมการบริโภคที่แตกต่างกันในแต่ละบุคคล หนึ่งในนั้นคือเพศ ซึ่งเป็นปัจจัยที่สำคัญปัจจัยหนึ่งที่ส่งผลโดยตรงต่อพฤติกรรมของผู้บริโภค สำหรับการเก็บข้อมูลเพศของผู้บริโภคในไทยนั้น ส่วนใหญ่มักใช้วิธีการระบุเพศจากข้อความต่างๆ ที่ผู้บริโภคเผยแพร่โดยใช้วิธีการระบุจากคำลงท้าย เช่น ครับ ค่ะ หรือคำสรรพนามแทนตัว เช่น ผม ดิฉัน ในการระบุเพศ สำหรับข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์ก ผู้วิจัยพบว่ามีข้อความเพียง 30% เท่านั้นที่มีคำที่สามารถระบุเพศได้ ซึ่งถ้าหากสามารถระบุเพศจากข้อความในส่วนที่ไม่มีคำเหล่านี้อีก 70% ได้ จะทำให้สามารถนำข้อมูลที่ได้ไปใช้ได้อย่างถูกต้องและมีประสิทธิภาพ ทำให้เกิดความได้เปรียบในด้านการตลาด ในงานวิจัยนี้ได้นำเสนอวิธีการจำแนกเพศของผู้เขียนข้อความแสดงความคิดเห็นสำหรับข้อความภาษาไทยบนโซเชียลเน็ตเวิร์ก โดยการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติกับการสกัดคุณลักษณะ ร่วมกับการสร้างแบบจำลองการเรียนรู้ของเครื่อง ให้ค่าความแม่นยำในการจำแนกเพศ 79.04%

คำสำคัญ : การจำแนกเพศ, การประมวลผลภาษาธรรมชาติ, การสกัดคุณลักษณะ, การเรียนรู้ของเครื่อง, ข้อความแสดงความคิดเห็นในภาษาไทย

Title	GENDER CLASSIFICATION FROM TEXT DATA IN SOCIAL NETWORK
Author	EKKAPOB POONSAWAT
Degree	MASTER OF SCIENCE
Academic Year	2020
Thesis Advisor	Assistant Professor Werayuth Charoenruengkit , Ph.D.

Data collection from various channels on the internet, especially social networks, is becoming an increasingly common practice amongst businesses. Companies conducting any form of marketing research will find valuable insights and gain a deeper understanding of their consumers through collecting data. Any company able to collect and effectively analyze the most consumer data will therefore gain a significant competitive advantage. There are a variety of factors that affect consumer behavior. One of the most important factors is gender, which is a key determinant of consumer behavior. In the collection of gender data from social networks in Thailand, a popular method of distinguishing gender of different texts can be done through analysis of suffixes such as "Krub" for males or "Ka" for females, or from pronouns such as "Phom" for males or "Di-Chan" for females. From the research, it was found that only 30% of the studied texts contained gender-related suffixes or pronouns. It stands to reason that if the remaining 70% of texts that do not use contemporary gender-related suffixes or pronouns could be adequately analyzed, it presents an attractive opportunity to gain a competitive advantage in consumer-targeted marketing. This research proposes that the gender classification method of Thai comment texts, through the application of natural language processing techniques and supplemented with machine learning models, with an accuracy of 79.04%.

Keyword : Gender classification, Natural Language Processing, Feature extraction, Machine learning, Thai comment texts

กิตติกรรมประกาศ

สารนิพนธ์นี้สามารถสำเร็จลุล่วงไปได้ด้วยความช่วยเหลือ ให้คำแนะนำและให้ข้อคิดเห็นจากหลายๆท่าน ผู้วิจัยขอกราบขอบพระคุณ

ผู้ช่วยศาสตราจารย์ ดร.วีรยุทธ เจริญเรืองกิจ ที่ได้ให้ความกรุณาเป็นที่ปรึกษา และให้คำแนะนำที่เป็นประโยชน์ต่อการทำสารนิพนธ์นี้ด้วยความเอาใจใส่ตลอดมา

ผู้ช่วยศาสตราจารย์ ดร.นุวีย์ วิวัฒนวัฒนา สำหรับความช่วยเหลือด้านเอกสารที่รวดเร็ว อาจารย์ ดร.โสภณ มงคลลักษมี ที่กรุณาเป็นกรรมการสอบตั้งแต่การสอบเค้าโครงถึงการสอบปากเปล่า อาจารย์ ดร.ศิริสรรพ เหล่าหะเกียรติ สำหรับคำแนะนำในการแก้ไขบทความวิจัย

ขอกราบขอบพระคุณคณาจารย์และกรรมการบริหารหลักสูตรเทคโนโลยีสารสนเทศ คณะวิทยาศาสตร์ มหาวิทยาลัยศรีนครินทรวิโรฒทุกท่าน ที่ได้ประสิทธิ์ประสาทวิชาความรู้ต่างๆ ให้แก่ผู้วิจัย

ขอขอบคุณ เจ้าหน้าที่ที่หลักสูตรและบัณฑิตวิทยาลัย สำหรับการให้ความช่วยเหลืออำนวยความสะดวกตลอดมา

ขอขอบคุณเพื่อนๆ พี่ๆ น้องๆ หลักสูตรเทคโนโลยีสารสนเทศ ครอบครัว และรวมถึงบุคคลที่ไม่ได้กล่าวนามไว้ ณ ที่นี้ ที่ให้ความช่วยเหลือและเป็นกำลังใจตลอดมา

สุดท้ายนี้ ผู้วิจัยขอน้อมรำลึกถึงบุญคุณของบิดามารดา ครูอาจารย์ที่อบรมสั่งสอนและให้การสนับสนุนด้วยดีตลอดมา

เอกภพ พูลสวัสดิ์

สารบัญ

หน้า

สารบัญตาราง

หน้า

สารบัญรูปภาพ

หน้า

บทที่ 1
บทนำ

ความสำคัญและที่มาของงานวิจัย

ทุกวันนี้ข้อมูลมากมายมหาศาลได้เกิดขึ้นทุกๆ วินาที ยิ่งเทคโนโลยีไปไกลมากเท่าไร ยิ่งจะทำให้เกิดการสร้างข้อมูลมากขึ้นเป็นทวีคูณ โดยเฉพาะข้อมูลในโลกออนไลน์ ถ้าใครสามารถนำข้อมูลเหล่านี้มาใช้ให้เกิดประโยชน์ได้ ก็ยิ่งจะทำให้เกิดการพัฒนาในด้านต่างๆ มากยิ่งขึ้น ทางด้านการตลาดก็เช่นกัน ทุกวันนี้หลายๆ บริษัทหันมาทำการตลาดดิจิตอล (Digital marketing) กันมากขึ้น เนื่องจากสามารถเก็บข้อมูลผู้ใช้งานได้โดยไม่ต้องทำการสำรวจการตลาดแบบออฟไลน์ (Offline marketing survey) เหมือนสมัยก่อน อีกทั้งยังสามารถทดลองแผนการตลาด การโฆษณา และสามารถวัดประสิทธิภาพของแผนการตลาดหรือการโฆษณานั้นๆ ได้ง่ายมากยิ่งขึ้น ทำให้เกิดการทำการตลาดส่วนบุคคล (Personalized marketing) ซึ่งเป็นวิธีการที่นักการตลาดพยายามนำเสนอสินค้า และบริการให้ตรงกับความต้องการของกลุ่มผู้บริโภคให้ได้มากที่สุด โดยไม่จำเป็นต้องเสนอสินค้าและบริการแบบเดียวกันให้กับทุกคนเหมือนการตลาดแบบเก่า แต่มุ่งเน้นไปที่ความต้องการของกลุ่มผู้บริโภคที่เป็นเป้าหมายเท่านั้น ดังนั้นการเก็บข้อมูลของผู้บริโภคให้ได้มากที่สุดจึงเป็นจุดเริ่มต้นที่สำคัญสำหรับการวิเคราะห์กลุ่มผู้บริโภค เพื่อให้ได้กลุ่มผู้บริโภคที่เป็นเป้าหมายต่อไป

ปัญหาที่สำคัญของการทำการตลาดส่วนบุคคลคือ ข้อมูลที่ได้มานั้นไม่ได้มีข้อมูลที่บ่งบอกรายละเอียดของบุคคลนั้นๆ อย่างชัดเจน ตัวอย่างเช่น ข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์ก มักจะมีแต่ข้อความที่แสดงอารมณ์ความรู้สึก ข้อความที่แสดงคำถามหรือตอบคำถาม ทำให้ไม่สามารถสกัดเอาข้อมูลส่วนที่เป็นที่ต้องการได้ อาทิเช่น เพศ อายุ ภูมิลำเนา เชื้อชาติ ภาษาและอื่นๆ จึงไม่สามารถนำไปวิเคราะห์หากลุ่มเป้าหมายได้อย่างที่ต้องการได้ เป็นที่มาของงานวิจัยทางด้านการจำแนกข้อมูลส่วนบุคคลของผู้เขียนจากงานเขียนหรือข้อความที่ผู้เขียนสร้างขึ้น (Author's Profiling) สิ่งที่ท้าทายสำหรับงานด้านการจำแนกข้อมูลส่วนบุคคล คือ คำที่อยู่ในข้อความนั้น ไม่สามารถบอกได้แน่ชัดว่าเป็นเพศอะไร อายุเท่าไร ภาษาแม่ที่ใช้หรือเชื้อชาติอะไร ซึ่งแตกต่างจากงานด้านการจำแนกข้อความประเภทอื่นๆ เช่น การจำแนกอารมณ์ (Sentiment Analysis) ที่คำแต่ละคำจะมีอารมณ์ ความรู้สึกของคำอยู่ในตัวเอง หรือการจำแนกหมวดหมู่ของข้อความ (Text Categorization) ก็มีคำที่สามารถบอกได้ว่าข้อความนั้นๆ อยู่ในหมวดหมู่อะไร เป็นต้น สำหรับงานด้านการจำแนกข้อมูลส่วนบุคคลในภาษาไทยนั้นยังไม่เป็นที่แพร่หลาย โดยเฉพาะการจำแนกเพศ เนื่องจากการที่ภาษาไทยมีคำสรรพนามบุรุษที่ 1 ที่สามารถบอกเพศ

ของผู้เขียนได้ เช่น ผม กระผม ดิฉัน ฉัน คำลงท้ายประโยคที่สามารถบอกเพศของผู้เขียนได้ เช่น ครับ คะ ค่ะ หรือคำที่ตั้งใจกร่อนเสียง บิดเสียงทางภาษา หรือคำที่พิมพ์ผิด ทั้งที่ตั้งใจพิมพ์ผิด หรือไม่ตั้งใจพิมพ์ผิดที่สามารถบอกเพศของผู้เขียนได้ เช่น ครัช ครัส คร่ะ ฮับ ป่ม นู่ เป็นต้น แต่ก็มีผู้เขียนจำนวนไม่น้อยที่ไม่ได้ใช้คำเหล่านี้ โดยเฉพาะอย่างยิ่งในโซเชียลเน็ตเวิร์ก เช่น เฟสบุ๊ค ทวิตเตอร์ อินสตาร์แกรม หรือเว็บบอร์ดต่างๆ ซึ่งมักใช้คำที่ไม่เป็นทางการ ซึ่งผู้วิจัยพบว่ามีข้อความเพียง 30% เท่านั้น ที่สามารถระบุเพศของผู้เขียนจากคำเหล่านี้ได้

จากที่กล่าวมาข้างต้นจะเห็นได้ว่าการจำแนกข้อมูลส่วนบุคคลนั้น ถ้าเราสามารถดึงข้อมูลที่จำเป็นออกมาได้ จะทำให้สามารถนำข้อมูลนั้นไปใช้ให้เกิดประโยชน์ ไม่เพียงแต่ด้านการตลาดเท่านั้น แต่สามารถนำไปใช้ให้เกิดประโยชน์ในด้านต่างๆ อีกมากมาย เพราะเราจะสามารถนำข้อมูลเหล่านี้ไปวิเคราะห์พฤติกรรมของผู้คนในสังคมต่อได้อีก ทำให้เกิดการพัฒนาในด้านต่างๆ มากยิ่งขึ้นไป สำหรับในงานวิจัยนี้ ผู้วิจัยต้องการนำเสนอวิธีการจำแนกเพศของผู้เขียนข้อความแสดงความคิดเห็นในภาษาไทย โดยใช้การสกัดคุณลักษณะ (Features Extraction) จากข้อความแสดงความคิดเห็นบนเว็บไซต์พันทิพ ร่วมกับการสร้างแบบจำลองสำหรับการจำแนก โดยใช้เทคนิคการเรียนรู้ของเครื่อง (Machine Learning techniques) และเทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing techniques)

วัตถุประสงค์ของงานวิจัย

1. เพื่อศึกษาและประยุกต์ใช้เทคนิคการเรียนรู้ของเครื่องร่วมกับเทคนิคการประมวลผลภาษาธรรมชาติ สำหรับการสร้างแบบจำลองการจำแนกเพศของผู้เขียนข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์ก

2. เพื่อศึกษาและประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ สำหรับการสกัดคุณลักษณะจากข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์ก

ขอบเขตของการวิจัย

1. ข้อมูลที่ใช้สำหรับการวิจัย เป็นข้อความแสดงความคิดเห็นในภาษาไทยจากเว็บไซต์พันทิพ ไม่น้อยกว่า 50,000 ข้อความแสดงความคิดเห็นสำหรับเพศชาย และ 50,000 ข้อความแสดงความคิดเห็นสำหรับเพศหญิง รวมทั้งสิ้นไม่น้อยกว่า 100,000 ข้อความแสดงความคิดเห็น

2. เทคนิคการเรียนรู้ของเครื่องที่ใช้สำหรับการสร้างแบบจำลองการจำแนก 3 เทคนิค ได้แก่ Logistic Regression, Naïve Bayes และ Random Forest

3. ใช้วิธีการวัดประสิทธิภาพด้วยตัววัดประสิทธิภาพ 4 ค่า ได้แก่ Accuracy, Precision, Recall และ F1 score

วิธีการดำเนินงานวิจัย

1. ทบทวนวรรณกรรมและงานวิจัย (Literature Review) ที่เกี่ยวข้อง
2. ศึกษาเครื่องมือที่และอัลกอริทึมจะนำมาใช้ในงานวิจัย
3. เก็บข้อมูลที่ใช้ในงานวิจัย
4. ทดลองวิธีการวิจัยตามที่ได้ศึกษามาจากงานวิจัยที่เกี่ยวข้อง
5. ศึกษา และทดลองเพิ่มเติม เพื่อพัฒนาประสิทธิภาพของงานวิจัย
6. วิเคราะห์ ประเมินผลและสรุปผลการวิจัย

สมมติฐานในการวิจัย

1. เพศหญิงและเพศชายมีการใช้คำศัพท์ในข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์กที่แตกต่างกัน จึงทำให้สามารถใช้เทคนิคการเรียนรู้ของเครื่อง และเทคนิคของการประมวลผลภาษาธรรมชาติในจำแนกเพศของผู้เขียนข้อความแสดงความคิดเห็นได้

2. เพศหญิงและเพศชายมีการใช้รูปแบบของคำในข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์กที่แตกต่างกัน จึงทำให้สามารถใช้เทคนิคการเรียนรู้ของเครื่อง และเทคนิคของการประมวลผลภาษาธรรมชาติในจำแนกเพศของผู้เขียนข้อความแสดงความคิดเห็นได้

3. เพศหญิงและเพศชายมีการใช้ชนิดของคำในข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์กที่แตกต่างกัน จึงทำให้สามารถใช้เทคนิคการเรียนรู้ของเครื่อง และเทคนิคของการประมวลผลภาษาธรรมชาติในจำแนกเพศของผู้เขียนข้อความแสดงความคิดเห็นได้

ประโยชน์ที่คาดว่าจะได้รับจากการวิจัย

1. สามารถนำแบบจำลองที่ได้จากงานวิจัยครั้งนี้ ไปใช้ในการจำแนกเพศกับข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์ก เพื่อนำข้อมูลเพศของผู้เขียนข้อความแสดงความคิดเห็นที่ได้ไปใช้ให้เกิดประโยชน์ทางการวิจัยและพัฒนาต่อไป เช่น การวิจัยทางด้านการตลาด การวิจัยพฤติกรรมของแต่ละเพศ เป็นต้น

2. สามารถนำคุณลักษณะที่สกัดจากข้อความแสดงความคิดเห็นและเทคนิคที่ใช้จากงานวิจัยครั้งนี้ ไปประยุกต์ใช้กับการจำแนกข้อความ หรืองานทางด้านการประมวลผลภาษาธรรมชาติอื่นๆ ต่อไป

บทที่ 2
เอกสารและงานวิจัยที่เกี่ยวข้อง

ในการวิจัยครั้งนี้ ผู้วิจัยได้ศึกษาเอกสารและงานวิจัยที่เกี่ยวข้อง และได้นำเสนอตามหัวข้อต่อไปนี้

1. การประมวลผลภาษาธรรมชาติ
2. ความแตกต่างของการใช้ภาษาระหว่างเพศหญิงและชาย
3. เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกที่ใช้ในงานวิจัย
4. วิธีการวัดประสิทธิภาพของการทดลองที่ใช้ประเมินผลการวิจัย
5. งานวิจัยที่เกี่ยวข้อง

การประมวลผลภาษาธรรมชาติ

การประมวลผลภาษาธรรมชาติ (Natural Language Processing: NLP) หมายถึง กระบวนการวิเคราะห์และประมวลผลทางด้านภาษา โดยการทำให้คอมพิวเตอร์เข้าใจภาษามนุษย์และสามารถนำไปประมวลผลต่อได้ การศึกษาการประมวลผลภาษาธรรมชาตินั้น สามารถแบ่งออกได้เป็น 7 ระดับ ดังนี้ (Jurafsky & Martin, 2000; Liddy, 2001)

1. การศึกษาในทางสัทศาสตร์ (Phonology) คือ การศึกษาและวิเคราะห์เสียงของมนุษย์ให้ออกมาเป็นคำ

2. การศึกษาในระดับวจีวิภาค (Morphology) คือ การศึกษาและวิเคราะห์หน่วยคำที่เล็กที่สุดที่มีความหมาย (Morpheme)

3. การศึกษาในระดับหน่วยศัพท์ (Lexical) คือ การศึกษาและวิเคราะห์ความหมายของคำโดด (the meaning of individual words) และชนิดของคำ (part of speech)

4. การศึกษาในระดับวากยสัมพันธ์ (Syntactic) คือ การศึกษาและวิเคราะห์โครงสร้างของประโยคและกฎไวยากรณ์ (grammatical structure)

5. การศึกษาในระดับอรรถศาสตร์ (Semantic) คือ การศึกษาและวิเคราะห์บริบทของคำและความหมายของคำในประโยค (the meanings of a sentence)

6. การศึกษาในระดับปริจเฉท (Discourse) คือ การศึกษาและวิเคราะห์ความเชื่อมโยงและขยายกันของหน่วยคำประโยค

7. การศึกษาในระดับวัจนปฏิบัติ (Pragmatics) คือ การตีความข้อความให้เห็นความหมายระดับลึกตามบริบท

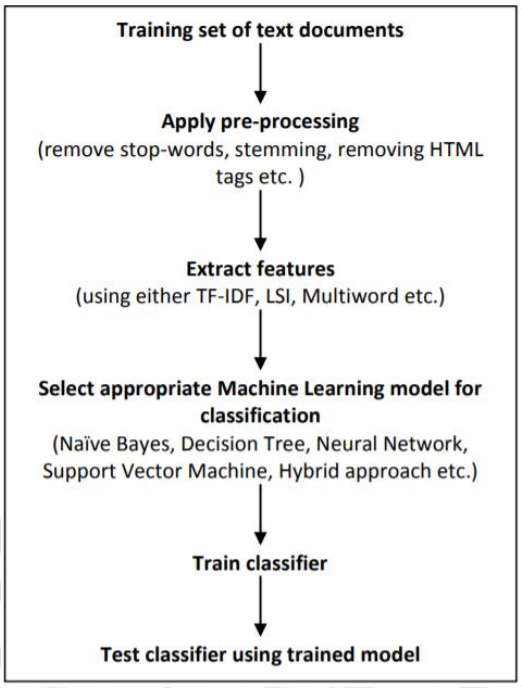
ตัวอย่างการใช้งานระบบประมวลผลภาษาธรรมชาติ (Natural Language Processing Applications)

1. การจำแนกข้อความ (Text classification)
2. การสรุปใจความสำคัญ (Text summarization)
3. การแปลภาษา (Machine translation)
4. การค้นคืนสารสนเทศ (Information retrieval)
5. การถามตอบ (Question answering)
6. การจดจำเสียงพูดอัตโนมัติ (Automatic speech recognition)
7. การแปลงข้อความให้เป็นเสียงพูด (Text-to-speech)
8. การจดจำตัวอักขระ (Optical character recognition)
9. การทำเหมืองข้อความ (Text mining)

เทคนิคการประมวลผลภาษาธรรมชาติ

สำหรับงานวิจัยนี้จะเป็นการนำเทคนิคการประมวลผลภาษาธรรมชาติมาใช้ในการจำแนกข้อความ สำหรับขั้นตอนพื้นทางของงานวิจัยทางด้านนี้ประกอบด้วย (Dalal & Zaveri, 2011; จิระวิชิตชัย, 2010) ดังภาพประกอบ 1 ซึ่งมีรายละเอียดดังนี้

1. ขั้นตอนการจัดเตรียมข้อมูล (pre-processing) เช่น การตัดคำ (word segmentation) การกำจัดคำหยุด (stop-word list removal) การหารากศัพท์ (stemming)
2. ขั้นตอนการสกัดคุณลักษณะ เป็นการเปลี่ยนข้อความให้อยู่ในรูปของเวกเตอร์
3. ขั้นตอนการเลือกแบบจำลองการเรียนรู้ของเครื่อง (machine learning models selection) สำหรับการจำแนกข้อความ
4. ขั้นตอนการฝึกสอนและประเมินประสิทธิภาพของแบบจำลอง (model training and evaluation)
5. ขั้นตอนการนำแบบจำลองไปประเมินผลกับชุดข้อมูลทดสอบ (model testing)



ภาพประกอบ 1 แสดงขั้นตอนการจำแนกข้อความโดยใช้เทคนิคทางด้านการประมวลผลภาษาธรรมชาติและเทคนิคการเรียนรู้ของเครื่อง

ที่มา : (Dalal & Zaveri, 2011)

อนึ่ง คำว่าภาษาศาสตร์คอมพิวเตอร์ หรือภาษาศาสตร์เชิงคำนวณ (Computational Linguistics) เป็นอีกหนึ่งคำที่อาจจะมีการเรียกสลับกันกับการประมวลผลภาษาธรรมชาติ ซึ่งแท้จริงแล้วคำทั้งสองคำนี้มีความหมายแตกต่างกันในด้านวัตถุประสงค์ของการศึกษา ภาษาศาสตร์คอมพิวเตอร์นั้นจะเป็นการศึกษาทางด้านภาษาศาสตร์โดยใช้คอมพิวเตอร์มาช่วยในการศึกษา จะเน้นการหาคำตอบทางด้านภาษาศาสตร์ ผ่านคลังข้อมูลและโมเดลทางคณิตศาสตร์ ส่วนการประมวลผลภาษาธรรมชาตินั้นจะเน้นไปที่การทำให้คอมพิวเตอร์เข้าใจภาษาของมนุษย์และทำหน้าที่ทางภาษาแทนมนุษย์ (Tsujii, 2011)

ความแตกต่างของการใช้ภาษาระหว่างเพศหญิงและชาย

เพศคือปัจจัยหนึ่งที่ทำให้เกิดความแตกต่างของการใช้ภาษา อันเนื่องมาจากการที่ผู้ใช้ภาษาต้องการแสดงความเป็นหญิงหรือชายของตนเอง หรืออาจเพื่อให้เกิดการตอบโต้ของผู้ที่มีปฏิสัมพันธ์ด้วยตามเพศของตนเอง เช่น เพศหญิงอาจต้องการให้คู่สนทนาใช้ภาษาอย่างสุภาพอ่อนโยน เนื่องจากตนเองเป็นเพศหญิง เป็นต้น

เฟ ครอสบี และลินดา ไนควิสต์ (Crosby & Nyquist, 1977) ได้ทำการศึกษาการใช้ภาษาของเพศหญิงตามสมมติฐานของเลคอฟที่ว่าเพศชายและเพศหญิงมีการใช้ภาษาที่แตกต่างกันเพราะเหตุผลทางสังคม พบว่าเพศหญิงมีการใช้ภาษาที่แตกต่างจากเพศชายอย่างมีนัยยะสำคัญในกรณีที่มีการใช้ภาษาอย่างเป็นทางการและมีความยาวของการใช้ภาษามากพอ กล่าวคือความแตกต่างเรื่องเพศในการใช้ภาษานั้นขึ้นอยู่กับบริบทของการใช้ภาษานั้นๆ ด้วย

สำหรับภาษาไทยนั้นความแตกต่างทางภาษาของเพศชายและเพศหญิงที่เห็นได้อย่างชัดเจนก็คือ สรรพนามที่ใช้เฉพาะเพศ เช่น ผม กระผม สำหรับเพศชาย และฉัน ดิฉัน หนูสำหรับเพศหญิง หรือคำลงท้าย ครับ สำหรับเพศชาย คะ ค่ะ สำหรับเพศหญิง เป็นต้น (ประสิทธิ์รัฐสินธุ์, 2545)

สำหรับการศึกษาความแตกต่างของการใช้ภาษาในเพศหญิงและเพศชายในภาษาไทยนั้น ได้มีผู้ทำการศึกษาไว้หลายท่าน อาทิเช่น มณทิรา ตาเมือง (ตาเมือง, 2555) ได้ทำการศึกษาการใช้ภาษาของนิสิตเพศหญิงและเพศชาย จากการศึกษาได้พบรูปแบบการใช้ภาษาทั้งสิ้น 10 รูปแบบ โดยเป็นรูปแบบของเพศชาย 3 รูปแบบและเพศหญิง 7 รูปแบบ ซึ่งแสดงให้เห็นถึงความแตกต่างของการใช้ภาษาของเพศหญิงและเพศชาย

นิตย์รดี ตาดทอง (ตาดทอง, 2558) ได้ศึกษาความแตกต่างของการใช้ภาษาระหว่างเพศชายและเพศหญิงบนเฟซบุ๊ค พบว่าโดยทั่วไปการใช้ภาษาของทั้งเพศชายและเพศหญิงมีความคล้ายคลึงกัน และมีแนวโน้มที่จะมีการใช้ภาษาแบบเสมอภาคกัน ส่วนที่มีความแตกต่างอย่างเห็นได้ชัดระหว่างการใช้ภาษาของเพศหญิงและเพศชาย คือ การใช้คำแสดงอารมณ์ความรู้สึก โดยเพศหญิงมักมีการใช้คำที่แสดงความเป็นกันเอง ส่วนเพศชายมักใช้คำที่แสดงความรุนแรง

โชติกา เศรษฐธัญการ (เศรษฐธัญการ, 2562) ได้ศึกษาการใช้ถ้อยคำเพื่อการอธิบายสิ่งต่างๆ โดยพบว่าเพศชายสามารถอธิบายสิ่งที่เป็นรูปธรรมได้มากกว่าเพศหญิง ในขณะที่เพศหญิงสามารถที่จะอธิบายสิ่งที่เป็นนามธรรมได้มากกว่าเพศชาย

ส่วน วรวรรณ เฟื่องขจรศักดิ์ (เฟื่องขจรศักดิ์, 2558) ได้มีการศึกษาการใช้ถ้อยคำแสดงความลังเล (Hedging Words) นั้น พบว่าเพศหญิงมีการใช้ถ้อยคำแสดงความมั่นใจมากกว่าเพศชาย ในขณะที่เพศชายจะใช้ถ้อยคำแสดงความไม่มั่นใจมากกว่าเพศหญิง

เทคนิคการเรียนรู้ของเครื่องสำหรับการจำแนกที่ใช้ในงานวิจัย

การจำแนก (Classification) เป็นหนึ่งในเทคนิคการเรียนรู้ของเครื่อง (Machine learning) ประเภทการเรียนรู้แบบมีผู้สอน (Supervised learning) ซึ่งจำเป็นต้องมีคำตอบสำหรับให้อัลกอริทึมได้เรียนรู้ก่อนที่จะนำแบบจำลองที่ได้ไปใช้ในการจำแนกต่อไป สำหรับในงานวิจัยนี้ได้

เลือกใช้อัลกอริทึมที่เป็นที่นิยมใช้สำหรับการจำแนก 3 อัลกอริทึม ได้ Logistic regression, Naive Bayes และ Random Forest (Müller & Guido, 2016)

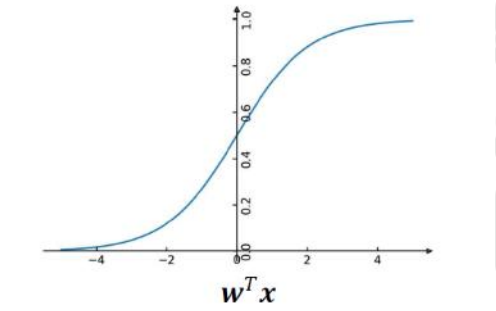
Logistic regression

สำหรับ Logistic regression นั้น เป็นเทคนิคสำหรับการจำแนกประเภท ซึ่งผลลัพธ์ของเทคนิคนี้อยู่ในรูปของความน่าจะเป็น การทำงานของอัลกอริทึมนี้จะแบ่งเป็น 2 ขั้นตอนคือ ทำการคำนวณ Linear combination ของ input กับ parameter ของ model จากนั้นนำผลลัพธ์จากขั้นแรกมาคำนวณ Sigmoid function (**σ**) ดังสมการ (1) ได้ออกมาเป็นความน่าจะเป็นที่จะเป็นมีค่าระหว่าง 0 ถึง 1

$$p(y = k|x) = \frac{1}{1+e^{-kw^T x}} \quad \text{เมื่อ } k \in \{-1, +1\} \tag{1}$$

โดยที่

$p(y = k|x)$ คือค่าความน่าจะที่ y จะเป็นคลาส k เมื่อมีตัวแปร x



ภาพประกอบ 2 กราฟแสดงผลลัพธ์ที่ได้จาก Sigmoid function

Naive Bayes

เทคนิค Naive Bayes นั้นจะใช้หลักการคำนวณความน่าจะเป็น โดยใช้สิ่งที่เรียกว่าทฤษฎีของเบย์ (Bayes theorem) ดังสมการ (2)

$$p(y|x) = \frac{p(x|y)p(y)}{p(x)} \tag{2}$$

โดยที่

$p(y|x)$ คือค่าความน่าจะเป็นที่ข้อมูลที่มีตัวแปรเป็น x จะมีคลาส y

$p(x|y)$ คือค่าความน่าจะเป็นที่ข้อมูลที่มีคลาส y และมีตัวแปร x

โดยที่ $x = x_1 \cap x_2 \cap \ldots \cap x_n$ โดยที่ n คือจำนวนตัวแปร

$p(y)$ คือค่าความน่าจะเป็นของคลาส y

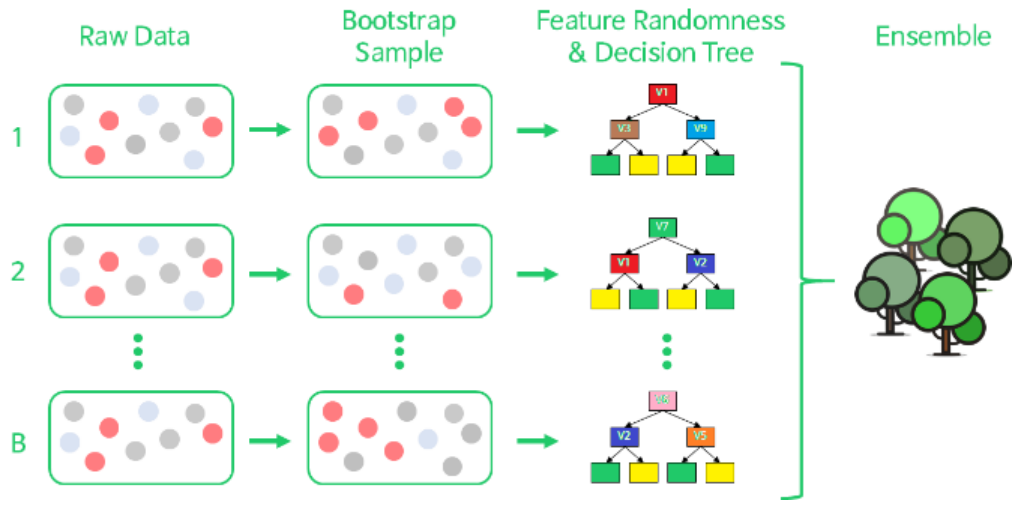
$p(x)$ คือค่าความน่าจะเป็นของตัวแปร x

แต่จากการที่ตัวแปร $x = x_1 \cap x_2 \cap \ldots \cap x_n$ จะเกิดขึ้นร่วมกันนั้น อาจจะมีน้อยหรือแทบไม่มีเลย ดังนั้นจึงใช้สมมติฐานที่ว่าตัวแปรแต่ละตัวไม่ขึ้นต่อกัน ทำให้สามารถเขียนสมการ $p(x|y)$ ใหม่ได้ ดังสมการ (3) ซึ่งเป็นสมการที่ใช้สำหรับเทคนิค Naive Bayes นี้

$$p(x|y) = p(x_1|y) \times p(x_2|y) \times \ldots \times p(x_n|y) \quad (3)$$

Random Forest

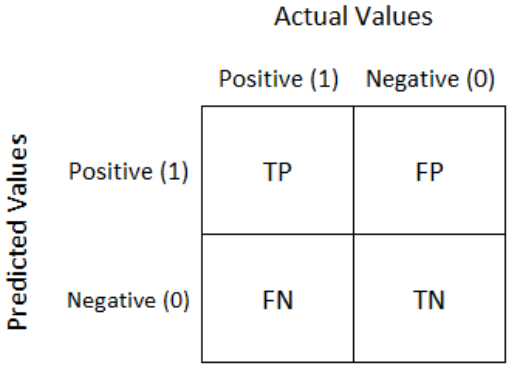
เป็นเทคนิคที่พัฒนาขึ้นจากอัลกอริทึม Decision Tree โดยเป็นการนำแบบจำลอง Decision Tree หลายๆ แบบจำลองมาใช้ทำนายร่วมกัน ซึ่งแต่ละแบบจำลองนั้นจะใช้ข้อมูลและคุณลักษณะที่นำมาสร้างแบบจำลองไม่เหมือนกัน หลังจากนั้นจึงนำผลการทำนายที่ได้จากแบบจำลองมาทำการลงคะแนน (Voting) เพื่อให้ได้ผลลัพธ์สุดท้าย ดังภาพประกอบ 3



ภาพประกอบ 3 การทำงานของอัลกอริทึม Random Forest

วิธีการวัดประสิทธิภาพของการทดลองที่ใช้ประเมินผลการวิจัย

Confusion Matrix เป็นอีกหนึ่งวิธีที่ดีในการวัดประสิทธิภาพแบบจำลองการทำนายสำหรับปัญหาการจำแนก โดยมีแนวคิดคือการนับจำนวนครั้งสำหรับการจำแนกประเภทถูกและการจำแนกประเภทผิดของแต่ในละประเภท (Class) ดังภาพประกอบ 4 ซึ่ง Confusion Matrix นั้นจะให้ข้อมูลที่หลากหลายไม่เพียงแค่ค่าความถูกต้อง (Accuracy) แต่ยังมีค่าอื่นๆ อีก โดยในงานวิจัยนี้นอกจากค่าความถูกต้องแล้ว จะใช้อีก 3 ค่าในการวัดประสิทธิภาพของแบบจำลองได้แก่ Precision, Recall และ F1Score



ภาพประกอบ 4 Confusion Matrix ขนาด 2x2

Accuracy

คือ การวัดความถูกต้องของโมเดล โดยใช้การเทียบผลลัพธ์การทำนายกับค่าจริงว่ามีความถูกต้องเท่าไร ดังสมการ (4)

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \quad (4)$$

โดยที่

True Positive (TP) คือ สิ่งที่โปรแกรมทำนายว่า “จริง” และ มีค่าเป็น “จริง”

True Negative (TN) คือ สิ่งที่โปรแกรมทำนายว่า “ไม่จริง” และ มีค่า “ไม่จริง”

False Positive (FP) คือ สิ่งที่โปรแกรมทำนายว่า “จริง” แต่ มีค่าเป็น “ไม่จริง”

False Negative (FN) คือ สิ่งที่โปรแกรมทำนายว่า “ไม่จริง” แต่ มีค่าเป็น “จริง”

Precision

คือ การวัดความถูกต้องของโมเดล โดยบอกว่าที่โมเดลทำนายว่าจริง ถูกต้องเท่าไร ดังสมการ (5)

$$Precision = \frac{TP}{(TP+FP)} \quad (5)$$

Recall

คือ การวัดความถูกต้องของโมเดล โดยบอกว่าที่โมเดลทำนายว่าจริง เป็นอัตราส่วนเท่าไรของจริงทั้งหมด ดังสมการ (6)

$$Recall = \frac{TP}{(TP+FN)} \quad (6)$$

F1 score

คือ ค่าเฉลี่ยแบบฮาโมนิคระหว่าง precision และ recall ดังสมการ (7)

$$F1 = 2 \times \frac{Precision \times Recall}{(Precision+Recall)} \quad (7)$$

งานวิจัยที่เกี่ยวข้อง

การทบทวนวรรณกรรมของงานวิจัยนี้ ได้ทำการศึกษางานวิจัยที่เกี่ยวข้องกับการจำแนกข้อมูลส่วนบุคคลของผู้เขียนจากงานเขียนหรือข้อความ บนเว็บบล็อก (Web blog) และโซเชียลเน็ตเวิร์กในภาษาต่างๆ และงานวิจัยที่เกี่ยวข้องกับการพัฒนาเทคนิคการประมวลผลภาษาธรรมชาติสำหรับภาษาไทย

1. บทความวิจัยเรื่อง Improving Gender Classification of Blog Authors (Mukherjee & Liu, 2010) ซึ่งตีพิมพ์ออกมาตั้งแต่ปี 2010 และยังคงเป็น state-of-the-art จนถึงทุกวันนี้ คณะผู้วิจัยได้ทำการพัฒนาวิธีการจำแนกเพศของผู้เขียนบทความภาษาอังกฤษบนเว็บบล็อกจำนวน 3,100 เว็บบล็อก แบ่งเป็นเว็บบล็อกของเพศชายจำนวน 1,588 เว็บบล็อก และเว็บบล็อกของเพศหญิงจำนวน 1,512 เว็บบล็อก โดยได้นำเสนอเทคนิคใหม่ 2 เทคนิค ได้แก่ เทคนิคการสกัดคุณลักษณะรูปแบบของชนิดของคำ (part-of-speech pattern features) และเทคนิคการ

เลือกคุณลักษณะแบบองค์รวม (ensemble of feature selection) ร่วมกับคุณลักษณะอื่นๆ อีก 4 แบบ คือ ค่าความถี่ชนิดของคำ (Frequency Measure) รูปแบบของคำ (Stylistic Features) รูปแบบของการใช้คำเฉพาะของแต่ละเพศ (Gender Preferential Features) และการวิเคราะห์ปัจจัยและชนิดของคำ (Factor Analysis and Word Classes) ส่วนการแปลงข้อความให้เป็นเวคเตอร์นั้นใช้วิธี TF (Term Frequency) จากนั้นนำไปสร้างแบบจำลองการจำแนกเพศด้วยอัลกอริทึม Naïve Bayes, SVM classification และ SVM regression ได้ค่าความถูกต้อง 73.57%, 86.24%, 88.56% ตามลำดับ

2. บทความวิจัยเรื่อง Gender Classification with Deep Learning (Bartle & Zheng, 2015) ได้นำเสนอการใช้อัลกอริทึม Windowed Recurrent Convolution Neural Network (WRCNN) สำหรับการจำแนกเพศของผู้เขียนบทความภาษาอังกฤษบนเว็บบล็อก และหนังสือช่วงศตวรรษที่ 19 และ 20 ที่เป็นภาษาอังกฤษ โดยทำการเปรียบเทียบประสิทธิภาพการจำแนกเพศของผู้เขียน กับแบบจำลองอื่นอีก 5 แบบ ได้แก่ Bag of Words with SVM, Average Embedding with single hidden layer network, paragraph2vec with SVM ซึ่งเป็น start-of-the-art ของ document classification ในขณะนั้น และเปรียบเทียบกับ POS features with SVM ของ Mukherjee และ Liu, 2010 โดยมีค่าความถูกต้องของ WRCNN สูงถึง 86% เป็นรองเพียงแค่ POS features with SVM เท่านั้น

3. บทความวิจัยเรื่อง Twitter Author Profiling Using Word Embeddings and Logistic Regression (Akhtyamova, Cardiff, & Ignatov, 2017) นำเสนอการใช้ pre-train model ที่สร้างจากอัลกอริทึม word2vec โดย pre-train model มีการฝึกฝนจากข้อความในทวิตเตอร์ เพื่อใช้สำหรับการแปลงข้อความเป็นเวคเตอร์ ซึ่งได้มีการทดลองการสร้างแบบจำลองสำหรับการจำแนกเพศของผู้เขียนบทความ โดยมีการทดลองหลากหลายอัลกอริทึม เช่น Random Forest, Linear Regression, Naive Bayes, SVMs และอื่นๆ กับข้อความบนทวิตเตอร์ 4 ภาษา ได้แก่ ภาษาอังกฤษ ภาษาสเปน ภาษาโปรตุเกส และภาษาอารบิก พบว่าแบบจำลองที่ให้ค่าแม่นยำมากที่สุดคือแบบจำลองที่สร้างจากอัลกอริทึม Logistic Regression โดยมีความถูกต้องสูงที่สุดในข้อความภาษาอังกฤษอยู่ที่ 74.46%

4. บทความวิจัยเรื่อง Using TF-IDF n-gram and Word Embedding Cluster Ensembles for Author Profiling (Poulston, Waseem, & Stevenson, 2017) ได้นำเสนอวิธีการสร้างแบบจำลองการจำแนกเพศของผู้เขียนจากข้อความบนทวิตเตอร์ 4 ภาษา ได้แก่ ภาษาอังกฤษ ภาษาสเปน ภาษาโปรตุเกส และภาษาอารบิก โดยได้มีการนำเสนอเอาไว้ 2 วิธีคือ

การใช้ TF-IDF n–grams กับอัลกอริทึม Logistic regression และอีกวิธีคือ การใช้ word embedding clusters กับ Gaussian process classifier พบว่าทั้ง 2 วิธีนี้ได้ค่าความถูกต้องของแบบจำลองการจำแนกเพศมากที่สุดอยู่ที่ภาษาโปรตุเกส โดยมีค่าความถูกต้องที่ 82.6% และ 83.9% ตามลำดับ

5. บทความวิจัยเรื่อง Word Unigram Weighing for Author Profiling (Veenhoven, Snijders, Hall, & Noord, 2018) ได้นำเสนอวิธีการสร้างแบบจำลองการจำแนกเพศของผู้เขียนจากข้อความบนทวิตเตอร์ 3 ภาษา ได้แก่ ภาษาอังกฤษ ภาษาสเปน และภาษาอารบิก โดยการใช้ 3 คุณลักษณะหลัก ได้แก่ Word Features, Character n-gram features และ Emoji Features และแปลงข้อความให้เป็นเวคเตอร์ด้วยวิธี TF-IDF weighting และใช้อัลกอริทึม Logistic Regression ในการสร้างแบบจำลอง พบว่าค่าความถูกต้องที่มีค่ามากที่สุดไม่ใช่การแปลงเวคเตอร์ด้วยวิธี TF-IDF weighting แต่เป็น TF-IDF แบบปกติ โดยมีความถูกต้องสูงสุด 77.8% ในภาษาอังกฤษ

6. บทความวิจัยเรื่อง Through a Gender Lens: Learning Usage Patterns of Emojis from Large-Scale Android Users (Chen et al., 2018) ได้มีการนำเสนอการใช้ emoji มาใช้ในการสร้างคุณลักษณะสำหรับการสร้างแบบจำลองการจำแนกเพศของผู้เขียน โดยใช้อัลกอริทึม 3 ชนิด ได้แก่ Random Forest Classifier, the Gradient Boosting Classifier และ SVM Classifier กับข้อมูลจำนวน 39,372 ผู้ใช้งาน ซึ่งผู้ใช้งานทั้งหมดนี้เลือกมาจากผู้ใช้งานที่มีการพิมพ์ข้อความที่มี emoji อยู่ด้วยไม่น้อยกว่า 100 ข้อความในแต่ละผู้ใช้งาน ซึ่งข้อความที่ใช้มาจาก 84 ภาษาทั่วโลก รวมถึงภาษาไทยด้วย โดยได้ค่าความถูกต้องสูงสุดอยู่ที่ภาษาโปรตุเกส 84.1% ส่วนภาษาไทยนั้นอยู่ที่ 80.8% (ดูตัวอย่าง 10 อันดับ emoji โดยเรียงลำดับจากค่า mutual information ได้จากตาราง 1)

ตาราง 1 ตัวอย่าง 10 อันดับ emoji โดยเรียงลำดับจากค่า mutual information

Rank	MI	Emoji e	p(Male|e)	p(Female|e)
1	0.0223	🎀	0.126	0.874
2	0.016	🎂	0.236	0.764
3	0.0145	🎉	0.275	0.725
4	0.0139	💄	0.232	0.768
5	0.0139	💋	0.267	0.733
6	0.012	🎈	0.225	0.775
7	0.0111	🌸	0.187	0.813
8	0.0104	💕	0.31	0.69
9	0.0096	💜	0.292	0.708
10	0.0094	💅	0.203	0.797

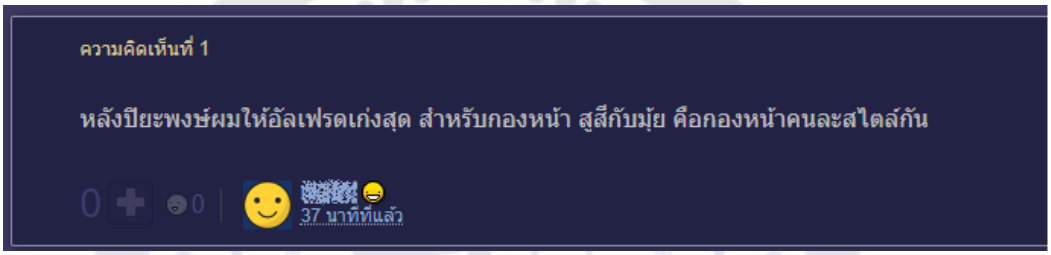
7. บทความวิจัยเรื่อง Gender Demography Classification on Instagram based on User's Comments Section (Reynaldo, Goenawan, Chanrico, Suhartono, & Purnomo, 2019) ได้มีการนำเสนอวิธีการสร้างแบบจำลองการจำแนกเพศของผู้เขียนจากข้อความแสดงความคิดเห็นบนอินสตาร์แกรมในภาษาอังกฤษ ซึ่งคณะผู้วิจัยได้ทำการกำกับเพศของของความจากรูปโปรไฟล์ของผู้ใช้งานโดยที่ผู้ใช้งานนั้นจะต้องมีรูปของตัวเองอย่างน้อย 10 รูป จากนั้นนำข้อความแสดงความคิดเห็นไปแปลงให้เป็นเวคเตอร์ด้วยวิธี bag-of-words แล้วนำไปสร้างแบบจำลองโดยใช้อัลกอริทึม 4 ชนิด ได้แก่ XGBoost, Naive Bayes, AdaBoost และ SVM โดยมีการเปลี่ยนสัดส่วนตัวอย่างของเพศชายและเพศหญิง ในการสร้างแบบจำลอง พบว่าได้ค่าความถูกต้องมากที่สุดเมื่อใช้อัลกอริทึม Naive Bayes กับข้อมูลที่มีสัดส่วนเพศจากต่อเพศหญิงเป็น 1:2 ได้ค่าความถูกต้อง 78.6%

8. บทความวิจัยเรื่อง A Comparative Study of Pre-trained Language Models on Thai Social Text Categorization (Horsuwan, Kanwatchara, Vateekul, & Kijsirikul, 2019) ได้ทำการพัฒนาแบบจำลองก่อนการฝึกฝน (pre-train modal) สำหรับภาษาไทย โดยได้ใช้ข้อความแสดงความคิดเห็นจากเว็บไซต์พันทิพ ซึ่งมีการพัฒนาพจนานุกรมโดยการเพิ่มเติมคำศัพท์ที่เป็นคำศัพท์ที่ใช้กันในโซเชียลเน็ตเวิร์ก สำหรับการตัดในพจนานุกรมของไลบรารี่ pyThaiNLP ด้วย และวัดประสิทธิภาพการตัดคำกับวิธีปกติของไลบรารี่ pyThaiNLP พบว่าได้ค่าความถูกต้องสำหรับการตัดคำเพิ่มขึ้นประมาณ 3.4% จากนั้นทางคณะผู้วิจัยได้ทำการสร้างแบบจำลองทางภาษา (Language model) จาก state-of-art ทั้ง 4 แบบ คือ ULMFiT, ELMo with biLSTM, OpenAI GPT และ BERT โดยทำการเปรียบเทียบประสิทธิภาพของแต่ละแบบจำลองทางภาษา ด้วยการเปรียบเทียบผลการจำแนกข้อมูลจากการแข่งขันการจำแนกข้อมูลที่เป็นภาษาไทยบนเว็บไซต์ Kaggle สองการแข่งขัน คือ Wongnai Challenge: Rating Review Prediction และ Wisesight Sentiment Analysis พบว่าแบบจำลองทางภาษาที่สร้างขึ้นจากข้อมูลในเว็บไซต์พันทิพกับพจนานุกรมที่พัฒนาขึ้นมาให้ผลการจำแนกที่ถูกต้องมากขึ้นอยู่ที่ประมาณ 1-2% สำหรับการแข่งขัน Wongnai Challenge: Rating Review Prediction และ 4-6% ในการแข่งขัน Wisesight Sentiment Analysis โดยผู้เข้าแข่งกันส่วนใหญ่ของทั้งสองการแข่งขันนี้มักจะใช้แบบจำลองก่อนการฝึกฝนที่ฝึกฝนมาจาก Thai Wiki Dump และตัดคำโดยใช้ไลบรารี่ pyThaiNLP ซึ่งกำลังเป็นที่นิยมกันในงานทางด้านการประมวลผลภาษาธรรมชาติของภาษาไทย

บทที่ 3
วิธีดำเนินการวิจัย

การเก็บข้อมูล (Data Acquisition)

ข้อมูลที่ใช้ในงานวิจัยนี้ คือข้อความแสดงความคิดเห็นบนเว็บไซต์พันทิพ ดังภาพประกอบ 5 และทำการเก็บข้อมูลด้วยวิธี Scraping โดยใช้ python library ที่ชื่อว่า selenium และ BeautifulSoup ตั้งแต่เดือน มิ.ย. 2562 ถึง ก.ค. 2562 มีจำนวนข้อความแสดงความคิดเห็นทั้งหมดก่อนเข้าสู่ขั้นตอนการทำความสะอาดข้อมูล ทั้งสิ้น 854,472 ข้อความแสดงความคิดเห็น โดยมีรายละเอียดของข้อมูลดังตาราง 2



ภาพประกอบ 5 ตัวอย่างข้อความแสดงความคิดเห็นบนเว็บไซต์พันทิพ

ตาราง 2 แสดงรายละเอียดของข้อมูลที่เก็บมาจากเว็บไซต์พันทิพ

Column	Description
_id	Primary key ของตาราง
category	ห้องในพันทิพ (forum)
comment	ข้อความแสดงความคิดเห็น
comment_id	id ของความคิดเห็น
content_id	id ของกระทู้
post_date	วันที่โพส
tags	แท็ก

จากตาราง 2 แสดงให้เห็นถึงรายละเอียดของข้อมูลที่เก็บมาจากเว็บไซต์พันทิพ โดยมีชื่อของคอลัมภ์ต่างๆ ของข้อมูลที่เก็บมาและรายละเอียดของแต่ละคอลัมภ์ว่าคือข้อมูลอะไร ดังภาพประกอบ 6 แสดงตัวอย่างข้อมูลบนโปรแกรมภาษาไพธอน

	_id	category	comment	comment_id	content_id	post_date	tags
0	{'$oid': '5d5500cbb3d30e2fdcfb33af'}	pantip_home	ผมอยู่สมุทรสาคร กระทุ่มแบน...หันเสาไปทางใบหยก2 ไม่มีสัญญาณเรย แต่หันไปสำโรง สมุทรปราการ กลับมีสัญญาณผมใช้เสาก้างปลา 14E ที่รับสัญญาณได้ไกล100กิโลเมตร แต่จุดที่ผมอยู่ห่างจากใบหยก2แค่ 25-26 กิโลฯ ดันไม่มีสัญญาณ เป็นเพราะอะไรคัฟ...แล้วเวลาลมพัดแรงๆ หรือฝนตก ไม่มีสัญญาณเรย(หันไปทางสำโรง)แต่ถ้าอากาศปกติ มีสัญญาณ แต่ก็ไม่ครบ หรือ บางทีสัญญาณอ่อนมาก	79776327	32789937	{'$date': '2018-06-30T22:04:56.000Z'}	เครื่องใช้ไฟฟ้า,ความบันเทิงในบ้าน,ทีวีดิจิตอล

ภาพประกอบ 6 ตัวอย่างข้อมูลที่เก็บมาจากเว็บไซต์พันทิพบนโปรแกรมภาษาไพธอน

การระบุประเภทข้อมูล (Data Labeling)

เนื่องจากข้อมูลที่ได้มานั้น ยังไม่มีการระบุประเภทของข้อมูลว่าข้อความแสดงความคิดเห็นข้อความไหนผู้เขียนข้อความเป็นเพศหญิงหรือเพศชาย ดังนั้นจึงต้องมีการระบุประเภทของข้อมูลก่อน โดยการดูจากคำสรรพนามบุรุษที่ 1 และคำลงท้ายประโยค (ประสิทธิ์รัฐสินธุ์, 2545) ดังตาราง 3 แสดงคำที่ใช้สำหรับระบุเพศของผู้เขียนข้อความแสดงความคิดเห็น โดยมีทั้งคำที่สะกดถูกต้องตามพจนานุกรม และสะกดผิด หลังจากนั้นจึงเลือกเฉพาะข้อความแสดงความคิดเห็นที่สามารถระบุเพศของผู้เขียนข้อความแสดงความคิดเห็นมาใช้สำหรับงานวิจัย ทั้งสิ้น 247,910 ข้อความแสดงความคิดเห็น ซึ่งจะเห็นได้ว่ามีเพียงประมาณ 30% ของข้อมูลเท่านั้นที่สามารถระบุเพศได้

ตาราง 3 คำที่ใช้สำหรับระบุเพศของผู้เขียนข้อความแสดงความคิดเห็น

คำที่ใช้	**เพศของผู้เขียนข้อความ**	label
ผม, ผ ม	ชาย	male
ครับ, ครัช, ครัส, คับ, ค่าบ	ชาย	male
ดิฉัน, ฉัน, หนู	หญิง	female
คะ, ค่ะ, ค้ะ, คร่ะ, ค ะ	หญิง	female
นะคะ, นะค่ะ	หญิง	female
จ๊ะ, จ้า	หญิง	female

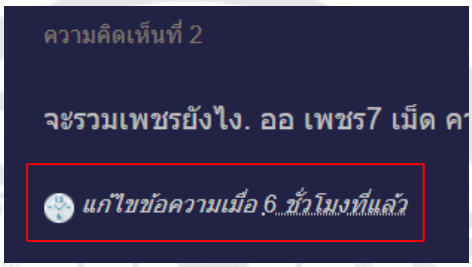
การทำความสะอาดข้อมูล (Data Cleaning)

ในขั้นตอนนี้จะเป็นการทำความสะอาดข้อมูล ซึ่งแบ่งออกได้เป็น 7 ขั้นตอนดังนี้

1. เลือกข้อมูลเฉพาะคอลัมภ์ที่ต้องการใช้ ได้แก่ comment

2. จัดการกับข้อความแสดงความคิดที่มีแต่การเคาะเว้นวรรคโดยไม่มีการพิมพ์ตัวหนังสือ ให้กลายเป็นข้อความว่าง

3. จัดการกับข้อความที่สร้างขึ้นโดยระบบของทางเว็บไซต์ ดังภาพประกอบ 7 แสดงข้อความที่ระบบสร้างขึ้นเมื่อมีการแก้ไขข้อความแสดงความคิดเห็น โดยการแทนที่ด้วยข้อความว่าง



ภาพประกอบ 7 ข้อความที่ระบบสร้างขึ้นเมื่อมีการแก้ไขข้อความแสดงความคิดเห็น

4. ลบข้อความแสดงความคิดเห็นที่ซ้ำกันทั้งความคิดเห็นออกไป จากสมมติฐานที่ว่าความคิดเห็นไม่ควรจะซ้ำกันทั้งความคิดเห็น ถ้ามีการซ้ำกันทั้งความคิดเห็นน่าจะเป็นข้อความโฆษณา หรือแอดมินของผลิตภัณฑ์ต่างๆ ที่คอยตอบคำถามอยู่ในเว็บไซต์พันทิพ หรือข้อความแสดงความคิดเห็นที่สั้น และเป็นคำสามัญซึ่งไม่สามารถบ่งบอกเพศได้เมื่อตัดคำลงท้ายออกไป เช่น ข้อความแสดงความคิดเห็นที่มีแต่คำว่า “ขอบคุณครับ” หรือ “ขอบคุณค่ะ” เมื่อตัดคำว่า “ครับ” และ “ค่ะ” ออก (ซึ่งการตัดคำที่บ่งบอกเพศออกนั้นจะเป็นขั้นตอนที่ทำหลังจากขั้นตอนการกำกับประเภทข้อความ) จะทำให้ข้อความแสดงความคิดเห็นเหล่านี้ซ้ำกันทั้งความคิดเห็น และไม่สามารถจำแนกเพศจากข้อความแสดงความคิดเห็นเหล่านี้ได้ ดังภาพประกอบ 8 แสดงตัวอย่างข้อความแสดงความคิดเห็นที่ซ้ำกันพร้อมกับจำนวนที่ซ้ำ

	comment_count
	10969
ขอบคุณครับ	1254
ขอบคุณค่ะ	827
ขอบคุณมากครับ	314
ขอบคุณมากค่ะ	285
ขอบคุณสำหรับข่าวครับ	159
ขอบคุณนะคะ	154
[Spoil] คลิกเพื่อดูข้อความที่ซ่อนไว้	124
ขอบคุณค่า	121
555	119

ภาพประกอบ 8 ตัวอย่างข้อความแสดงความคิดเห็นที่ซ้ำกันพร้อมกับจำนวนที่ซ้ำ

5. จัดการกับ Emoji ซึ่งมาในรูปของตัวหนังสือ โดยการ Tag คำว่า emoji ลงไปที่ด้านหน้าเพื่อให้ทราบว่าเป็น Emoji ไม่ใช่ตัวข้อความ ดังภาพประกอบ 9 แสดงข้อความที่มี Emoji ที่มาในรูปแบบข้อความบนข้อความแสดงความคิดเห็น จะเห็นว่า Emoji ที่ได้มานั้น จะอยู่ในรูปของข้อความที่ทางเว็บไซต์พันทิพจัดทำขึ้นมาสำหรับใช้บนเว็บไซต์

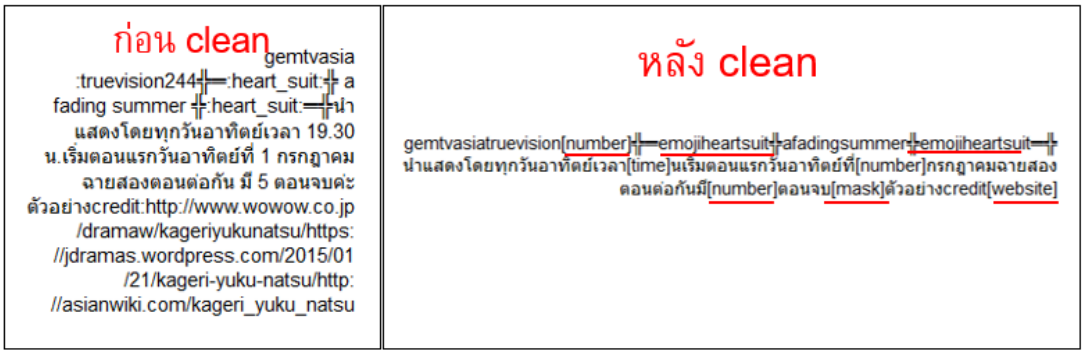
comment
ลืมไปหนึ่งเรื่องค่ะ ช่วงนี้วุ่นๆเลยเพิ่งได้มา อัพGemtvasia :Truevision244╬═:heart_suit:╬ ZERO -THE BRAVEST MONEY GAME- ╬:heart_suit:═╬นำแสดงโดย[Spoil] คลิกเพื่อดูข้อความที่ซ่อนไว้ทุกวันเสาร์ เวลา 20.30 น.เริ่มตอนแรกวันเสาร์ที่ 21 กรกฎาคมฉายสัปดาห์เดียวกับญี่ปุ่น ตัวอย่างhttps://mydramalist.com/28856-zero-ikkaku-senkin-game?trailer=1Credit :[Spoil] คลิกเพื่อดูข้อความที่ซ่อนไว้http://www.ntv.co.jp/0/http://asianwiki.com /Zero:_The_Bravest_Money_Gamehttps://mydramalist.com/28856-zero-ikkaku-senkin-game?trailer=1แก้ไขข้อความเมื่อ16 กรกฎาคม 2561 เวลา 23:21 น.

ภาพประกอบ 9 ตัวอย่างข้อความที่มี Emoji ที่มาในรูปแบบข้อความ

6. จัดการกับรูปแบบคำเฉพาะ เช่น ชื่อเว็บไซต์ วันที่ เวลา เบอร์โทรศัพท์ ที่เป็นตัวเลขโดยการเปลี่ยนให้เป็น Tag ของคำนั้นๆ ดังภาพประกอบ 10

7. จัดการกับสัญลักษณ์ต่างๆ ที่ไม่มีความหมาย โดยการเปลี่ยนเป็นข้อความว่าง

8. จัดการคำกับที่ใช้ระบุเพศของผู้เขียนข้อความทีเป็นคำลงท้าย เช่น “ครับ” “ค่ะ” โดยปกปิดคำ (Masking) ด้วยการใช้ [mask] แทนคำที่สามารถระบุเพศ ดังภาพประกอบ 10



ภาพประกอบ 10 เปรียบเทียบข้อความแสดงความคิดเห็นก่อนและหลังทำความสะอาด

การเตรียมข้อมูล (Data Pre-Processing)

1. การเตรียมพจนานุกรมคำศัพท์สำหรับใช้ในการตัดคำ

ในงานวิจัยนี้จะใช้พจนานุกรมตั้งต้นของ (Horsuwan et al., 2019) และทำการเพิ่มคำศัพท์ที่ได้จากการศึกษางานวิจัยที่เกี่ยวข้องได้แก่ คำแสดงความลังเล คำหยุด และ emoji ที่ได้จากขั้นตอนการทำความสะอาดข้อมูล

1.1 คำแสดงความลังเล คือ คำที่ทำหน้าที่ลดความชัดเจนหรือความมั่นใจของผู้ใช้ถ้อยคำ มักถูกใช้เพื่อป้องกันการโจมตีจากผู้ฟังหรือผู้อ่านที่มีความรู้มากกว่า (ทองพูล, 2559; นาคพันธุ์, 2561) โดยใช้คำที่มาจาก (Gillett; Smith, 2020) ซึ่งเป็นคำในภาษาอังกฤษ และใช้โปรแกรมแปลภาษา Google Translate แปลเป็นภาษาไทย มีทั้งสิ้น 200 คำ ดังตาราง 4 แสดงตัวอย่างคำแสดงความลังเลในภาษาอังกฤษและภาษาไทยที่ได้จากโปรแกรมแปลภาษา

ตาราง 4 ตัวอย่างคำแสดงความลังเลในภาษาอังกฤษและภาษาไทย

Hedging Word	**คำแสดงความลังเล**
assume	สมมุติ
believe	เชื่อว่า
definitely	อย่างแน่นอน
doubt	สงสัยว่า
estimate	ประมาณ

1.2 คำหยุด คือ คำทั่วๆ ไปที่สามารถพบได้บ่อยในประโยค เป็นคำที่ไม่มีนัยยะสำคัญต่อความหมายในประโยค (จิระวิชิตชัย, 2010) เมื่อตัดออกจะไม่ทำให้ใจความสำคัญของประโยคเปลี่ยนไป เช่นคำว่า นั่นไง เหตุนั้น มาก ช้านาน เพียงใด ฯลฯ เป็นต้น ซึ่งการกำจัดคำหยุดนั้นเป็นหนึ่งในขั้นตอนการจัดเตรียมข้อมูลสำหรับการประมวลผลภาษาธรรมชาติ แต่ในงานวิจัยนี้ผู้วิจัยจะนำคำหยุดมาสร้างเป็นคุณลักษณะสำหรับการสร้างแบบจำลอง โดยใช้คำหยุดภาษาไทยที่มีอยู่ในไลบรารี่ pyThaiNLP ซึ่งมีจำนวนทั้งสิ้น 1,030 คำ

1.3 Emoji คือ สัญลักษณ์แทนอารมณ์ ความรู้สึก หรือแทนสิ่งต่างๆ เช่น 😄 ☺ ❤ 🍔 ซึ่งหลังจากทำการเก็บข้อมูลจะได้มาเป็นข้อความแทนที่สัญลักษณ์ เช่น :heart_suit: แทนสัญลักษณ์ ❤ ดังภาพประกอบ 9 ดังนั้นผู้วิจัยจึงต้องทำการแยกคำปกติกับ Emojis ออกจากกัน โดยทำการสกัดข้อความ Emoji แล้วทำการรวบรวมเป็นพจนานุกรม Emoji โดยมีทั้งสิ้น 789 emojis เพื่อใช้สำหรับการตัดคำ และสร้างเป็นคุณลักษณะสำหรับการสร้างแบบจำลอง ซึ่งมีวิธีการสกัดข้อความ Emojis ดังภาพประกอบ 11

```
emoji_comment = df_selected[df_selected.comment.str.contains('[:][a-z_-]*[:]', regex=True)][['comment']]
```

```
emoji_dict_temp = []
emoji_dict = []
for i in emoji_comment.comment:
    for j in re.findall ('[:]([a-z_-]*?)[:]', i):
        if('http' in j):
            pass
        else:
            emoji_dict_temp.append(f"emoji{j}")

emoji_dict_temp = list(set(emoji_dict_temp))
emoji_dict_temp.remove('emoji')

for i in emoji_dict_temp:
    emoji_dict.append(re.sub( r"[_]","", i))
```

```
len(emoji_dict)
```

```
789
```

```
emoji_dict[10:20]
```

```
['emojiframedpicture',
 'emojiwhistle',
 'emojiavocado',
 'emojismilingfacewithhalo',
 'emojihammer',
 'emojimansurfingmedium-lightskintone',
```

ภาพประกอบ 11 ตัวอย่างโค้ดไพธอนสำหรับการสร้างพจนานุกรม Emoji (บน) และตัวอย่างคำศัพท์ Emojis (ล่าง)

2. **การตัดคำ** (Word Segmentation)

การตัดคำคือการแบ่งคำ ซึ่งในภาษาไทยนั้นจะไม่เหมือนกับการตัดคำในภาษาอังกฤษ เพราะว่าภาษาอังกฤษมีการเว้นระหว่างคำ จึงสามารถใช้ช่องว่าในการตัดคำ ส่วนในภาษาไทยนั้นจะต้องมีวิธีการตัดที่ต่างออกไป

ในงานวิจัยนี้ผู้เขียนได้เลือกใช้ไลบรารีสำหรับการประมวลผลภาษาธรรมชาติในภาษาไทยที่ชื่อว่า pyThaiNLP ซึ่งมีโมดูลสำหรับการตัดคำภาษาไทยอยู่ ผู้วิจัยเลือกใช้วิธีตัดคำแบบใช้พจนานุกรมเป็นพื้นฐาน (Dictionary base) ร่วมกับวิธีการตัดคำแบบสอดคล้องมากที่สุด (Maximal Matching) โดยทำการตัดคำสองแบบ คือ แบบที่หนึ่งใช้เฉพาะพจนานุกรมที่มาจาก (Horsuwan et al., 2019) เพื่อใช้สร้างเป็นแบบจำลองพื้นฐานสำหรับการวัดประสิทธิภาพ (baseline) และแบบที่สองเพิ่มคำศัพท์ของผู้วิจัยเองอีก 2 ชุดคำศัพท์ลงในพจนานุกรม คือ คำแสดงความลังเล และคำศัพท์ Emoji เนื่องจากการตัดคำโดยใช้พจนานุกรมเป็นพื้นฐานนั้น จะตัดคำจากพจนานุกรมที่ใช้ก่อน กรณีที่ไม่มีคำในพจนานุกรมจึงจะใช้วิธีการตัดคำแบบสอดคล้องมากที่สุดในการตัดคำ เพราะฉะนั้นคำแสดงความลังเลที่ไม่มีในพจนานุกรมจะถูกตัดแยกออกจากกัน เช่น คำว่า “ประมาณนี้” เมื่อใช้การตัดคำโดยใช้พจนานุกรมจาก (Horsuwan et al., 2019)จะตัดได้เป็นสองคำคือ “ประมาณ” และ “นี้” ส่วนถ้าตัดด้วยพจนานุกรมที่เพิ่มคำแสดงความลังลงไปในพจนานุกรมจะได้เพียงหนึ่งคำ คือ “ประมาณนี้” ทำให้เกิดเป็นคุณลักษณะที่แตกต่างกัน ซึ่งมีผลต่อการสร้างแบบจำลองสำหรับการจำแนกเพศ นอกจากนี้ยังมี Tagging ต่างๆ ที่ใช้ในขั้นตอนการทำความสะอาดข้อมูลเพิ่มเข้าไปด้วย ดังตาราง 5 แสดงตัวอย่างข้อความแสดงความคิดเห็นที่ทำความสะอาดแล้ว

ตาราง 5 เปรียบเทียบข้อความแสดงความคิดที่ทำความสะอาดแล้วก่อนตัดคำและหลังตัดคำ ด้วยพจนานุกรมทั้งสองแบบ

ข้อความที่ทำความสะอาด	ตัดคำด้วยพจนานุกรมแบบที่หนึ่ง	ตัดคำด้วยพจนานุกรมแบบที่สอง
สาเหตุที่เที่ยงวันติดยาเพราะประชดพ่อแม่พ่อแม่ไม่มีเวลาให้ประมาณนี้หรือเปล่า[mask]หรือสาเหตุหลักมาจากอะไร[mask]emojiheartsuit	[สาเหตุ, ที่, เที่ยงวัน, ติดยา, เพราะ, ประชด, พ่อแม่, พ่อแม่, ไม่มีเวลา, ให้, ประมาณ, นี้, หรือเปล่า, [mask], หรือ, สาเหตุ, หลัก, มาจาก, อะไร, [mask]], emojiheartsuit	[สาเหตุ, ที่, เที่ยงวัน, ติดยา, เพราะ, ประชด, พ่อแม่, พ่อแม่, ไม่มีเวลา, ให้, ประมาณนี้, หรือเปล่า, [mask], หรือ, สาเหตุ, หลัก, มาจาก, อะไร, [mask]], emojiheartsuit

3. การระบุชนิดของคำ (Part-of-Speech Tagging)

เป็นการระบุชนิดของคำตามหลักไวยกรณ์ โดยอิงหน้าที่ของคำเป็นหลัก ในงานวิจัยนี้ผู้เขียนได้ใช้โมดูลของ pyThaiNLP และเลือกคลังคำศัพท์ ORCHID (Virach, Naoto, & Hitoshi, 1999) ในการระบุชนิดของคำ ดังตาราง 6 แสดงชนิดของคำและตัวอย่างคำในคลังคำศัพท์ ORCHID และตาราง 7 แสดงตัวอย่างข้อความแสดงความคิดที่ระบุชนิดของคำแล้ว

ตาราง 6 ชนิดของคำและตัวอย่างคำในคลังคำศัพท์ ORCHID

ตัวย่อ	**ชนิดของคำ (Part-of-Speech tag)**	**ตัวอย่างคำ**
NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, โค้ก
NCNM	Cardinal number	หนึ่ง, สอง, สาม, 1, 2, 10
NONM	Ordinal number	ที่หนึ่ง, ที่สอง, ที่สาม, ที่1, ที่2
NLBL	Label noun	1, 2, 3, 4, ก, ข, a, b
NCMN	Common noun	หนังสือ, อาหาร, อาคาร, คน
NTTL	Title noun	ครู, พลเอก
PPRS	Personal pronoun	คุณ, เขา, ฉัน
PDMN	Demonstrative pronoun	นี่, นั้น, ที่นั่น, ที่นี่
PNTR	Interrogative pronoun	ใคร, อะไร, อย่างไร
PREL	Relative pronoun	ที่, ซึ่ง, อัน, ผู้
VACT	Active verb	ทำงาน, ร้องเพลง, กิน
VSTA	Stative verb	เห็น, รู้, คือ
VATT	Attributive verb	อ้วน, ดี, สวย
XVBM	Pre-verb auxiliary, before negator “ไม่”	เกิด, เกือบ, กำลัง
XVAM	Pre-verb auxiliary, after negator “ไม่”	ค่อย, น่า, ได้
XVMM	Pre-verb, before or after negator “ไม่”	ควร, เคย, ต้อง
XVBB	Pre-verb auxiliary, in imperative mood	กรุณา, จง, เชิญ, อย่า, ห้าม

ตาราง 6 (ต่อ)

ตัวย่อ	ชนิดของคำ (Part-of-Speech tag)	ตัวอย่างคำ
XVAE	Post-verb auxiliary	ไป, มา, ขึ้น
DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
DDBQ	Definite determiner, between noun and classifier or preceding quantitative expression	ทั้ง, อีก, เพียง
DDAQ	Definite determiner, following quantitative expression	พอดี, ถ้วน
DIAC	Indefinite determiner, following noun; allowing classifier in between	ไหน, อื่น, ต่างๆ
DIBQ	Indefinite determiner, between noun and classifier or preceding quantitative expression	บาง, ประมาณ, เกือบ
DIAQ	Indefinite determiner, following quantitative expression	กว่า, เศษ
DCNM	Determiner, cardinal number expression	หนึ่งคน, เสือ, 2 ตัว
DONM	Determiner, ordinal number expression	ที่หนึ่ง, ที่สอง, ที่สุดท้าย
ADVN	Adverb with normal form	เก่ง, เร็ว, ช้า, สม่ำเสมอ
ADVI	Adverb with iterative form	เร็วๆ , เสมอๆ , ช้าๆ
ADVP	Adverb with prefixed form	โดยเร็ว
ADVS	Sentential adverb	โดยปกติ, ธรรมดา
CNIT	Unit classifier	ตัว, คน, เล่ม
CLTV	Collective classifier	คู่, กลุ่ม, ฝูง, เชิง, ทาง,
CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง
CFQC	Frequency classifier	ครั้ง, เที่ยว

ตาราง 6 (ต่อ)

ตัวย่อ	ชนิดของคำ (Part-of-Speech tag)	ตัวอย่างคำ
CVBL	Verbal classifier	ม้วน, มัด
JCRG	Coordinating conjunction	และ, หรือ, แต่
JCMP	Comparative conjunction	กว่า, เหมือนกับ, เท่ากับ
JSBR	Subordinating conjunction	เพราะว่า, เนื่องจาก ที่, แม้ว่า, ถ้า
RPRE	Preposition	จาก, ละ, ของ, ใต้, บน
INT	Interjection	โอ้บ, โอ้, เออ, เอ๋, อ๋อ
FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
FIXV	Adverbial prefix	อย่างเร็ว
EAFF	Ending for affirmative sentence	จ๊ะ, จ้ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
PUNC	Punctuation	(,), “, ,, ;

ตาราง 7 เปรียบเทียบข้อความแสดงความคิดที่ทำความสะอาด กับข้อความแสดงความคิดที่ทำการตัดคำด้วยวิธีที่สองและระบุชนิดของคำแล้ว

Clean Comment	POS Tagging
เค้าไม่ได้เข้าไปในถ้ำหายไปในอุทยานแห่งชาติ[mask]	[(เค้า, NCMN), (ไม่ได้, NEG), (เข้าไปใน, VACT), (ถ้ำ, NCMN), (หายไป, VSTA), (ใน, RPRE), (อุทยานแห่งชาติ, NCMN)]

4. **การปกปิดคำสรรพนามแทนตัว** (Personal Pronoun Masking)

ขั้นตอนนี้เป็นหนึ่งในขั้นตอนการทำความสะอาดข้อมูล แต่เนื่องจากเราไม่สามารถระบุคำสรรพนามตัวที่บ่งบอกเพศได้ทั้งหมด จึงต้องทำการปกปิดคำสรรพนามแทนตัวหลังจากที่ทำการระบุชนิดของคำได้แล้ว โดยเราจะเลือกคำที่มีชนิดเป็นคำสรรพนาม (PPRS) ทั้งหมด และจะทำการปกปิดคำสรรพนามแทนตัวด้วยการใช้ [PPRS] แทนคำสรรพนามแทนตัวที่นั้น เพื่อป้องกัน

การลำเอียง (bias) ของแบบจำลองจากคำเหล่านี้ ดังตาราง 8 แสดงตัวอย่างข้อความหลังการปกปิดคำสรรพนามแทนตัวแล้ว

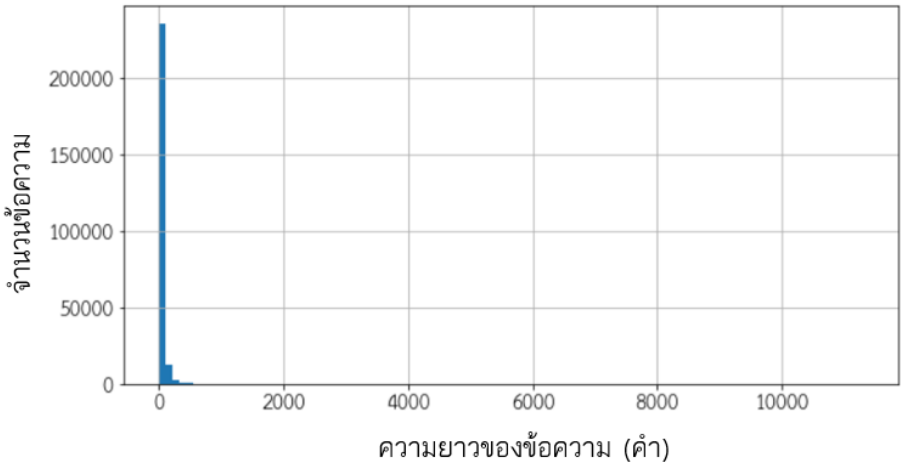
ตาราง 8 เปรียบเทียบข้อความแสดงความคิดที่ทำความสะอาด กับข้อความแสดงความคิดที่ตัดคำและทำการปกปิดคำสรรพนามแล้ว

Clean Comment	Word Segmentation with [PPRS]
ใช่[mask]ผมว่ามีเสน่ห์มากจริงๆ ปอเล่นเรื่องไหนก็เข้าถึงตีบทแตกหมดนะเราว่าแต่เรื่องนี้เราไม่ค่อยได้เห็นไงเลยรู้สึกเซอร์ไพรสมากและชอบบทนี้ของบิ๊กเอ็มมากจริงๆ [mask]	[ใช่, [mask], [PPRS], ว่า, มีเสน่ห์, มาก, จริงๆ , ปอ, เล่น, เรื่อง, ไหน, ก็, เข้าถึง, ตีบทแตก, หมด, นะ, [PPRS], ว่าแต่, เรื่อง, นี้, [PPRS], ไม่ค่อย, ได้เห็น, ไง, เลย, รู้สึก, เซ, อร, ไพ, รส, มาก, และ, ชอบ, บท, นี้, ของ, บิ๊ก, เอ็ม, มาก, จริงๆ , [mask]]

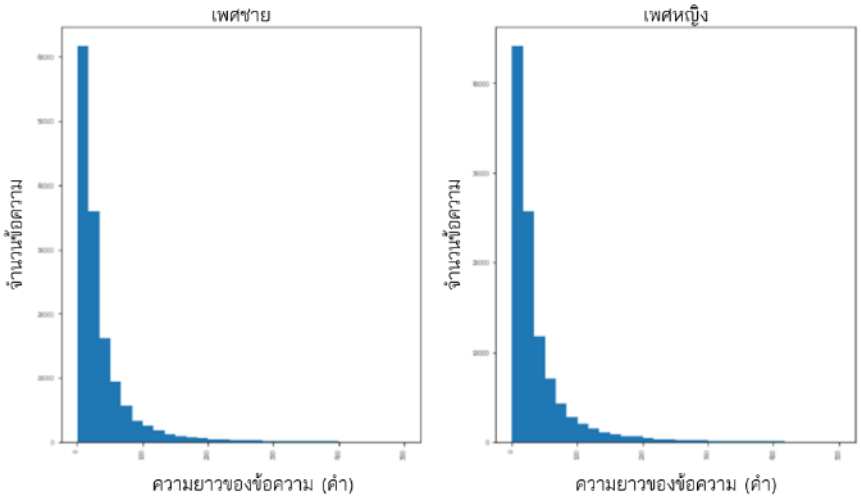
การสำรวจข้อมูลเบื้องต้น (Exploratory Data Analysis)

เมื่อทำการสำรวจการกระจายตัวของจำนวนคำของข้อความแสดงความคิดทั้งหมดที่จะนำมาใช้ในการจำแนก พบว่าจำนวนคำของข้อความแสดงความคิดกระจุกตัวกันที่อยู่ที่จำนวนคำไม่มาก ดังภาพประกอบ 12 และภาพประกอบ 13 โดยมีค่าสถิติแสดงดังตาราง 9 จะเห็นได้ว่าเพศชายและเพศหญิงมีการกระจายตัวของจำนวนคำค่อนข้างใกล้เคียงกันมาก

เนื่องจากทั้งสองเพศมีการกระจายตัวของจำนวนคำที่มีการแจกแจงไม่ปกติ ดังนั้นการจะเปรียบเทียบค่าสถิติของทั้งเพศชายและหญิงจึงต้องทำการแปลงข้อมูลเพื่อให้เปลี่ยนเป็นการแจกแจงปกติ ซึ่งจะทำให้สามารถใช้การวิเคราะห์แบบ parametric ทดสอบสมมติฐานได้ (จิรวัฒน์กุล, 2552) ผู้วิจัยจึงได้ทำการแปลงจำนวนคำด้วย Log ทำให้ได้การกระจายตัวของจำนวนคำที่มีการแจกแจงแบบปกติ ดังภาพประกอบ 14 จากนั้นจึงนำข้อมูลที่แปลงแล้วมาทำการทดสอบ T-Test พบว่าทั้งเพศชายและเพศหญิงมีค่าเฉลี่ยของจำนวนคำไม่แตกต่างกันอย่างมีค่านัยสำคัญทางสถิติ และทำการทดสอบ F-Test พบว่ามีค่าความแปรปรวนของจำนวนคำไม่มีความแตกต่างกันอย่างมีค่านัยสำคัญทางสถิติ ซึ่งมีผลการทดสอบทางสถิติ ดังภาพประกอบ 15



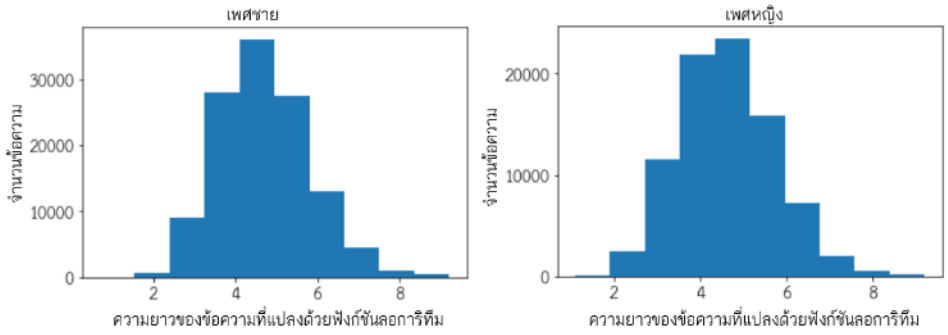
ภาพประกอบ 12 การกระจายตัวของจำนวนคำ



ภาพประกอบ 13 การกระจายตัวของจำนวนคำสำหรับข้อความแสดงความคิดเห็นที่มีจำนวนคำไม่เกิน 500 คำ เพศชาย (ซ้าย) เพศหญิง (ขวา)

ตาราง 9 เปรียบเทียบค่าสถิติการกระจายตัวของจำนวนคำในเพศชายและเพศหญิง

	เพศชาย	เพศหญิง	รวม
จำนวนความคิดเห็น	143,006	104,904	247,910
จำนวนคำเฉลี่ย	40.4	43.3	41.6
ส่วนเบี่ยงเบนมาตราฐานของจำนวนคำ	77.9	98.8	87.4
จำนวนคำต่ำสุด	1	1	1
จำนวนคำที่เปอร์เซนไทล์ที่ 25	11	11	11
จำนวนคำที่เปอร์เซนไทล์ที่ 50	21	22	21
จำนวนคำที่เปอร์เซนไทล์ที่ 75	43	46	44
จำนวนคำสูงสุด	2,793	8,850	8,850



ภาพประกอบ 14 แสดงการกระจายตัวของจำนวนคำที่แปลงค่าด้วย Log เพศชาย (ซ้าย) และเพศหญิง (ขวา)

```
from scipy.stats import ttest_ind
ttest_ind(male_len_xform, female_len_xform)
```

```
Ttest_indResult(statistic=-9.260396781705062, pvalue=2.0522491456402616e-20)
```

```
from scipy.stats import f_oneway
f_oneway(male_len_xform,female_len_xform)
```

```
F_onewayResult(statistic=85.75494855461359, pvalue=2.0522491453999795e-20)
```

ภาพประกอบ 15 ผลการทดทางสถิติ T-Test และ F-Test ของจำนวนคำระหว่างเพศชายและเพศหญิง

การแบ่งตัวอย่างข้อมูลสำหรับการสร้างแบบจำลอง (Data Splitting)

อ้างอิงจาก (Crosby & Nyquist, 1977) ที่กล่าวว่าความแตกต่างของการใช้ภาษาระหว่างเพศหญิงและชายจะแตกต่างอย่างมีนัยยะก็ต่อเมื่อมีความยาวของการใช้ภาษามากพอ ผู้วิจัยจึงได้ทำการแบ่งข้อมูลออกตามจำนวนคำ ออกเป็น 4 กลุ่ม ดังตาราง 10 โดยใช้ข้อมูลจากการกระจายตัวของจำนวนคำที่ได้จากขั้นตอนการสำรวจข้อมูลดังนี้

1. ข้อความที่มีจำนวนคำน้อยกว่าเปอร์เซนไทล์ที่ 25 (ไม่เกิน 10 คำ)
2. ข้อความที่มีจำนวนคำมากกว่าเปอร์เซนไทล์ที่ 25 ถึง Q3+1.5IQR (จำนวนคำระหว่าง 11-96)
3. ข้อความที่มีจำนวนคำมากกว่า 96 ถึง 200 คำ (จำนวนคำระหว่าง 97-200)
4. ข้อความที่มีจำนวนคำมากกว่า 200 คำ

หลังจากนั้นจึงแบ่งตัวอย่างในแต่ละกลุ่มออกเป็นชุดข้อมูลสำหรับฝึกฝน (Training dataset) และชุดข้อมูลสำหรับทดสอบ (Testing dataset) มีอัตราส่วน 80:20 โดยใช้วิธี Stratify ซึ่งเป็นการแบ่งข้อมูลโดยให้คงสัดส่วนของข้อมูลในแต่ละเพศเท่ากันทั้งชุดข้อมูลสำหรับฝึกฝนและชุดข้อมูลสำหรับทดสอบ

ตาราง 10 จำนวนข้อความแสดงความคิดเห็นของแต่ละกลุ่มตัวอย่างแยกเพศชายและเพศหญิง

จำนวนคำ	จำนวนคำเฉลี่ย	จำนวนข้อความแสดงความคิดเห็น		
		เพศชาย	เพศหญิง	รวม
<= 10	6.8	34,841	24,783	59,624
11-96	32.7	96,392	70,238	166,630
97-200	133.5	8,577	7,218	15,795
> 200	400.2	3,196	2,665	5,861

การสกัดคุณลักษณะ (Features Extraction / Features Engineering)

เป็นการเปลี่ยนข้อมูลที่มีให้กลายเป็นคุณลักษณะสำหรับการสร้างแบบจำลองการจำแนก จากการศึกษางานวิจัยที่เกี่ยวข้อง ผู้วิจัยพบว่าในงานวิจัยทางด้านการจำแนกเพศของผู้เขียนข้อความนั้น ผู้วิจัยส่วนใหญ่มักจะเลือกใช้ใช้เทคนิค TF หรือ TF-IDF เนื่องจากเป็นเทคนิคเกี่ยวข้องกับคำศัพท์ในข้อความโดยตรง (แทนแต่ละคำศัพท์ด้วยคุณลักษณะ) แม้กระทั่งงานวิจัยที่เป็น State-of-Art ในภาษาอังกฤษ (Mukherjee & Liu, 2010) ก็ยังเลือกใช้เทคนิค TF แม้ว่าหลังจากนั้นจะมีผู้วิจัยหลายท่านพยายามที่จะใช้อัลกอริทึมเป็นสถาปัตยกรรมใหม่ๆ ในงานวิจัยการจำแนกเพศจากข้อความ ก็ยังไม่มีใครสามารถทำผลลัพธ์การจำแนกเพศได้มากกว่าถึงแม้จะผ่านมามากกว่า 10 ปีแล้วก็ตาม ซึ่งอัลกอริทึมที่เป็นสถาปัตยกรรมใหม่ๆ นั้นจะเน้นที่ความหมายของประโยค ซึ่งจะเหมาะกับงานด้านอื่นๆ ที่ใช้ระบบการประมวลผลภาษาธรรมชาติ เช่น การวิเคราะห์ความรู้สึกจากข้อความ (Sentiment Analysis) หรือการจำแนกหัวข้อของข้อความ (Topic Classification) เป็นต้น ดังนั้นในงานวิจัยนี้ผู้วิจัยจึงได้เลือกใช้เทคนิค TF-IDF (n-gram 1,2) ในการสกัดคุณลักษณะ โดยเทคนิค n-gram 1,2 นั้น จะเป็นการนำเอาคำหนึ่งคำและคำสองคำที่อยู่ติดกันมาสกัดเป็นคุณลักษณะ ซึ่งมาจากสมมติฐานของผู้วิจัยที่ตั้งสมมติฐานว่าเพศชายและเพศหญิงมีการใช้รูปแบบของคำที่แตกต่างกัน จึงไม่ใช้เพียงแค่คำๆ เดียวเท่านั้น แต่จะใช้สองคำที่อยู่ติดกันเพื่อสร้างรูปแบบการใช้คำ และสกัดเป็นคุณลักษณะขึ้นมา ซึ่งแบ่งตามคุณลักษณะดังต่อไปนี้

1. สกัดคุณลักษณะจากคำศัพท์ทั่วไป ได้คุณลักษณะจำนวน 5,000 คุณลักษณะ ดังภาพประกอบ 16 แสดงตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-

gram 1,2) กับคำศัพท์ทั่วไป ตัวเลขที่ได้จากการสกัดคุณลักษณะเป็นตัวเลขที่คำนวณมาจากวิธี TF-IDF ซึ่งค่า 0 หมายถึงไม่มีคำๆ นั้น อยู่ในข้อความแสดงความคิดเห็นนั้นๆ ดังภาพประกอบ 16

[date]	[date] [mask]	[date] [number]	[date] เวลา	[date] แผนก	[email]	[mask]	[mask] [number]	[mask] [website]	[mask] tt	...
0.0	0.0	0.0	0.0	0.0	0.0	0.047868	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	0.0	0.0	0.032545	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	0.0	0.0	0.023495	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	0.0	0.0	0.021493	0.0	0.0	0.0	...
0.0	0.0	0.0	0.0	0.0	0.0	0.057785	0.0	0.0	0.0	...

ภาพประกอบ 16 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับคำศัพท์ทั่วไป

2. สกัดคุณลักษณะจาก Emoji ได้คุณลักษณะจำนวน 789 คุณลักษณะ ดังภาพประกอบ 17 แสดงตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-gram 1,2) กับ emojis

emojipotoffood	emojiwhiteheavycheckmark	emojichildrencrossing	emojiyinyang	emojimonkeyface
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0

ภาพประกอบ 17 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับ emojis

3. สกัดคุณลักษณะจากคำแสดงความลังเลได้คุณลักษณะจำนวน 200 คุณลักษณะ ดังภาพประกอบ 18 แสดงตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-gram 1,2) กับคำแสดงความลังเล

ก็	ก็ ตาม ที่	กระจ่าง แจ้ง	กระจ่าง ชัด	กะ จ่าย เลย	กิจวัตร	ขอ ให้	คา หนัง คาเขา	คง	คง จะ	...	อะไร อย่างงี้	อะไร อย่าง นั้น	อะไร อย่างนี้
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

ภาพประกอบ 18 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับคำแสดงความลังเล

4. สกัดคุณลักษณะจากคำหยุดได้คุณลักษณะจำนวน 1,030 คุณลักษณะ ดังภาพประกอบ 19 แสดงตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-gram 1,2) กับคำหยุด

เช่นที่	เล็ก	อันได้แก่	เหตุ	ทุกที่	ซะ	รับรอง	แยะๆ	สู่	เป็นการ	...	เกี่ยวเนื่อง	เช่นไร	เฉกเช่น
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

ภาพประกอบ 19 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับคำหยุด

5. สกัดคุณลักษณะจากชนิดของคำได้คุณลักษณะจำนวน 1,239 1,722 1,615 และ 1,606 คุณลักษณะ สำหรับแต่ละกลุ่มตัวอย่างที่มีจำนวนคำไม่เกิน 10 คำ, 11-96 คำ, 97-200 คำ และมากกว่า 200 คำ ตามลำดับ ซึ่งตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-gram 1,2) กับชนิดของคำ แสดงดังภาพประกอบ 20

ADVI	ADVI ADVN	ADVI ADVP	ADVI ADVS	ADVI CFQC	ADVI CNIT	ADVI DCNM	ADVI DDAC	ADVI DDAN	ADVI DDBQ	...	XVMM PPRS	XVMM PREL	XVMM RPRE
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

ภาพประกอบ 20 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับชนิดของคำ

การสร้างแบบจำลองการจำแนก (Classification Model Training and Evaluation)

ในขั้นตอนการสร้างแบบจำลอง ผู้วิจัยได้ทำการเลือกอัลกอริทึมสำหรับงานประเภทการจำแนกจำนวน 3 อัลกอริทึม คือ Logistic Regression, Naive Bayes และ Random Forest โดยจะใช้คุณลักษณะทั้ง 5 ประเภทที่สกัดได้ มาทำการจัดกลุ่ม (combination) ซึ่งแต่ละกลุ่มจะมีคุณลักษณะตั้งแต่ 1 คุณลักษณะ จนถึง 5 คุณลักษณะ โดยในกลุ่มคุณลักษณะที่มีตั้งแต่ 2 คุณลักษณะขึ้นไปจะใช้คุณลักษณะที่ได้จากวิธี TF-IDF ที่ใช้พจนานุกรมที่เพิ่มคำศัพท์ใหม่เข้าไปเป็นคุณลักษณะพื้นฐานสำหรับทุกกลุ่มคุณลักษณะ และเมื่อรวมกับคุณลักษณะที่ได้จากวิธี TF-IDF ที่ใช้พจนานุกรมจาก (Horsuwan และคนอื่น ๆ, 2019) จะได้กลุ่มคุณลักษณะ 21 กลุ่มคุณลักษณะ จากนั้นจึงนำชุดคุณลักษณะที่ได้มาทำการทดลองเพื่อหาว่าชุดคุณลักษณะและอัลกอริทึมใดที่ให้ประสิทธิภาพในการจำแนกได้ดีที่สุด และยังใช้อัลกอริทึมสำหรับการเลือกคุณลักษณะอีก 2 อัลกอริทึม ได้แก่ SelectKBest และ Recursive Feature Elimination with Cross-Validation (RFECV) กับชุดคุณลักษณะที่มีคุณลักษณะทั้ง 5 คุณลักษณะ จึงมีกลุ่มคุณลักษณะรวมทั้งสิ้น 23 กลุ่มคุณลักษณะที่ใช้ในการสร้างแบบจำลอง หลังจากนั้นจึงทำการวัดประสิทธิภาพของแบบจำลองโดยใช้ cross-validation score บนชุดข้อมูลฝึกฝน โดยใช้วิธี Stratified K-Fold (5 folds) ซึ่งมีชุดคุณลักษณะที่ใช้ทดลองดังนี้

1. emoji (E)
2. คำแสดงความลังเล (H)
3. คำหยุด (S)
4. ชนิดของคำ (T)
5. TF-IDF (สำหรับค่าประสิทธิภาพพื้นฐาน(Baseline)) (TFIDF1)
6. TF-IDF+คำศัพท์ใหม่ (TFIDF2)
7. TF-IDF+คำศัพท์ใหม่, ชนิดของคำ

8. TF-IDF+คำศัพท์ใหม่, คำหยุด

9. TF-IDF+คำศัพท์ใหม่, ชนิดของคำ, คำหยุด

10. TF-IDF+คำศัพท์ใหม่, emoji

11. TF-IDF+คำศัพท์ใหม่, emoji, ชนิดของคำ

12. TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด

13. TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด, ชนิดของคำ

14. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล

15. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, ชนิดของคำ

16. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, คำหยุด

17. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, ชนิดของคำ, คำหยุด

18. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji

19. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, ชนิดของคำ

20. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด

21. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ชนิดของคำ

22. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ชนิดของคำ (SelectKBest)

23. TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ชนิดของคำ (RFECV)

การหาค่าความสำคัญของคุณลักษณะ (Feature Importance)

ค่าความสำคัญของคุณลักษณะนั้น สามารถบอกถึงคุณลักษณะที่มีผลต่อการจำแนกเพศในแบบจำลอง ในแต่ละอัลกอริทึมจะมีวิธีหาที่แตกต่างกัน ซึ่งสามารถหาค่าความสำคัญของคุณลักษณะของแต่ละอัลกอริทึมได้ดังนี้

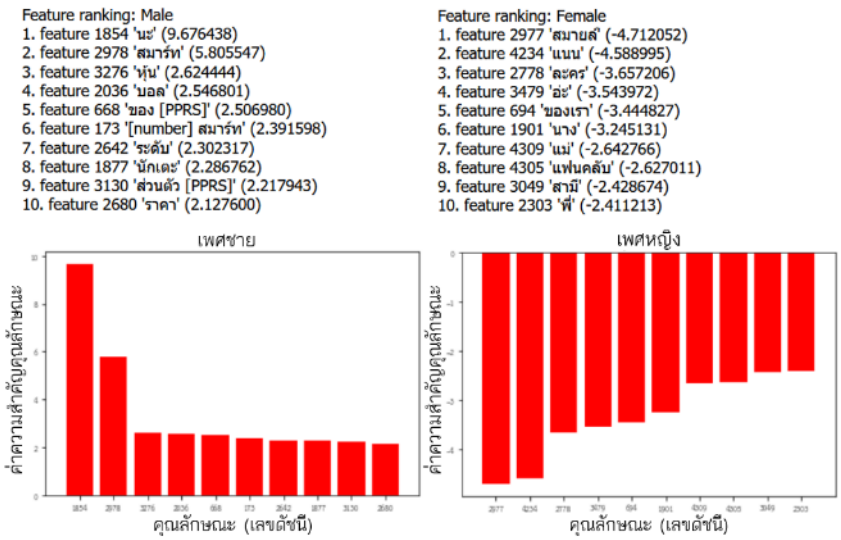
1. ค่าความสำคัญจากอัลกอริทึม Logistic Regression

เนื่องจาก Logistic Regression เป็นแบบจำลองเชิงเส้น ค่าความสำคัญของคุณลักษณะจะสามารถหาได้จากค่าสัมประสิทธิ์ของคุณลักษณะ (Coefficient of Features) ซึ่งในโมดูล Logistic Regression ของ scikit-learn มีฟังก์ชั่นที่สามารถหาค่าสัมประสิทธิ์ของแต่ละคุณลักษณะได้ ดังภาพประกอบ 21 แสดงตัวอย่างโค้ดโปรแกรมภาษาไพธอนในการหาค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรก ซึ่งความหมายของค่าสัมประสิทธิ์กับการจำแนกคือ ค่าสัมประสิทธิ์ของคุณลักษณะใดมีค่ามาก คุณลักษณะนั้นจะมีผลต่อการจำแนกสำหรับเพศ

ชายมาก ในขณะเดียวกันถ้าค่าสัมประสิทธิ์ของคุณลักษณะใดมีค่าน้อย จะมีผลต่อการจำแนกสำหรับเพศหญิงมาก ดังภาพประกอบ 22 แสดง 10 อันดับแรกของคุณลักษณะที่มีผลต่อการจำแนก และค่าความสำคัญของคุณลักษณะที่ได้จากค่าสัมประสิทธิ์ของแต่ละคุณลักษณะจากแบบจำลอง

```
feature_names = tfidf_wHES.get_feature_names()

importances = clf20.coef_[0]
indices = np.argsort(importances)[::-1]

# Print the feature ranking
print("Feature ranking: Male")

for f in range(10):
    print("%d. feature %d '%s' (%f)" % (f + 1, indices[f], feature_names[indices[f]], importances[indices[f]]))

plt.figure()
plt.title("Feature importances: Male", fontsize=20)
plt.bar(range(10), importances[indices[0:10]],
        color="r", align="center")
plt.xticks(range(10), indices[0:10])
plt.xlim([-1, 10])
plt.show()
```

ภาพประกอบ 21 ตัวอย่างโค้ดโปรแกรมภาษาไพธอน สำหรับการหาค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรกสำหรับอัลกอริทึม Logistic Regression



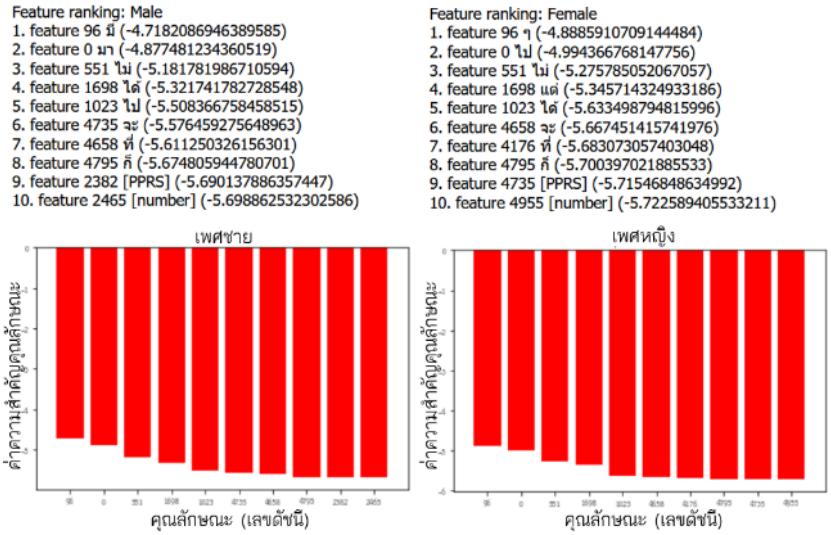
ภาพประกอบ 22 คุณลักษณะที่มีผลต่อการจำแนกและค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรกที่ได้จากค่าสัมประสิทธิ์ของแต่ละคุณลักษณะจากแบบจำลอง

2. ค่าความสำคัญจากอัลกอริทึม Naïve Bayes

อัลกอริทึม Naïve Bayes เป็นแบบจำลองที่ใช้หลักการคำนวณความน่าจะเป็นในการจำแนก ค่าความสำคัญของคุณลักษณะจะสามารถหาได้จากค่าความน่าจะเป็นของคุณลักษณะ (Probability of Features) ซึ่งในโมดูล Naïve Bayes ของ scikit-learn มีฟังก์ชั่นที่สามารถหาค่าความน่าจะเป็นของแต่ละคุณลักษณะได้ ดังภาพประกอบ 23 แสดงตัวอย่างโค้ดโปรแกรมภาษาไพธอนในการหาค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรก ซึ่งความหมายของค่าความน่าจะเป็นกับการจำแนก คือ ค่าความน่าจะเป็นของคุณลักษณะใดมีค่ามาก คุณลักษณะนั้นจะมีผลต่อการจำแนกสำหรับเพศชายมาก ในขณะเดียวกันถ้าค่าความน่าจะเป็นของคุณลักษณะใดมีค่าน้อย จะมีผลต่อการจำแนกสำหรับเพศหญิงมาก ดังภาพประกอบ 24 แสดงตัวอย่าง 10 อันดับคุณลักษณะที่มีผลต่อการจำแนก และค่าความสำคัญของคุณลักษณะที่ได้จากค่าความน่าจะเป็นของแต่ละคุณลักษณะจากแบบจำลอง

```
pos_class_prob_sorted = clf20.feature_log_prob_[1, :].argsort()[-10:]
male_feature = np.take(tfidf_wHES.get_feature_names(), pos_class_prob_sorted)

print("Feature ranking: Male")

plot_prob = []
for i in pos_class_prob_sorted:
    plot_prob.append(clf20.feature_log_prob_[1, i])

plot_prob.reverse()
pos_class_prob_sorted = pos_class_prob_sorted.tolist()
pos_class_prob_sorted.reverse()

for f in range(10):
    print(f'{f+1}. feature {pos_class_prob_sorted[f]} {male_feature[f]} ({plot_prob[f]})')

plt.figure()
plt.title("Feature importances: Male", fontsize=20)
plt.bar(range(10), plot_prob,
        color="r", align="center")
plt.xticks(range(10), pos_class_prob_sorted)
plt.xlim([-1, 10])
plt.show()
```

ภาพประกอบ 23 ตัวอย่างโค้ดโปรแกรมภาษาไพธอน สำหรับการหาค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรกสำหรับอัลกอริทึม Naïve Bayes



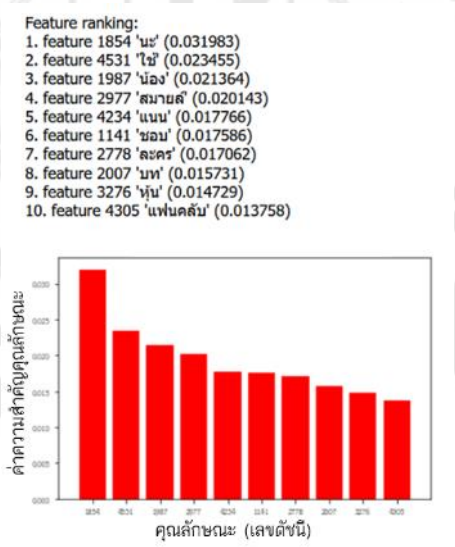
ภาพประกอบ 24 คุณลักษณะที่มีผลต่อการจำแนก และค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรกที่ได้จากค่าความน่าจะเป็นของแต่ละคุณลักษณะจากแบบจำลอง

3. ค่าความสำคัญจากอัลกอริทึม Random Forest

อัลกอริทึม Random Forest เป็นเพียงอัลกอริทึมเดียวจากสามอัลกอริทึมที่ใช้ในงานวิจัยนี้ที่ไม่สามารถบอกได้ว่าคุณลักษณะใดมีผลต่อการจำแนกเพศใด เนื่องจากการหาค่าความสำคัญของคุณลักษณะของอัลกอริทึม Random Forest จะดูจากค่าความบริสุทธิ์ (Impurity) ซึ่งสามารถบอกได้เพียงแค่ว่าคุณลักษณะใดสามารถจำแนกทั้งสองเพศออกจากกันได้ดี ซึ่งในโมดูล Random Forest ของ scikit-learn มีฟังก์ชั่นที่สามารถหาค่าความบริสุทธิ์ของแต่ละคุณลักษณะได้ ดังภาพประกอบ 25 แสดงตัวอย่างโค้ดโปรแกรมภาษาไพธอนในการหาค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรก ซึ่งความหมายของค่าความบริสุทธิ์กับการจำแนก คือ ค่าความบริสุทธิ์ของคุณลักษณะใดมีค่ามาก คุณลักษณะนั้นจะมีผลต่อการจำแนกมาก ดังภาพประกอบ 26 แสดงตัวอย่าง 10 อันดับคุณลักษณะที่มีผลต่อการจำแนก และค่าความสำคัญของคุณลักษณะที่ได้จากค่าความบริสุทธิ์ของแต่ละคุณลักษณะจากแบบจำลอง

```
feature_names = tfidf_wHES.get_feature_names()

importances = clf20.feature_importances_
indices = np.argsort(importances)[::-1]

# Print the feature ranking
print("Feature ranking: Male")

for f in range(10):
    print("%d. feature %d '%s' (%f)" % (f + 1, indices[f], feature_names[indices[f]], importances[indices[f]]))

# Plot the impurity-based feature importances of the forest
plt.figure()
plt.title("Feature importances: Male", fontsize=20)
plt.bar(range(10), importances[indices[0:10]],
        color="r", align="center")
plt.xticks(range(10), indices[0:10])
plt.xlim([-1, 10])
plt.show()
```

ภาพประกอบ 25 ตัวอย่างโค้ดโปรแกรมภาษาไพธอน สำหรับการหาค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรกสำหรับอัลกอริทึม Random Forest



ภาพประกอบ 26 คุณลักษณะที่มีผลต่อการจำแนก และค่าความสำคัญของคุณลักษณะสูงสุด 10 อันดับแรกที่ได้จากค่าความบริสุทธิ์ของแต่ละคุณลักษณะจากแบบจำลอง

การทดสอบแบบจำลอง (Model Testing)

นำแบบจำลองที่ได้ทั้งหมด 276 แบบจำลอง (จาก 3 อัลกอริทึม 4 ชุดข้อมูล 23 ชุดคุณลักษณะ) ไปทำการทดสอบกับชุดข้อมูลทดสอบที่ได้จากการแบ่งข้อมูล 20% จากแต่ละชุดข้อมูลในขั้นตอนการแบ่งตัวอย่างข้อมูลสำหรับการสร้างแบบจำลองโดยวิธี Stratify ซึ่งเป็นการแบ่ง

ข้อมูลโดยให้คงสัดส่วนของข้อมูลในแต่ละเพศเท่ากับชุดข้อมูลสำหรับฝึกฝน เพื่อดูว่าแบบจำลองใดที่สามารถให้ค่าประสิทธิภาพที่ดีที่สุด โดยใช้ Confusion Matrix ในการวัดประสิทธิภาพของแบบจำลอง ซึ่งมีทั้งหมด 4 ค่า ได้แก่ Accuracy, Precision, Recall และ F1 Score

บทที่ 4
ผลการดำเนินงานวิจัย

ในการวิจัยการจำแนกเพศจากข้อความบนโซเชียลเน็ตเวิร์ก ผู้วิจัยได้ใช้ข้อความแสดงความคิดเห็นบนเว็บไซต์พันทิพภายในระยะเวลา 2 เดือน ตั้งแต่เดือนมิถุนายน พ.ศ.2562 ถึงเดือนกรกฎาคม พ.ศ.2562 และเลือกเฉพาะข้อความแสดงความคิดเห็นที่สามารถระบุเพศของผู้เขียนข้อความแสดงความคิดเห็นได้มาใช้ในงานวิจัย จำนวนทั้งสิ้น 247,910 ข้อความ โดยมีการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติในการสกัดคุณลักษณะ และทำการสร้างแบบจำลองการจำแนกเพศโดยใช้เทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้ดำเนินการวิจัยโดยการศึกษาตามกระบวนการและขั้นตอนต่างๆ ตลอดจนการวัดประสิทธิภาพ เพื่อให้บรรลุวัตถุประสงค์ของการวิจัยที่ได้กำหนดไว้ โดยมีผลการดำเนินงานวิจัยดังนี้

1. ประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำไม่เกิน 10 คำ
2. ประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 11 – 96 คำ
3. ประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 97 – 200 คำ
4. ประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำมากกว่า 200 คำ

ประสิทธิภาพของแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำไม่เกิน 10 คำ

ประสิทธิภาพของแบบจำลองแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

จากการสร้างแบบจำลองการจำแนกเพศด้วยอัลกอริทึม 3 อัลกอริทึม ได้แก่ Logistic Regression, Naïve Bayes และ Random Forest บนชุดข้อมูลสำหรับการสร้างแบบจำลองที่มีจำนวนคำไม่เกิน 10 คำ และทำการทดสอบประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำไม่เกิน 10 คำ โดยใช้ค่า Weighted average ของ Accuracy, Precision, Recall และ F1 Score เป็นตัววัดประสิทธิภาพของแบบจำลอง ได้ผลลัพธ์ดังตาราง 11 ตาราง 12 และตาราง 13 จะเห็นได้ว่าค่าประสิทธิภาพของแบบจำลองที่ดีที่สุดมีค่า Accuracy 67.23%, Precision 66.73%, Recall 67.23% และ F1 Score 66.42% ซึ่งใช้คุณลักษณะ TF-IDF ที่สกัดจากคำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด และชนิดของคำ ร่วมกับอัลกอริทึม Logistic Regression และใช้อัลกอริทึม SelectKBest ในการเลือกคุณลักษณะที่ดีที่สุดในการสร้างแบบจำลอง

ตาราง 11 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Logistic Regression บนชุดข้อมูลทดสอบที่มีจำนวนคำไม่เกิน 10 คำ

คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	59.54%	67.08%	59.54%	46.25%
คำแสดงความลังเล	200	58.38%	54.67%	58.38%	45.34%
คำหยุด	1,030	59.79%	58.07%	59.79%	55.87%
ชนิดของคำ	1,239	59.25%	57.32%	59.25%	51.92%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	65.64%	65.06%	65.64%	64.28%
TF-IDF(คำศัพท์ใหม่)	5,000	65.66%	65.10%	65.66%	64.25%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,239	65.66%	65.06%	65.66%	64.45%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	66.81%	66.28%	66.81%	65.95%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,239	66.85%	66.33%	66.85%	66.10%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	66.10%	65.60%	66.10%	64.73%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,239	66.16%	65.63%	66.16%	64.92%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	67.13%	66.63%	67.13%	66.23%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,239	67.13%	66.62%	67.13%	66.36%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	65.62%	65.04%	65.62%	64.27%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,239	65.65%	65.04%	65.65%	64.50%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	66.94%	66.42%	66.94%	66.11%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,239	66.95%	66.44%	66.95%	66.22%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	65.95%	65.40%	65.95%	64.61%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,239	65.96%	65.40%	65.96%	64.74%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	67.08%	66.57%	67.08%	66.25%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,239	67.02%	66.51%	67.02%	66.26%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	**4,239**	**67.23%**	**66.73%**	**67.23%**	**66.42%**
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	4,373	67.09%	66.59%	67.09%	66.31%

ตาราง 12 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes บนชุดข้อมูลทดสอบที่มีจำนวนคำไม่เกิน 10 คำ

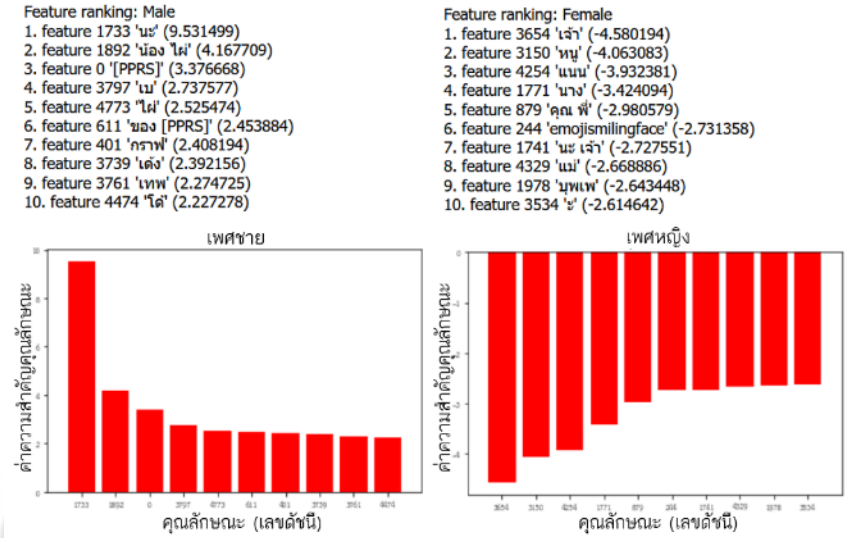
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	59.42%	67.82%	59.42%	45.84%
คำแสดงความลังเล	200	58.32%	54.27%	58.32%	45.42%
คำหยุด	1,030	59.87%	58.34%	59.87%	53.75%
ชนิดของคำ	1,239	58.97%	57.10%	58.97%	48.90%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	65.60%	65.18%	65.60%	63.76%
TF-IDF(คำศัพท์ใหม่)	5,000	65.63%	65.22%	65.63%	63.75%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,239	65.47%	64.89%	65.47%	64.03%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	66.61%	66.27%	65.93%	65.00%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,239	66.43%	65.93%	66.43%	65.21%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	66.07%	65.74%	66.07%	64.25%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,239	65.93%	65.41%	65.93%	64.52%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	66.78%	66.50%	66.78%	65.14%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,239	66.83%	66.39%	66.83%	65.58%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	65.58%	65.15%	65.58%	63.73%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,239	65.49%	64.91%	65.49%	64.08%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	66.74%	66.43%	66.74%	65.14%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,239	66.49%	65.99%	66.49%	65.28%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	65.79%	65.39%	65.79%	63.99%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,239	65.84%	65.29%	65.84%	64.45%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	**5,000**	**66.94%**	**66.66%**	**66.94%**	**65.32%**
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,239	66.79%	66.33%	66.79%	65.58%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	5,204	66.83%	66.37%	66.83%	65.61%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,239	66.79%	66.33%	66.79%	65.58%

ตาราง 13 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Random Forest บนชุดข้อมูลทดสอบที่มีจำนวนคำไม่เกิน 10 คำ

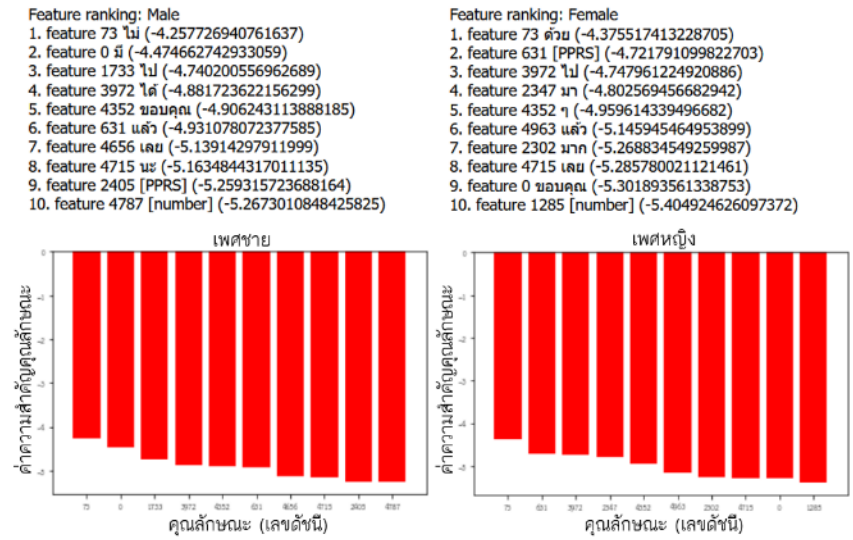
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	58.88%	69.03%	58.88%	44.29%
คำแสดงความลังเล	200	58.50%	64.19%	58.50%	43.33%
คำหยุด	1,030	58.57%	67.77%	58.57%	43.48%
ชนิดของคำ	1,239	58.49%	62.61%	58.49%	43.32%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	58.88%	74.41%	58.88%	44.13%
TF-IDF(คำศัพท์ใหม่)	5,000	58.56%	75.75%	58.56%	43.39%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,239	58.63%	74.18%	58.63%	43.57%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	58.67%	72.13%	58.67%	43.68%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,239	58.72%	71.09%	58.72%	43.82%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	58.58%	73.69%	58.58%	43.46%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,239	58.50%	75.73%	58.50%	43.25%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	58.71%	73.56%	58.71%	43.76%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,239	58.68%	71.42%	58.68%	43.73%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	58.60%	75.77%	58.60%	43.48%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,239	58.59%	75.77%	58.59%	43.46%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	58.76%	74.81%	58.76%	43.85%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,239	58.61%	75.77%	58.61%	43.50%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	58.58%	73.69%	58.58%	43.46%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,239	58.56%	75.75%	58.56%	43.39%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	58.92%	73.93%	58.92%	44.24%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,239	58.61%	75.77%	58.61%	43.50%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	3,400	58.84%	71.24%	58.84%	44.13%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	**19**	**60.96%**	**62.16%**	**60.96%**	**52.57%**

คุณลักษณะที่มีผลต่อการจำแนกแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

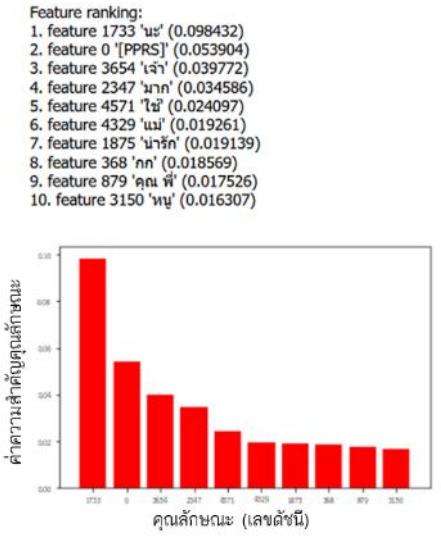
คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ 10 อันดับแรกจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดในแต่ละอัลกอริทึม ซึ่งแสดงดังภาพประกอบ 27 ภาพประกอบ 28 และภาพประกอบ 29



ภาพประกอบ 27 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Logistic Regression



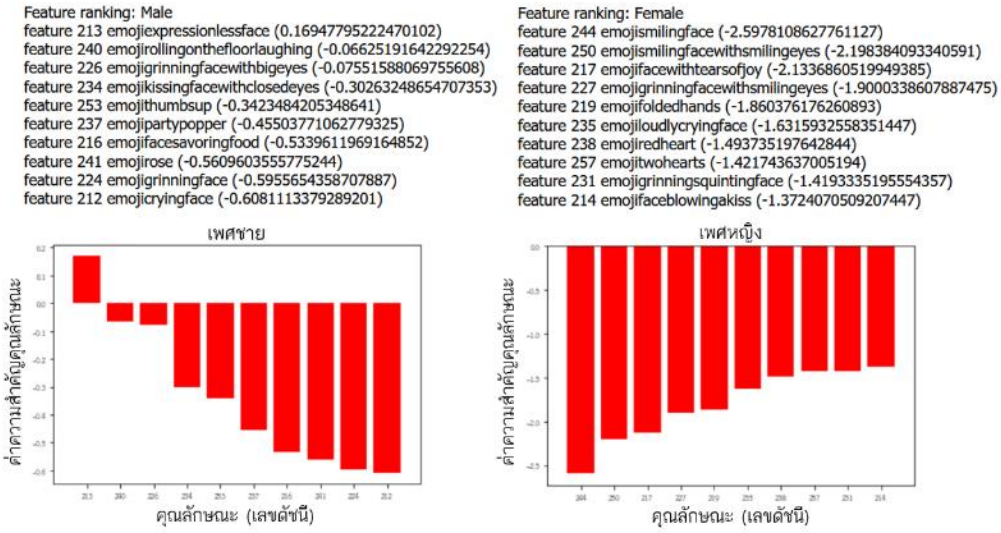
ภาพประกอบ 28 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes



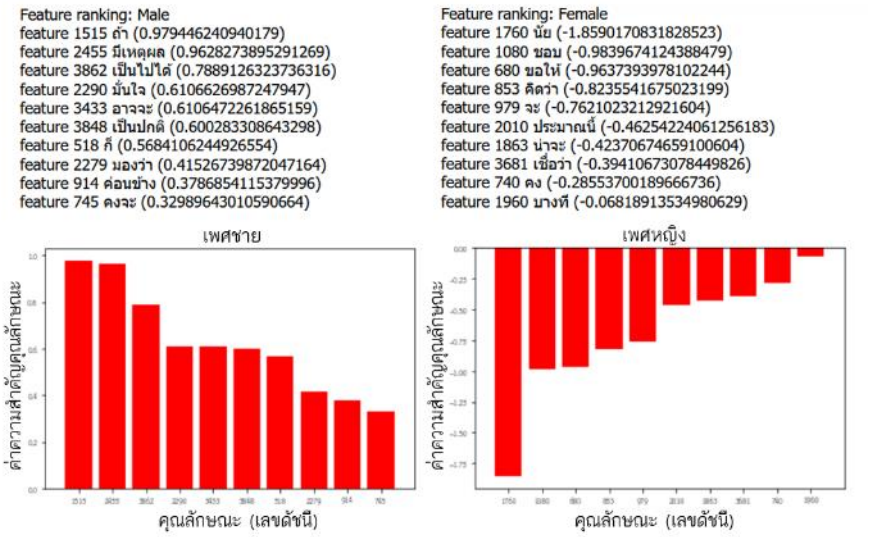
ภาพประกอบ 29 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Random Forest

คุณลักษณะที่มีผลต่อการจำแนกแยกตามประเภทคุณลักษณะ

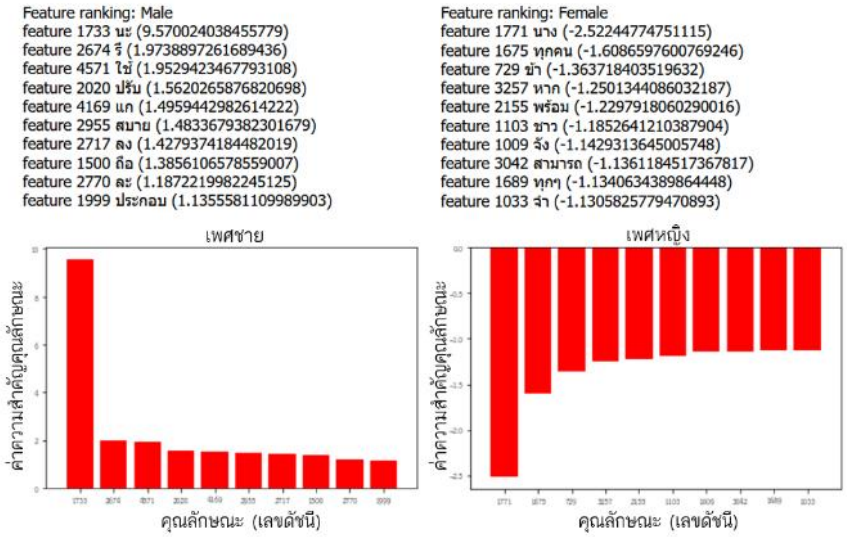
คุณลักษณะที่มีผลต่อการจำแนกเพศจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ 10 อันดับแรก แยกตามประเภทคุณลักษณะ ซึ่งแสดงดังภาพประกอบ 30 ภาพประกอบ 31 ภาพประกอบ 32 และภาพประกอบ 33



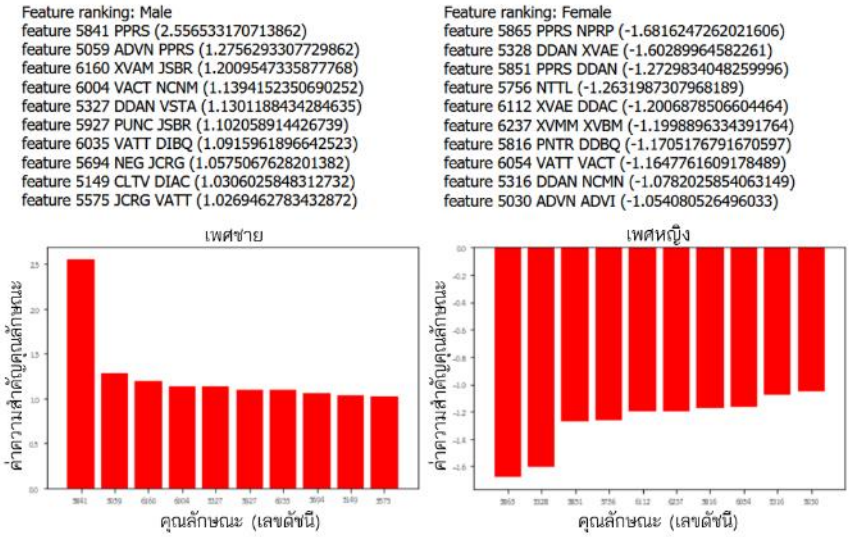
ภาพประกอบ 30 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ จากคุณลักษณะ emoji 10 อันดับแรก



ภาพประกอบ 31 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ จากคุณลักษณะคำแสดงความลังเล 10 อันดับแรก



ภาพประกอบ 32 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ จากคุณลักษณะคำหยุด 10 อันดับแรก



ภาพประกอบ 33 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำไม่เกิน 10 คำ จากคุณลักษณะชนิดของคำ 10 อันดับแรก

ประสิทธิภาพของแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 11 – 96 คำ

ประสิทธิภาพของแบบจำลองแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

จากการสร้างแบบจำลองการจำแนกเพศด้วยอัลกอริทึม 3 อัลกอริทึม ได้แก่ Logistic Regression, Naïve Bayes และ Random Forest บนชุดข้อมูลสำหรับการสร้างแบบจำลองที่มีจำนวนคำระหว่าง 11 – 96 คำ และทำการทดสอบประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 11 – 96 คำ โดยใช้ค่า Weighted average ของ Accuracy, Precision, Recall และ F1 Score เป็นตัววัดประสิทธิภาพของแบบจำลอง ได้ผลลัพธ์ดังตาราง 14 ตาราง 15 และตาราง 16 จะเห็นได้ว่าค่าประสิทธิภาพของแบบจำลองที่ดีที่สุดมีค่า Accuracy 76.62%, Precision 76.49%, Recall 76.62% และ F1 Score 76.40% ซึ่งใช้คุณลักษณะ TF-IDF ที่สกัดจากคำศัพท์ใหม่, emoji, คำหยุด และชนิดของคำ ร่วมกับอัลกอริทึม Logistic Regression ในการสร้างแบบจำลอง

ตาราง 14 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Logistic Regression บนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 11 – 96 คำ

คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	59.34%	68.95%	59.34%	46.29%
คำแสดงความลังเล	200	59.20%	59.38%	59.20%	49.36%
คำหยุด	1,030	66.12%	65.63%	66.12%	65.34%
ชนิดของคำ	1,239	60.66%	59.53%	60.66%	57.66%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	74.50%	74.38%	74.50%	74.12%
TF-IDF(คำศัพท์ใหม่)	5,000	74.42%	74.29%	74.42%	74.03%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,239	74.80%	74.67%	74.80%	74.45%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	76.34%	76.21%	76.34%	76.11%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,239	76.49%	76.36%	76.49%	76.28%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	74.68%	74.57%	74.68%	74.30%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,239	75.02%	74.91%	75.02%	74.68%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	76.47%	76.34%	76.47%	76.23%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	**6,239**	**76.62%**	**76.49%**	**76.62%**	**76.40%**
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	74.51%	74.39%	74.51%	74.13%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,239	74.74%	74.61%	74.74%	74.39%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	76.27%	76.14%	76.27%	76.03%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,239	76.41%	76.28%	76.41%	76.19%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	74.69%	74.58%	74.69%	74.31%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,239	74.90%	74.78%	74.90%	74.55%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	76.41%	76.29%	76.41%	76.17%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,239	76.55%	76.43%	76.55%	76.33%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,011	76.40%	76.27%	76.40%	76.16%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	5,378	76.57%	76.44%	76.57%	76.35%

ตาราง 15 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes บนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 11 – 96 คำ

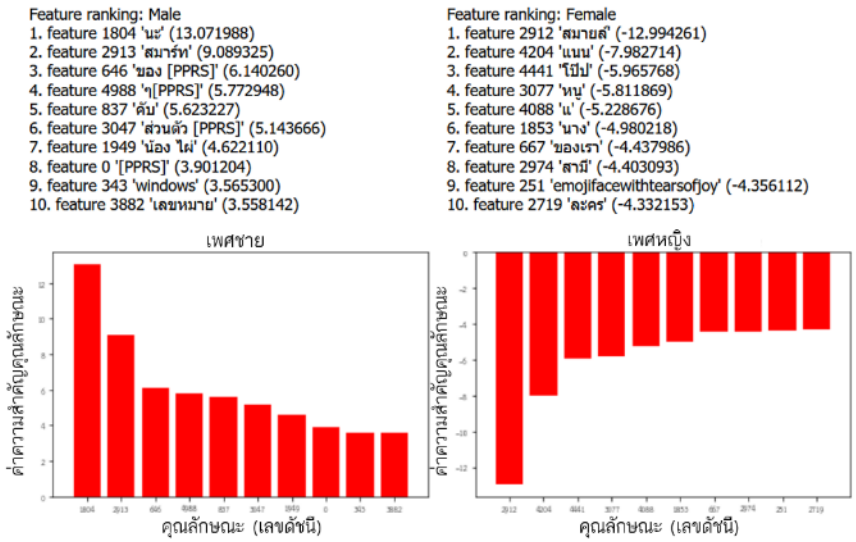
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	59.11%	69.81%	59.11%	45.62%
คำแสดงความลังเล	200	59.15%	59.40%	59.15%	49.08%
คำหยุด	1,030	63.04%	65.20%	63.04%	57.12%
ชนิดของคำ	1,239	59.06%	61.71%	59.06%	47.12%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	73.08%	73.04%	73.08%	72.48%
TF-IDF(คำศัพท์ใหม่)	5,000	73.03%	72.98%	73.03%	72.42%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,239	73.18%	73.12%	73.18%	72.61%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	73.74%	73.89%	73.74%	73.01%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,239	73.63%	73.72%	73.63%	72.95%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	73.28%	73.25%	73.28%	72.67%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,239	73.45%	73.41%	73.45%	72.88%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	73.91%	74.07%	73.91%	73.19%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,239	73.73%	73.81%	73.73%	73.06%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	73.09%	73.06%	73.09%	72.48%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,239	73.26%	73.21%	73.26%	72.68%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	73.61%	73.79%	73.61%	72.85%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,239	73.46%	73.53%	73.46%	72.78%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	73.32%	73.29%	73.32%	72.71%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,239	73.39%	73.34%	73.39%	72.82%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	73.80%	73.99%	73.80%	73.05%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,239	73.66%	73.74%	73.66%	72.98%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	3,943	73.62%	73.70%	73.62%	72.94%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,722	73.66%	73.74%	73.66%	72.98%

ตาราง 16 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Random Forest บนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 11 – 96 คำ

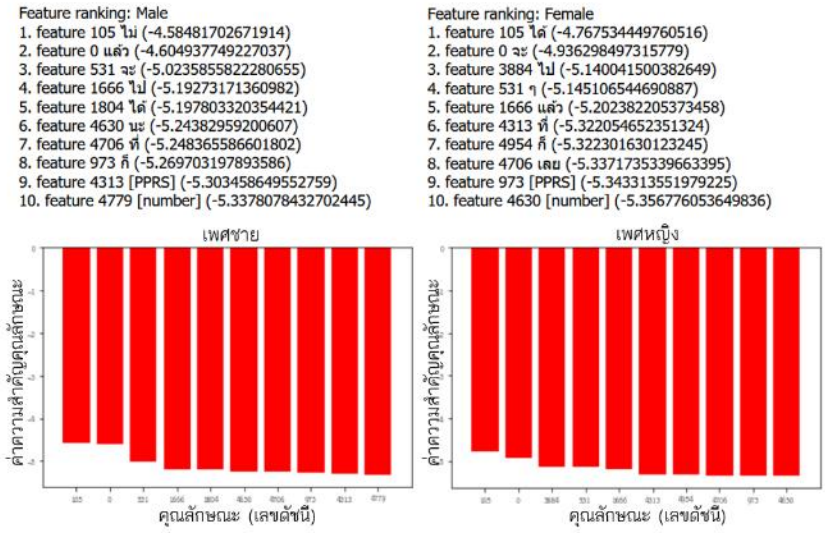
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	58.95%	70.30%	58.95%	45.17%
คำแสดงความลังเล	200	59.03%	61.17%	59.03%	47.26%
คำหยุด	1,030	59.25%	69.08%	59.25%	46.04%
ชนิดของคำ	1,239	58.48%	68.34%	58.48%	44.11%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	60.40%	71.57%	60.40%	48.48%
TF-IDF(คำศัพท์ใหม่)	5,000	60.26%	72.28%	60.26%	48.06%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,239	59.96%	72.27%	59.96%	47.39%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	61.38%	72.55%	61.38%	50.50%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,239	60.65%	72.54%	60.65%	48.90%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	60.63%	71.73%	60.63%	48.97%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,239	60.04%	72.16%	60.04%	47.59%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	61.14%	73.01%	61.14%	49.90%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,239	60.60%	72.95%	60.60%	48.74%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	60.40%	72.54%	60.40%	48.34%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,239	60.38%	72.35%	60.38%	48.32%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	61.04%	72.96%	61.04%	49.70%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,239	61.21%	72.60%	61.21%	50.13%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	60.42%	72.32%	60.42%	48.41%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,239	60.40%	72.33%	60.40%	48.38%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	61.41%	72.82%	61.41%	50.53%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,239	61.14%	73.33%	61.14%	49.86%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	2,958	61.42%	73.26%	61.42%	50.46%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	338	65.64%	71.62%	65.64%	59.55%

คุณลักษณะที่มีผลต่อการจำแนกแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

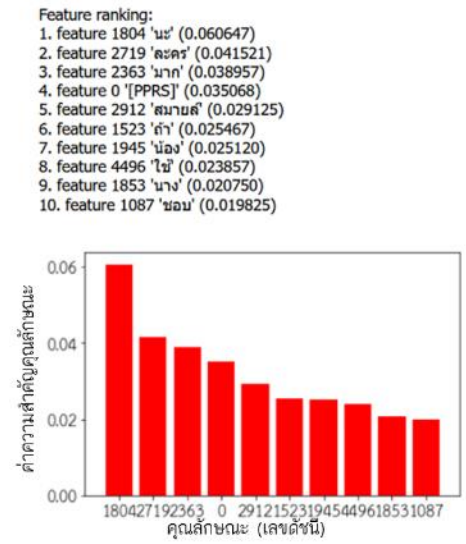
คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ 10 อันดับแรกจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดในแต่ละอัลกอริทึม ซึ่งแสดงดังภาพประกอบ 34 ภาพประกอบ 35 และภาพประกอบ 36



ภาพประกอบ 34 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Logistic Regression



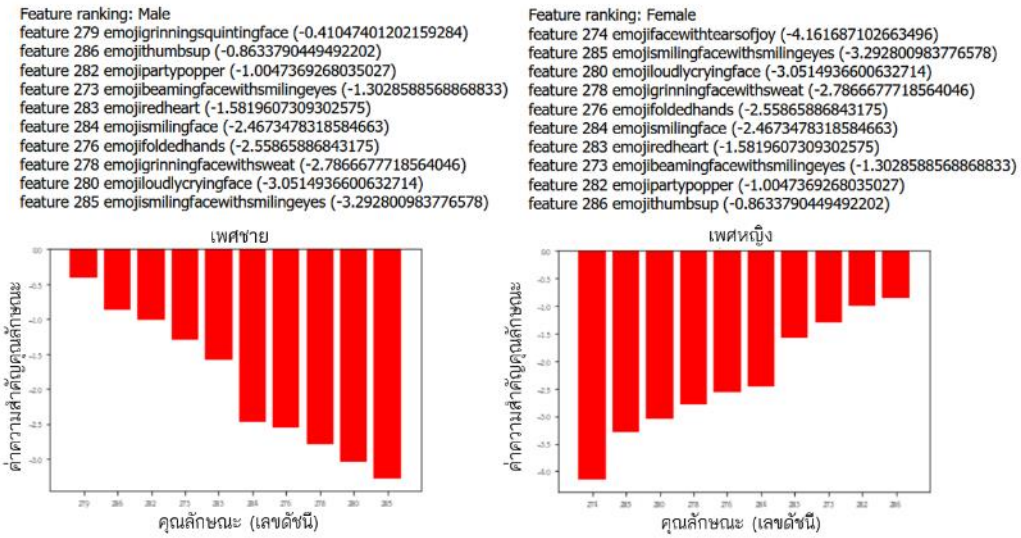
ภาพประกอบ 35 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes



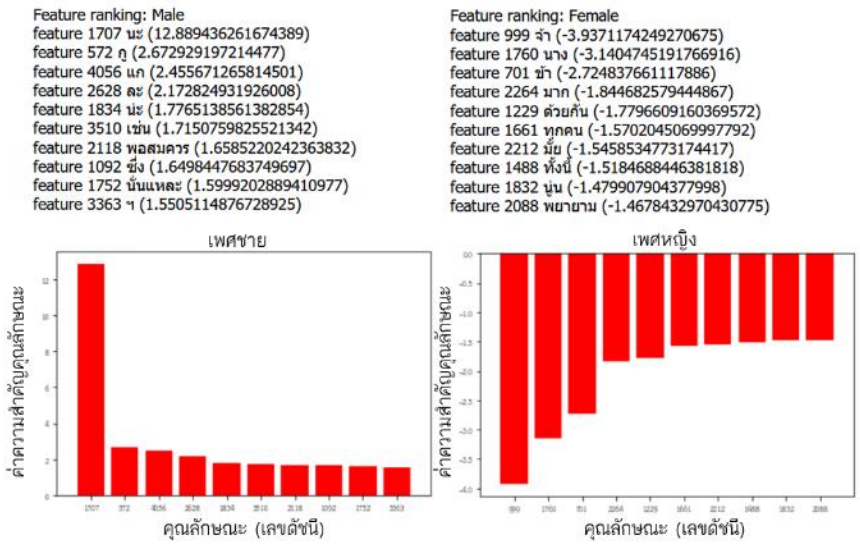
ภาพประกอบ 36 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Random Forest

คุณลักษณะที่มีผลต่อการจำแนกแยกตามประเภทคุณลักษณะ

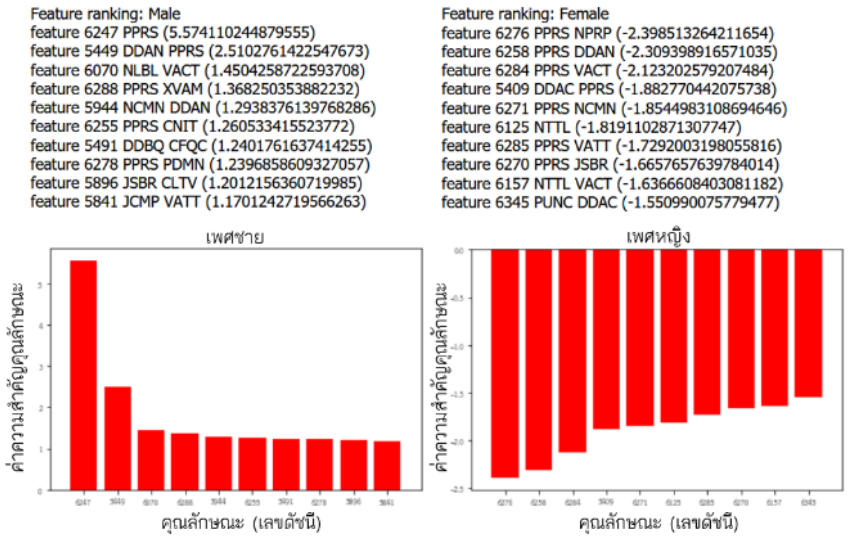
คุณลักษณะที่มีผลต่อการจำแนกเพศจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ10 อันดับแรก แยกตามประเภทคุณลักษณะ ซึ่งแสดงดังภาพประกอบ 37 ภาพประกอบ 38 และภาพประกอบ 39



ภาพประกอบ 37 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ จากคุณลักษณะ emoji 10 อันดับแรก



ภาพประกอบ 38 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ จากคุณลักษณะคำหยุด 10 อันดับแรก



ภาพประกอบ 39 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 11 – 96 คำ จากคุณลักษณะชนิดของคำ 10 อันดับแรก

ประสิทธิภาพของแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 97 - 200 คำ

ประสิทธิภาพของแบบจำลองแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

จากการสร้างแบบจำลองการจำแนกเพศด้วยอัลกอริทึม 3 อัลกอริทึม ได้แก่ Logistic Regression, Naïve Bayes และ Random Forest บนชุดข้อมูลสำหรับการสร้างแบบจำลองที่มีจำนวนคำระหว่าง 97 – 200 คำ และทำการทดสอบประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 97 – 200 คำ โดยใช้ค่า Weighted average ของ Accuracy, Precision, Recall และ F1 Score เป็นตัววัดประสิทธิภาพของแบบจำลอง ได้ผลลัพธ์ดังตาราง 17 ตาราง 18 และตาราง 19 จะเห็นได้ว่าค่าประสิทธิภาพของแบบจำลองที่ดีที่สุดมีค่า Accuracy 79.04%, Precision 79.03%, Recall 79.04% และ F1 Score 78.98% ซึ่งใช้คุณลักษณะ TF-IDF ที่สกัดจากคำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด และชนิดของคำ ร่วมกับอัลกอริทึม Logistic Regression และใช้อัลกอริทึม SelectKBest ในการเลือกคุณลักษณะที่ดีที่สุดในการสร้างแบบจำลอง

ตาราง 17 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Logistic Regression บนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 97 – 200 คำ

คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	55.90%	66.89%	55.90%	42.52%
คำแสดงความลังเล	200	59.42%	59.22%	59.42%	57.77%
คำหยุด	1,030	68.69%	68.61%	68.69%	68.62%
ชนิดของคำ	1,615	61.63%	61.42%	61.63%	60.89%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	77.52%	77.50%	77.52%	77.45%
TF-IDF(คำศัพท์ใหม่)	5,000	77.75%	77.73%	77.75%	77.67%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,615	78.51%	78.48%	78.51%	78.46%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	78.73%	78.71%	78.73%	78.67%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,615	78.92%	78.89%	78.92%	78.86%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	77.81%	77.79%	77.81%	77.74%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,615	78.19%	78.16%	78.19%	78.14%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	78.89%	78.87%	78.89%	78.82%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,615	78.95%	78.93%	78.95%	78.90%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	77.75%	77.72%	77.75%	77.68%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,615	78.38%	78.36%	78.38%	78.32%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	78.51%	78.49%	78.51%	78.44%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,615	78.76%	78.73%	78.76%	78.71%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	77.62%	77.60%	77.62%	77.55%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,615	78.19%	78.17%	78.19%	78.13%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	78.38%	78.36%	78.38%	78.31%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,615	78.89%	78.86%	78.89%	78.84%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	**4,206**	**79.04%**	**79.03%**	**79.04%**	**78.98%**
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	2,985	78.98%	78.96%	78.98%	78.92%

ตาราง 18 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes บนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 97 – 200 คำ

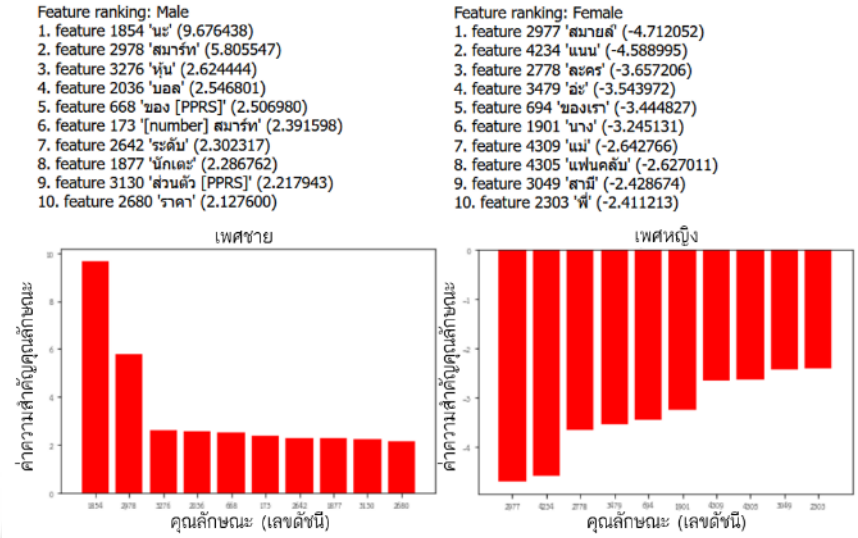
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	55.90%	67.59%	55.90%	42.42%
คำแสดงความลังเล	200	58.88%	60.58%	58.88%	53.65%
คำหยุด	1,030	64.86%	66.09%	64.86%	62.91%
ชนิดของคำ	1,615	56.22%	63.78%	56.22%	43.99%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	75.82%	75.77%	75.82%	75.77%
TF-IDF(คำศัพท์ใหม่)	5,000	75.78%	75.74%	75.78%	75.74%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,615	75.59%	75.56%	75.59%	75.51%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	75.69%	75.68%	75.69%	75.57%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,615	75.34%	75.35%	75.34%	75.20%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	75.85%	75.81%	75.85%	75.80%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,615	75.56%	75.52%	75.56%	75.48%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	**5,000**	**75.88%**	**75.87%**	**75.88%**	**75.77%**
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,615	75.47%	75.48%	75.47%	75.33%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	75.78%	75.74%	75.78%	75.74%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,615	75.25%	75.20%	75.25%	75.17%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	75.37%	75.36%	75.37%	75.25%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,615	74.99%	75.02%	74.99%	74.83%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	75.85%	75.81%	75.85%	75.80%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,615	75.44%	75.40%	75.44%	75.36%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	75.50%	75.49%	75.50%	75.38%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,615	75.18%	75.21%	75.18%	75.02%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	3,685	75.12%	75.15%	75.12%	74.96%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,615	75.18%	75.21%	75.18%	75.02%

ตาราง 19 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Random Forest บนชุดข้อมูลทดสอบที่มีจำนวนคำระหว่าง 97 – 200 คำ

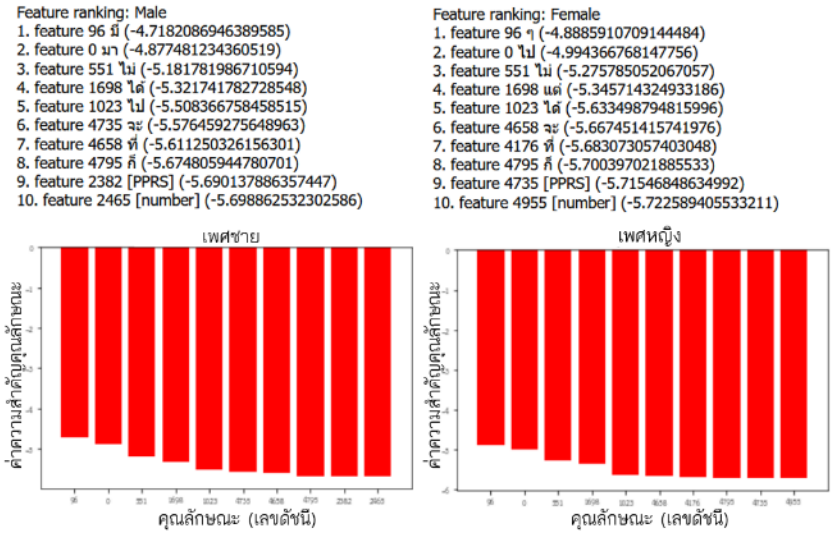
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	55.59%	66.73%	55.59%	41.71%
คำแสดงความลังเล	200	58.63%	61.75%	58.63%	51.81%
คำหยุด	1,030	65.18%	69.55%	65.18%	61.46%
ชนิดของคำ	1,615	59.16%	63.23%	59.16%	52.20%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	69.07%	73.64%	69.07%	66.38%
TF-IDF(คำศัพท์ใหม่)	5,000	68.95%	73.46%	68.95%	66.24%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,615	69.20%	74.57%	69.20%	66.26%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	70.43%	75.03%	70.43%	68.01%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,615	70.43%	74.92%	70.43%	68.05%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	69.01%	73.55%	69.01%	66.31%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,615	68.44%	73.06%	68.44%	65.57%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	69.39%	73.81%	69.39%	66.81%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,615	70.15%	74.86%	70.15%	67.64%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	69.77%	74.49%	69.77%	67.17%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,615	69.17%	74.16%	69.17%	66.35%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	70.85%	75.54%	70.85%	68.48%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,615	69.23%	73.96%	69.23%	66.51%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	69.86%	74.65%	69.86%	67.27%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,615	69.33%	74.32%	69.33%	66.54%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	70.72%	75.13%	70.72%	68.41%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,615	69.61%	74.29%	69.61%	66.99%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	4,300	70.18%	74.46%	70.18%	67.82%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	**345**	**72.55%**	**74.60%**	**72.55%**	**71.34%**

คุณลักษณะที่มีผลต่อการจำแนกแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

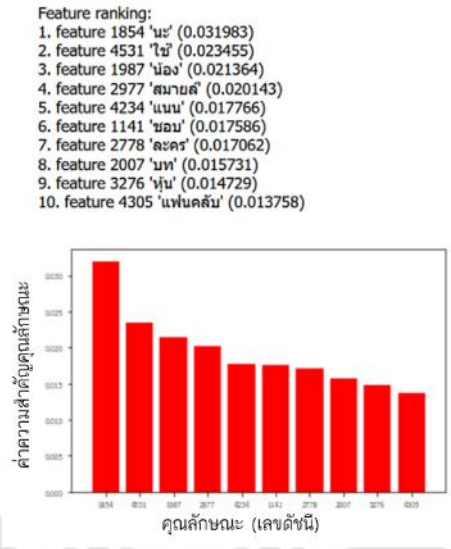
คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ 10 อันดับแรกจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดในแต่ละอัลกอริทึม ซึ่งแสดงดังภาพประกอบ 40 ภาพประกอบ 41 และภาพประกอบ 42



ภาพประกอบ 40 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Logistic Regression



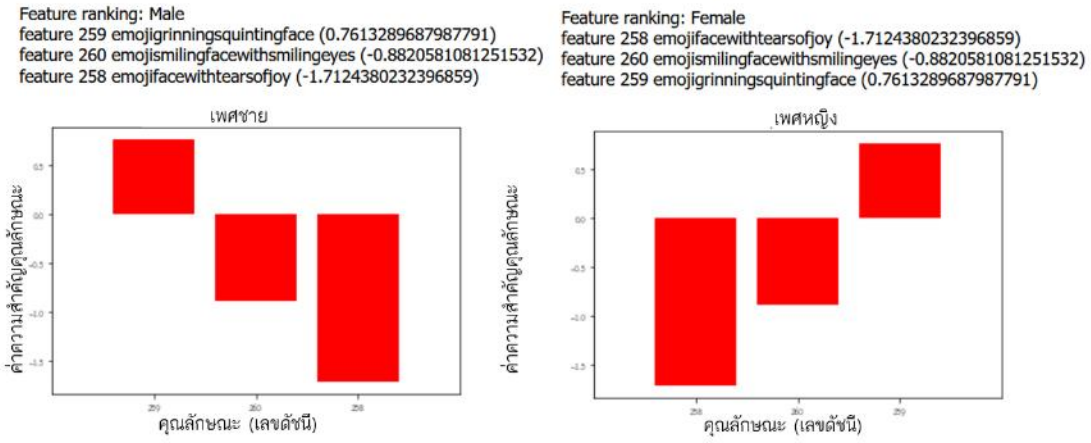
ภาพประกอบ 41 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes



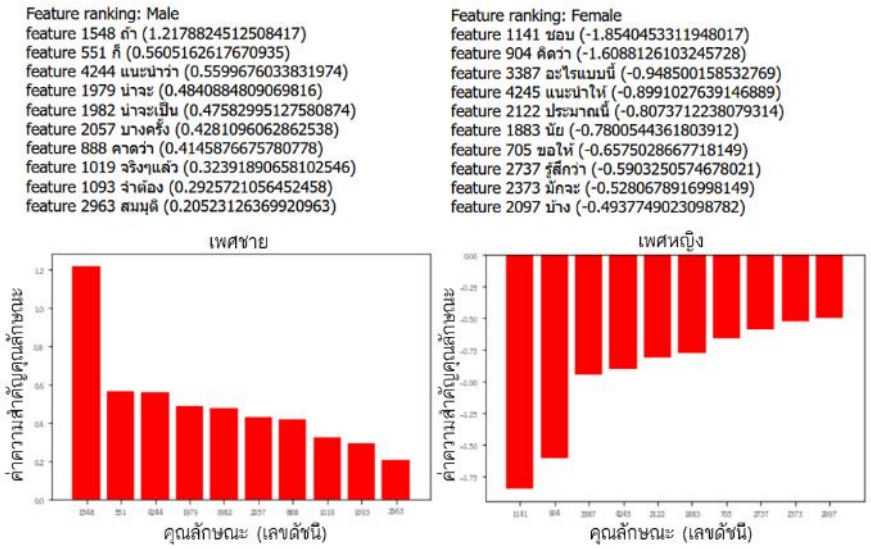
ภาพประกอบ 42 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Random Forest

คุณลักษณะที่มีผลต่อการจำแนกแยกตามประเภทคุณลักษณะ

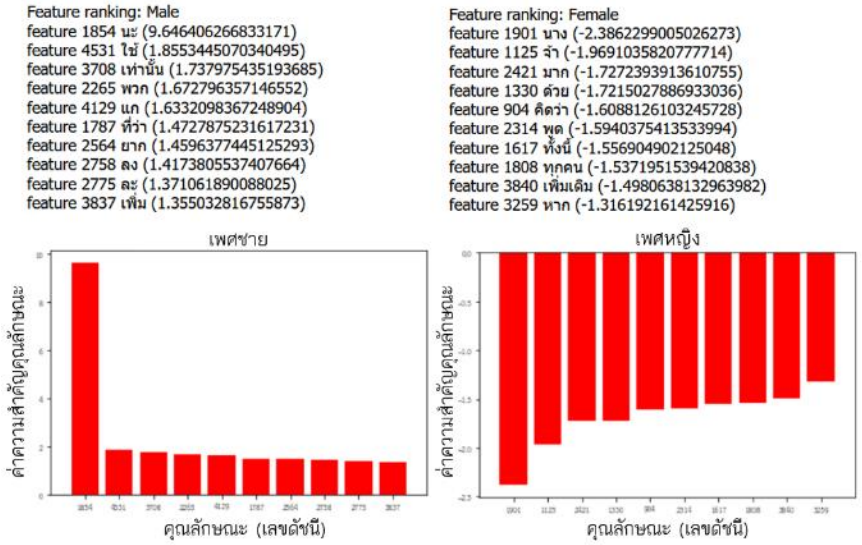
คุณลักษณะที่มีผลต่อการจำแนกเพศจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ 10 อันดับแรก แยกตามประเภทคุณลักษณะ ซึ่งแสดงดังภาพประกอบ 43 ภาพประกอบ 44 ภาพประกอบ 45 และภาพประกอบ 46 แต่สำหรับ emoji จะมีเพียง 3 อันดับเท่านั้น เนื่องจากแบบจำลองนี้มี emoji เพียง 3 คุณลักษณะเท่านั้นจากคุณลักษณะทั้งหมด



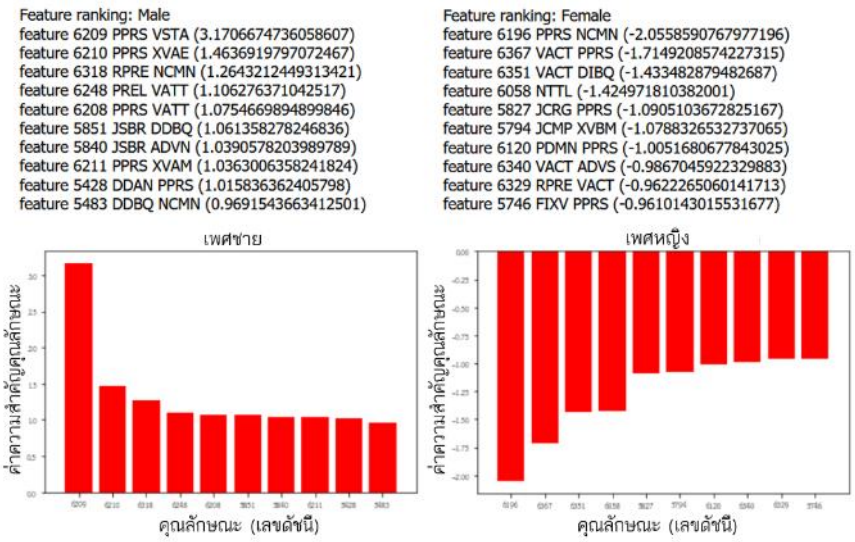
ภาพประกอบ 43 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ จากคุณลักษณะ emoji 3 อันดับแรก



ภาพประกอบ 44 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ จากคุณลักษณะคำแสดงความลังเล 10 อันดับแรก



ภาพประกอบ 45 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ จากคุณลักษณะคำหยุด 10 อันดับแรก



ภาพประกอบ 46 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำระหว่าง 97 – 200 คำ จากคุณลักษณะชนิดของคำ 10 อันดับแรก

ประสิทธิภาพของแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำมากกว่า 200 คำ

ประสิทธิภาพของแบบจำลองแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

จากการสร้างแบบจำลองการจำแนกเพศด้วยอัลกอริทึม 3 อัลกอริทึม ได้แก่ Logistic Regression, Naïve Bayes และ Random Forest บนชุดข้อมูลสำหรับการสร้างแบบจำลองที่มีจำนวนคำมากกว่า 200 คำ และทำการทดสอบประสิทธิภาพแบบจำลองบนชุดข้อมูลทดสอบที่มีจำนวนคำมากกว่า 200 คำ โดยใช้ค่า Weighted average ของ Accuracy, Precision, Recall และ F1 Score เป็นตัววัดประสิทธิภาพของแบบจำลอง ได้ผลลัพธ์ดังตาราง 20 ตาราง 21 และตาราง 22 จะเห็นได้ว่าค่าประสิทธิภาพของแบบจำลองที่ดีที่สุดมีค่า Accuracy 79.03%, Precision 79.11%, Recall 79.03% และ F1 Score 78.89% ซึ่งใช้คุณลักษณะ TF-IDF ที่สกัดจากคำศัพท์ใหม่, คำแสดงความลังเล, คำหยุด และชนิดของคำ ร่วมกับอัลกอริทึม Logistic Regression ในการสร้างแบบจำลอง

ตาราง 20 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Logistic Regression บนชุดข้อมูลทดสอบที่มีจำนวนคำมากกว่า 200 คำ

คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	56.61%	63.67%	56.61%	44.69%
คำแสดงความลังเล	200	62.66%	62.69%	62.66%	61.41%
คำหยุด	1,030	72.29%	72.34%	72.29%	72.01%
ชนิดของคำ	1,606	62.57%	63.34%	62.57%	60.29%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	76.04%	76.09%	76.04%	75.87%
TF-IDF(คำศัพท์ใหม่)	5,000	76.21%	76.24%	76.21%	76.06%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,606	77.15%	77.17%	77.15%	77.02%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	77.66%	77.71%	77.66%	77.52%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,606	78.52%	78.57%	78.52%	78.39%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	76.13%	76.16%	76.13%	75.97%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,606	77.32%	77.35%	77.32%	77.19%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	77.83%	77.88%	77.83%	77.70%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,606	78.43%	78.48%	78.43%	78.30%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	76.21%	76.25%	76.21%	76.05%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,606	76.56%	76.58%	76.56%	76.41%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	78.35%	78.41%	78.35%	78.21%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,606	79.03%	79.11%	79.03%	78.89%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	75.96%	76.00%	75.96%	75.79%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,606	76.81%	76.85%	76.81%	76.66%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	78.35%	78.41%	78.35%	78.21%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,606	78.94%	79.03%	78.94%	78.80%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	5,693	78.52%	78.63%	78.52%	78.35%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	2,316	78.52%	78.62%	78.52%	78.36%

ตาราง 21 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes บนชุดข้อมูลทดสอบที่มีจำนวนคำมากกว่า 200 คำ

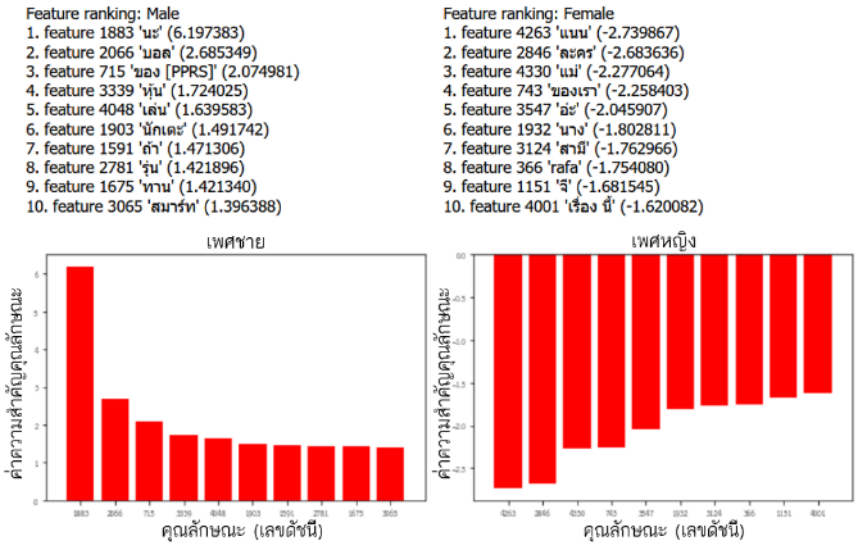
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	56.78%	65.05%	56.78%	44.79%
คำแสดงความลังเล	200	58.40%	59.78%	58.40%	52.70%
คำหยุด	1,030	66.75%	70.01%	66.75%	63.93%
ชนิดของคำ	1,606	55.84%	69.12%	55.84%	41.74%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	72.98%	72.98%	72.98%	72.76%
TF-IDF(คำศัพท์ใหม่)	5,000	73.32%	73.32%	73.32%	73.10%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,606	73.57%	73.72%	73.57%	73.26%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	74.51%	74.66%	74.51%	74.22%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,606	74.08%	74.40%	74.08%	73.68%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	73.23%	73.24%	73.23%	73.01%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,606	73.49%	73.64%	73.49%	73.16%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	74.42%	74.57%	74.42%	74.14%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,606	74.00%	74.32%	74.00%	73.59%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	73.32%	73.31%	73.32%	73.12%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,606	73.74%	73.93%	73.74%	73.41%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	74.17%	74.33%	74.17%	73.86%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,606	74.25%	74.60%	74.25%	73.84%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	73.15%	73.14%	73.15%	72.94%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,606	73.83%	74.01%	73.83%	73.50%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	74.25%	74.41%	74.25%	73.96%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,606	74.34%	74.66%	74.34%	73.95%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	4,494	74.34%	74.64%	74.34%	73.96%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	5,946	74.25%	74.54%	74.25%	73.88%

ตาราง 22 ผลลัพธ์ของประสิทธิภาพแบบจำลองที่ใช้อัลกอริทึม Random Forest บนชุดข้อมูลทดสอบที่มีจำนวนคำมากกว่า 200 คำ

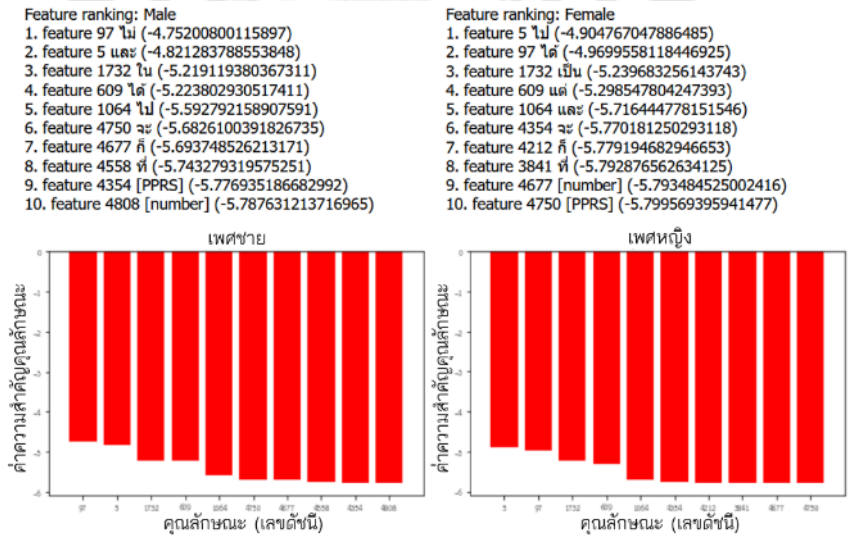
คุณลักษณะ	จำนวนคุณลักษณะ	Accuracy	Precision	Recall	F-1
emoji	789	56.69%	65.28%	56.69%	44.51%
คำแสดงความลังเล	200	60.78%	63.30%	60.78%	55.95%
คำหยุด	1,030	68.03%	71.07%	68.03%	65.63%
ชนิดของคำ	1,606	60.19%	63.25%	60.19%	54.51%
TF-IDF(คำศัพท์เดิม สำหรับ baseline)	5,000	67.77%	70.27%	67.77%	65.57%
TF-IDF(คำศัพท์ใหม่)	5,000	66.67%	69.26%	66.67%	64.16%
TF-IDF(คำศัพท์ใหม่),ชนิดของคำ	6,606	66.92%	69.71%	66.92%	64.38%
TF-IDF(คำศัพท์ใหม่),คำหยุด	5,000	68.71%	71.30%	68.71%	66.65%
TF-IDF(คำศัพท์ใหม่),คำหยุด,ชนิดของคำ	6,606	68.88%	71.88%	68.88%	66.68%
TF-IDF(คำศัพท์ใหม่),emoji	5,000	66.84%	69.93%	66.84%	64.12%
TF-IDF(คำศัพท์ใหม่),emoji,ชนิดของคำ	6,606	67.09%	70.01%	67.09%	64.53%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด	5,000	68.46%	71.15%	68.46%	66.30%
TF-IDF(คำศัพท์ใหม่),emoji,คำหยุด,ชนิดของคำ	6,606	68.03%	71.22%	68.03%	65.56%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล	5,000	67.60%	70.32%	67.60%	65.25%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,ชนิดของคำ	6,606	67.26%	69.95%	67.26%	64.85%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด	5,000	68.71%	71.30%	68.71%	66.65%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,คำหยุด,ชนิดของคำ	6,606	68.03%	71.14%	68.03%	65.59%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji	5,000	66.58%	69.25%	66.58%	64.01%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,ชนิดของคำ	6,606	65.98%	69.06%	65.98%	63.04%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด	5,000	68.97%	71.88%	68.97%	66.82%
TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	6,606	68.20%	71.61%	68.20%	65.67%
(SelectKBest)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	3,245	69.48%	72.31%	69.48%	67.46%
(RFECV)TF-IDF(คำศัพท์ใหม่),คำแสดงความลังเล,emoji,คำหยุด,ชนิดของคำ	**336**	**71.53%**	**72.85%**	**71.53%**	**70.45%**

คุณลักษณะที่มีผลต่อการจำแนกแยกตามอัลกอริทึมที่ใช้สร้างแบบจำลอง

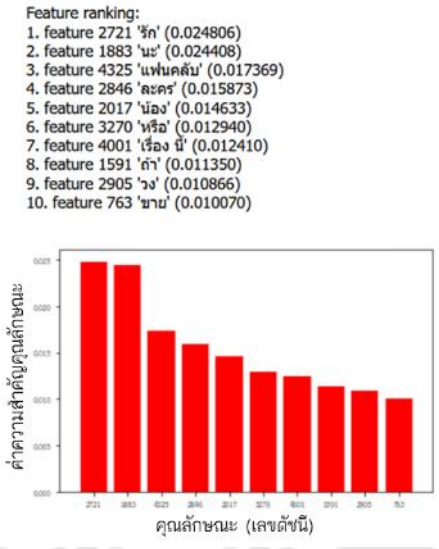
คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ 10 อันดับแรกจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดในแต่ละอัลกอริทึม ซึ่งแสดงดังภาพประกอบ 47 ภาพประกอบ 48 และภาพประกอบ 49



ภาพประกอบ 47 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Logistic Regression



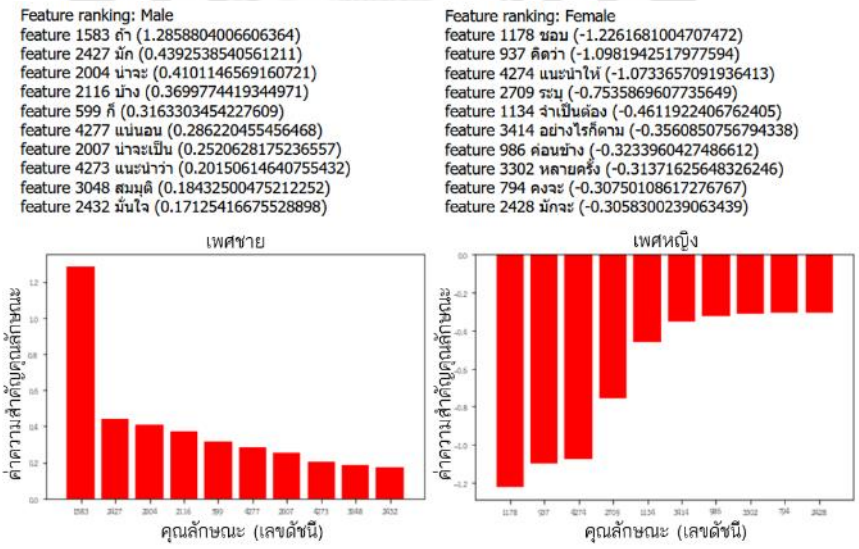
ภาพประกอบ 48 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Naïve Bayes



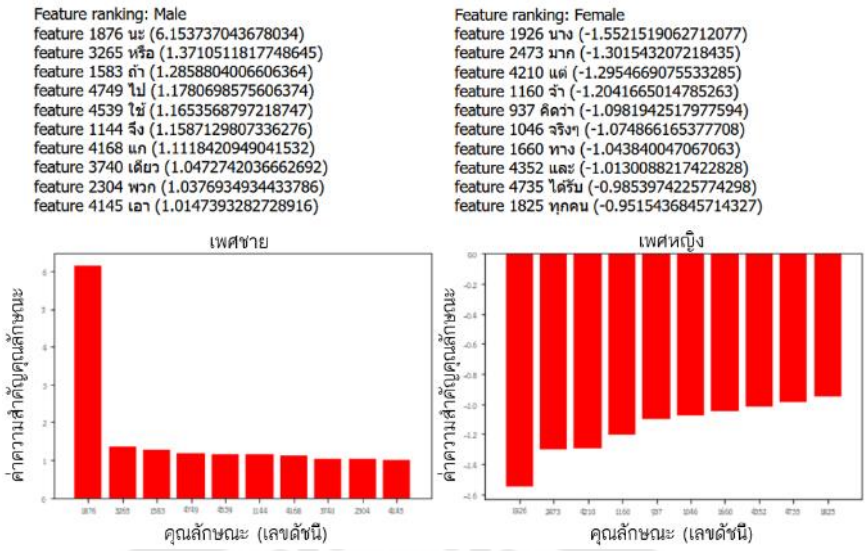
ภาพประกอบ 49 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ 10 อันดับแรกจากแบบจำลองที่ใช้อัลกอริทึม Random Forest

คุณลักษณะที่มีผลต่อการจำแนกแยกตามประเภทคุณลักษณะ

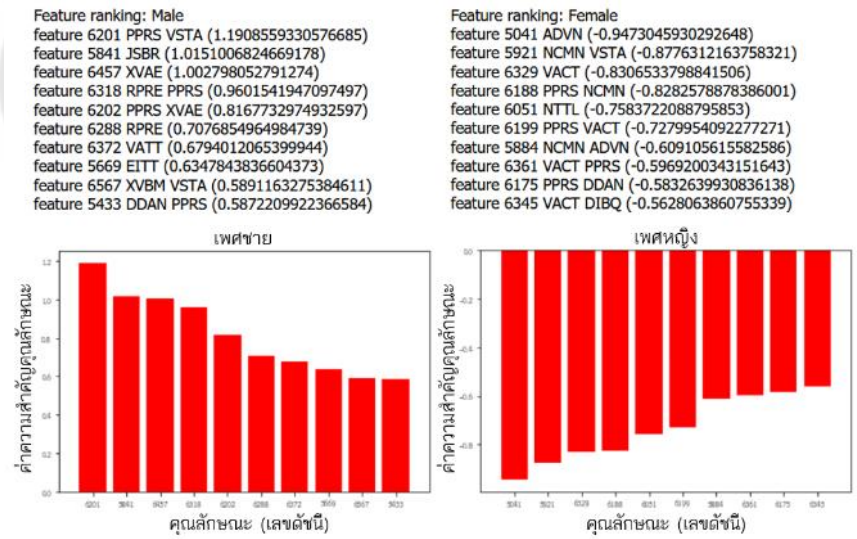
คุณลักษณะที่มีผลต่อการจำแนกเพศจากแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ 10 อันดับแรก แยกตามประเภทคุณลักษณะ ซึ่งแสดงดังภาพประกอบ 50 ภาพประกอบ 51 และภาพประกอบ 52



ภาพประกอบ 50 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ จากคุณลักษณะคำแสดงความลังเล 10 อันดับแรก



ภาพประกอบ 51 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ จากคุณลักษณะคำหยุด 10 อันดับแรก



ภาพประกอบ 52 คุณลักษณะที่มีผลต่อการจำแนกเพศบนชุดข้อมูลที่มีจำนวนคำมากกว่า 200 คำ จากคุณลักษณะชนิดของคำ 10 อันดับแรก

บทที่ 5
สรุปผลการวิจัย อภิปรายผล และข้อเสนอแนะ

ในการวิจัยการจำแนกเพศจากข้อความบนโซเชียลเน็ตเวิร์ก ซึ่งผู้วิจัยได้มีการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติในการสกัดคุณลักษณะ และทำการสร้างแบบจำลองการจำแนกเพศโดยใช้เทคนิคการเรียนรู้ของเครื่อง ผู้วิจัยได้วัดประสิทธิภาพของแบบจำลองแต่ละอัลกอริทึมเพื่อนำมาเปรียบเทียบและสรุปผล โดยสามารถแบ่งหัวข้อในการสรุปผลได้ดังต่อไปนี้

1. สรุปผลการวิจัย
2. อภิปรายผลการวิจัย
3. ข้อจำกัดและสิ่งที่ต้องพัฒนาในงานวิจัยนี้
4. ข้อเสนอแนะ

สรุปผลการวิจัย

ในปัจจุบันการเก็บข้อมูลจากช่องทางต่างๆ บนอินเตอร์เน็ต โดยเฉพาะโซเชียลเน็ตเวิร์กนั้นกำลังเป็นที่แพร่หลาย โดยเฉพาะอย่างยิ่งการเก็บข้อมูลเพื่อใช้สำหรับวิจัยทางด้านการตลาด ซึ่งจะทำให้สามารถเข้าใจผู้บริโภคได้มากยิ่งขึ้น บริษัทใดที่สามารถเก็บข้อมูลผู้บริโภคได้มากจะทำให้มีความสามารถในการแข่งขันในตลาดได้มากกว่าบริษัทที่เก็บข้อมูลได้น้อย ซึ่งข้อมูลผู้บริโภคนั้นก็มีหลากหลายปัจจัยที่ส่งผลให้เกิดพฤติกรรมการบริโภคที่แตกต่างกันในแต่ละบุคคล หนึ่งในนั้นคือเพศ ซึ่งเป็นปัจจัยที่สำคัญปัจจัยหนึ่งที่ส่งผลโดยตรงต่อพฤติกรรมของผู้บริโภค สำหรับการเก็บข้อมูลเพศของผู้บริโภคในไทยนั้น ส่วนใหญ่มักใช้วิธีการระบุเพศจากข้อความต่างๆ ที่ผู้บริโภคเผยแพร่โดยใช้วิธีการระบุจากคำลงท้าย เช่น ครับ ค่ะ หรือคำสรรพนามแทนตัว เช่น ผม ดิฉัน ในการระบุเพศ ซึ่งผู้วิจัยพบว่ามีข้อความเพียง 30% เท่านั้นที่มีคำที่สามารถระบุเพศได้ ซึ่งถ้าหากสามารถระบุเพศจากข้อความในส่วนที่ไม่มีคำเหล่านี้อีก 70% ได้ จะทำให้สามารถนำข้อมูลที่ได้ไปใช้ได้อย่างถูกต้องและมีประสิทธิภาพ ทำให้เกิดความได้เปรียบในด้านการตลาด

ในงานวิจัยนี้ผู้วิจัยได้ศึกษาวิธีการจำแนกเพศของผู้เขียนข้อความแสดงความคิดเห็นบนโซเชียลเน็ตเวิร์ก โดยการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติในการสกัดคุณลักษณะจากข้อความแสดงความคิดเห็น ร่วมกับการสร้างแบบจำลองสำหรับการจำแนก และเลือกใช้อัลกอริทึมการเรียนรู้ของเครื่อง 3 อัลกอริทึม ได้แก่ Logistic Regression, Naive Bayes และ Random Forest โดยได้แบ่งชุดข้อมูลที่ใช้ศึกษาออกเป็น 4 ชุด ตามจำนวนคำ โดยมีค่า accuracy ของแต่ละชุดข้อมูล โดยแบ่งตามอัลกอริทึมและกลุ่มคุณลักษณะที่ใช้ ดังตาราง 23

ตาราง 23 เปรียบเทียบค่า accuracy ในแต่ละชุดข้อมูล โดยแบ่งตามอัลกอริทึมและกลุ่มคุณลักษณะที่ใช้

คุณลักษณะ (คำย่อ)	ชุดข้อมูลที่มีจำนวนคำ <= 10			ชุดข้อมูลที่มีจำนวนคำ 11-96			ชุดข้อมูลที่มีจำนวนคำ 97-200			ชุดข้อมูลที่มีจำนวนคำ > 200		
	LR	NB	RF	LR	NB	RF	LR	NB	RF	LR	NB	RF
Emoji (E)	59.54%	59.42%	58.88%	59.34%	59.11%	58.95%	55.90%	55.90%	55.59%	56.61%	56.78%	56.69%
คำแสดงความลังเล (H)	58.38%	58.32%	58.50%	59.20%	59.15%	59.03%	59.42%	58.88%	58.63%	62.66%	58.40%	60.78%
คำหยุด (S)	59.79%	59.87%	58.57%	66.12%	63.04%	59.25%	68.69%	64.86%	65.18%	72.29%	66.75%	68.03%
ประเภทของคำ (T)	59.25%	58.97%	58.49%	60.66%	59.06%	58.48%	61.63%	56.22%	59.16%	62.57%	55.84%	60.19%
TF-IDF(คำศัพท์เดิม) (TFIDF1)	65.64%	65.60%	58.88%	74.50%	73.08%	60.40%	77.52%	75.82%	69.07%	76.04%	72.98%	67.77%
TF-IDF(คำศัพท์ใหม่) (TFIDF2)	65.66%	65.63%	58.56%	74.42%	73.03%	60.26%	77.75%	75.78%	68.95%	76.21%	73.32%	66.67%
TFIDF2, T	65.66%	65.47%	58.63%	74.80%	73.18%	59.96%	78.51%	75.59%	69.20%	77.15%	73.57%	66.92%
TFIDF2, S	66.81%	66.61%	58.67%	76.34%	73.74%	61.38%	78.73%	75.69%	70.43%	77.66%	74.51%	68.71%
TFIDF2, T, S	66.85%	66.43%	58.72%	76.49%	73.63%	60.65%	78.92%	75.34%	70.43%	78.52%	74.08%	68.88%
TFIDF2, E	66.10%	66.07%	58.58%	74.68%	73.28%	60.63%	77.81%	75.85%	69.01%	76.13%	73.23%	66.84%
TFIDF2, E, T	66.16%	65.93%	58.50%	75.02%	73.45%	60.04%	78.19%	75.56%	68.44%	77.32%	73.49%	67.09%
TFIDF2, E, S	67.13%	66.78%	58.71%	76.47%	73.91%	61.14%	78.89%	75.88%	69.39%	77.83%	74.42%	68.46%
TFIDF2, E, S, T	67.13%	66.83%	58.68%	76.62%	73.73%	60.60%	78.95%	75.47%	70.15%	78.43%	74.00%	68.03%
TFIDF2, H	65.62%	65.58%	58.60%	74.51%	73.09%	60.40%	77.75%	75.78%	69.77%	76.21%	73.32%	67.60%
TFIDF2, H ,T	65.65%	65.49%	58.59%	74.74%	73.26%	60.38%	78.38%	75.25%	69.17%	76.56%	73.74%	67.26%
TFIDF2, H, S	66.94%	66.74%	58.76%	76.27%	73.61%	61.04%	78.51%	75.37%	70.85%	78.35%	74.17%	68.71%
TFIDF2, H, S, T	66.95%	66.49%	58.61%	76.41%	73.46%	61.21%	78.76%	74.99%	69.23%	79.03%	74.25%	68.03%
TFIDF2, H, E	65.95%	65.79%	58.58%	74.69%	73.32%	60.42%	77.62%	75.85%	69.86%	75.96%	73.15%	66.58%
TFIDF2, H, E, T	65.96%	65.84%	58.56%	74.90%	73.39%	60.40%	78.19%	75.44%	69.33%	76.81%	73.83%	65.98%

ตาราง 23 (ต่อ)

คุณลักษณะ (คำย่อ)	ชุดข้อมูลที่มีจำนวนคำ <= 10			ชุดข้อมูลที่มีจำนวนคำ 11-96			ชุดข้อมูลที่มีจำนวนคำ 97-200			ชุดข้อมูลที่มีจำนวนคำ > 200		
	LR	NB	RF	LR	NB	RF	LR	NB	RF	LR	NB	RF
TFIDF2, H, E, S	67.08%	66.94%	58.92%	76.41%	73.80%	61.41%	78.38%	75.50%	70.72%	78.35%	74.25%	68.97%
TFIDF2, H, E, S, T	67.02%	66.79%	58.61%	76.55%	73.66%	61.14%	78.89%	75.18%	69.61%	78.94%	74.34%	68.20%
(SelectKBest) TFIDF2, H, E, S, T	67.23%	66.83%	58.84%	76.40%	73.62%	61.42%	79.04%	75.12%	70.18%	78.52%	74.34%	69.48%
(RFECV) TFIDF2, H, E, S, T	67.09%	66.79%	60.96%	76.57%	73.66%	65.64%	78.98%	75.18%	72.55%	78.52%	74.25%	71.53%

จากผลการวิจัยในตาราง 23 สรุปได้ว่าแบบจำลองที่ให้ค่าประสิทธิภาพการจำแนกที่ดีที่สุดคือ แบบจำลองที่ใช้อัลกอริทึม Logistic Regression โดยมีการใช้คุณลักษณะ TF-IDF ที่สกัดจากคำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด และชนิดของคำ และใช้อัลกอริทึม SelectKBest ในการเลือกคุณลักษณะที่ดีที่สุดในการสร้างแบบจำลอง ซึ่งมีค่า Accuracy 79.04% บนชุดข้อมูลที่มีความยาวระหว่าง 97-200 คำ โดยมีค่า accuracy มากกว่าแบบจำลองที่ใช้คุณลักษณะที่สกัดจากพจนานุกรมจากงานวิจัย (Horsuwan และคนอื่น ๆ, 2019) อยู่ 1.52% บนชุดข้อมูลและอัลกอริทึมเดียวกัน

ส่วนคุณลักษณะที่มีผลต่อการจำแนกที่ใช้ในแบบจำลองนี้ 10 อันดับแรก สำหรับเพศชาย ได้แก่ “นะ” “สมาร์ท” “หุ้น” “บอล” “ของ [PPRS]” “[number] สมาร์ท” “ระดับ” “นักเตะ” “ส่วนตัว [PPRS]” และ “ราคา” สำหรับเพศหญิง ได้แก่ “สมายล์” “แนน” “ละคร” “อ่ะ” “ของเรา” “นาง” “แม่” “แฟนคลับ” “สามี” และ “พี่” ซึ่งแสดงดังภาพประกอบ 40

เมื่อพิจารณาแยกตามประเภทคุณลักษณะ โดยพิจารณาเฉพาะคุณลักษณะที่ใช้ในแบบจำลองนี้ จะพบว่าคุณลักษณะ emoji มีเพียง 3 คุณลักษณะ จากทั้งหมด 4,206 คุณลักษณะที่ใช้ในแบบจำลองนี้ แสดงดังภาพประกอบ 43 ได้แก่ emojigrinningsquintingface 😆 emojismilingfacewithsmilingeyes 😊 emojifacewithtearsofjoy 😂

ส่วน 10 อันดับแรกของคุณลักษณะคำแสดงความลังเลสำหรับเพศชาย ได้แก่ “ถ้า” “ก็” “แนะนำว่า” “น่าจะ” “น่าจะเป็น” “บางครั้ง” “คาดว่า” “จริงๆ แล้ว” “จำต้อง” “สมมุติ” สำหรับเพศหญิง ได้แก่ “ชอบ” “คิดว่า” “อะไรแบบนี้” “แนะนำให้” “ประมาณนี้” “นัย” “ขอให้” “รู้สึกว่า” “มักจะ” “บ้าง” ดังภาพประกอบ 44

ในส่วนของ 10 อันดับแรกของคุณลักษณะคำหยุดสำหรับเพศชาย ได้แก่ “นะ” “ใช้” “เท่านั้น” “พวก” “แก” “ที่ว่า” “ยาง” “ลง” “ละ” “เพิ่ม” สำหรับเพศหญิง ได้แก่ “นาง” “จ้า” “มาก” “ด้วย” “คิดว่า” “พูด” “ทั้งนี้” “ทุกคน” “เพิ่มเติม” “หาก” ดังภาพประกอบ 45

และ 10 อันดับแรกของคุณลักษณะชนิดของคำสำหรับเพศชาย ได้แก่ “PPRS VSTA” “PPRS XVAE” “RPRE NCMN” “PREL VATT” “PPRS VATT” “JSBR DDBQ” “JSBR ADVN” “PPRS XVAM” “DDAN PPRS” “DDBQ NCMN” สำหรับเพศหญิง ได้แก่ “PPRS NCMN” “VACT PPRS” “VACT DIBQ” “NTTL” “JCRG PPRS” “JCMP XVBM” “PDMN PPRS” “VACT ADVS” “RPRE VACT” “FIXV PPRS” ดังภาพประกอบ 46

อภิปรายผลการวิจัย

เมื่อเปรียบเทียบค่าประสิทธิภาพของงานวิจัยนี้ กับงานวิจัยอื่นๆ ที่ทำการศึกษา ดังตาราง 24 พบว่าค่า accuracy สูงสุดของแบบจำลอง มีค่ามากกว่าค่า accuracy ของแบบจำลองที่ใช้คุณลักษณะที่สกัดจากพจนานุกรมจาก (Horsuwan et al., 2019) อยู่ 1.52% บนชุดข้อมูลและอัลกอริทึมเดียวกัน และเมื่อเปรียบเทียบกับ state-of-the-art ในภาษาอังกฤษ (Mukherjee & Liu, 2010) จะได้ค่าประสิทธิภาพของแบบจำลองน้อยกว่า 9.52% อาจเป็นเพราะข้อความที่นำมาใช้ในงานวิจัยนี้ ยังมีการระบุเพศของผู้เขียนข้อความได้ไม่ถูกต้องเท่าที่ควร ในขณะที่งานวิจัย state-of-the-art (Mukherjee & Liu, 2010) จะใช้ข้อมูลส่วนตัวของผู้เขียนข้อความ และ/หรือรูปโปรไฟล์ของผู้เขียนข้อความในการระบุเพศ และยังมีการระบุเพศทีละข้อความ ทำให้มีความถูกต้องของข้อมูลที่นำมาใช้ในการวิจัยเป็นอย่างมาก

ในภาพรวมค่าประสิทธิภาพจะเพิ่มขึ้นตามจำนวนคำ เมื่อเปรียบเทียบในอัลกอริทึมเดียวกันและชุดคุณลักษณะเดียวกัน ซึ่งสอดคล้องกับ (Crosby & Nyquist, 1977) ที่กล่าวว่าความแตกต่างของการใช้ภาษาระหว่างเพศหญิงและชายจะแตกต่างอย่างมีนัยยะก็ต่อเมื่อมีความยาวของการใช้ภาษามากพอ คือเมื่อข้อความที่นำมาวิเคราะห์มีความยาวไม่เพียงพอ จะทำให้ค่าประสิทธิภาพที่ได้มีค่าไม่มาก (ไม่มีความแตกต่างอย่างมีนัยยะ)

ตาราง 24 เปรียบเทียบค่าประสิทธิภาพกับงานวิจัยอื่นที่ทำการศึกษา

งานวิจัย	เทคนิคหลัก	คุณลักษณะ	ภาษา	accuracy
(Mukherjee & Liu, 2010)	TF	Frequency measure, Stylistic, Gender Preferential, Factor Analysis, Word Classes	อังกฤษ	88.56%
(Bartle & Zheng, 2015)	WRCNN	Normal word	อังกฤษ	86.00%
(Poulston et al., 2017)	TF-IDF, Word embedding	Normal word	อังกฤษ, สเปน, โปรตุเกส, อารบิก	83.90%
งานวิจัยนี้	TF-IDF	Normal word, Heading word, Stop word, Emoji, Part-of-Speech	ไทย	79.04%
(Reynaldo et al., 2019)	Bag of Word	Normal word	อังกฤษ	78.60%
(Veenhoven et al., 2018)	TF-IDF, TF-IDF weighting	Normal word, Character, Emoji	อังกฤษ, สเปน, อารบิก	77.80%
(Horsuwan et al., 2019)	TF-IDF	Normal word	ไทย	77.52%
(Akhtyamova et al., 2017)	Word2vec	Normal word	อังกฤษ, สเปน, โปรตุเกส, อารบิก	74.46%

เมื่อพิจารณาคุณลักษณะที่มีผลต่อการจำแนกเพศ จะเห็นได้ว่าคุณลักษณะที่เป็นคำศัพท์สามารถบ่งบอกถึงสิ่งที่แต่ละเพศสนใจได้เป็นอย่างดี อาทิเช่น เพศชายสนใจในเรื่องกีฬา ฟุตบอล หุ้นและโทรศัพท์มือถือ (สมาร์ทโฟน) ส่วนเพศหญิงสนใจในเรื่องความบันเทิง เช่น ละคร ดารา การประกวดต่างๆ เป็นต้น แต่คุณลักษณะ emoji ยังไม่ชัดเจนเนื่องจากมีเพียง 3 คุณลักษณะเท่านั้นที่ใช้ในแบบจำลอง ซึ่งสนับสนุนสมมติฐานข้อ 1 ที่ว่าเพศหญิงและเพศชายมีการใช้คำศัพท์ในข้อความบนโซเชียลเน็ตเวิร์กที่แตกต่างกัน

ในส่วนของคุณลักษณะคำแสดงความลังเล คำที่เด่นชัดเนื่องจากมีค่าความสำคัญคุณลักษณะแตกต่างจากคำอื่นอย่างชัดเจนในเพศชายคือคำว่า “ถ้า” ส่วนเพศหญิงคือคำว่า “ชอบ” กับคำว่า “คิดว่า” คุณลักษณะคำหยุดจะพบว่าเพศชายมักใช้คำว่า “นะ” และเพศหญิงจะใช้คำว่า “นาง” กับคำว่า “จ้า” ซึ่งสนับสนุนสมมติฐานข้อ 1 ที่ว่าเพศหญิงและเพศชายมีการใช้คำศัพท์ในข้อความบนโซเชียลเน็ตเวิร์กที่แตกต่างกัน

เมื่อพิจารณาในส่วนของคุณลักษณะชนิดของคำใน 10 อันดับแรกในภาพรวมจะพบว่าเพศชายและเพศหญิงจะใช้คำกริยาและคำสันธานที่แตกต่างกันอย่างชัดเจน กล่าวคือ เพศชายมักจะใช้คำกริยาประเภท VSTA (คำกริยาที่เกี่ยวกับความรู้สึก สภาวะความเป็นอยู่ เช่น รู้สึก ได้ยิน เห็น) และ VATT (คำกริยาที่เป็นคำคุณศัพท์ สภาวะต่างๆ เช่น ดี สวย อ้วน ผอม) ส่วนเพศหญิงจะใช้คำกริยาประเภท VACT (คำกริยาที่แสดงการกระทำ เช่น กิน เดิน วิ่ง) ในส่วนของคำสันธาน เพศชายมักจะใช้คำสันธานประเภท JSBR (คำเชื่อมประโยคหลักกับประโยคย่อย เช่น เพราะว่า เนื่องจาก แม้ว่า) ส่วนเพศหญิงจะใช้คำสันธานประเภท JCRG (คำเชื่อมประโยคหลักกับประโยคหลัก เช่น และ หรือ แต่) และ JCMP (คำเชื่อมประโยคเปรียบเทียบ เช่น กว่า เหมือนกับ เท่ากับ) ซึ่งสนับสนุนสมมติฐานข้อ 3 ที่ว่าเพศหญิงและเพศชายมีการใช้ชนิดของคำในข้อความบนโซเชียลเน็ตเวิร์กที่แตกต่างกัน และเมื่อพิจารณาถึงประเภทของคำเป็นคู่ของชนิดคำซึ่งแสดงถึงรูปแบบการใช้คำ ดังภาพประกอบ 46 จะเห็นได้ว่าทั้งเพศหญิงแล้วเพศชายมีรูปแบบการใช้คำที่ต่างกัน ซึ่งสนับสนุนสมมติฐานข้อ 2 ที่ว่าเพศหญิงและเพศชายมีการใช้รูปแบบของคำในข้อความบนโซเชียลเน็ตเวิร์กที่แตกต่างกัน

ข้อจำกัดและสิ่งที่ต้องพัฒนาในงานวิจัยนี้

1. ในงานวิจัยนี้ใช้การระบุเพศของผู้เขียนข้อความแสดงความคิดเห็นจากคำลงท้ายประโยค อาจจะไม่มีความถูกต้องของข้อมูลเท่าที่ควร เนื่องจากในความเป็นจริง เพศชายสามารถเขียนข้อความลงท้ายด้วย คะ ค่ะ ส่วนเพศหญิงก็สามารถเขียนข้อความลงท้ายด้วยคำว่าครับ ถ้ามี

ข้อมูลที่ระบุเพศของผู้เขียนข้อมูลได้ถูกต้องมาใช้ในการสร้างแบบจำลอง อาจจะช่วยเพิ่มประสิทธิภาพของแบบจำลองให้ดีมากยิ่งขึ้น

2. คุณลักษณะ เช่น emoji ที่ใช้ในงานวิจัยนี้ยังไม่สามารถจำแนกเพศได้เนื่องจากมีเพียง 3 คุณลักษณะเท่านั้นที่ถูกใช้ในแบบจำลองที่ให้ค่าประสิทธิภาพที่ดีที่สุดในงานวิจัยนี้ อาจหาวิธีการสกัดคุณลักษณะอื่นๆ จาก emoji รวมถึงคุณลักษณะประเภทอื่นๆ ซึ่งอาจมีคุณลักษณะใหม่ๆ ที่สามารถใช้ในการจำแนกเพศได้ดีกว่าคุณลักษณะที่ใช้ในงานวิจัยนี้

3. เนื่องจากข้อมูลในโซเชียลเน็ตเวิร์กช่องทางต่างๆ ไม่เฉพาะแค่ในพันทิพเท่านั้น จะมีการเปลี่ยนแปลงตามสิ่งที่กำลังเป็นที่นิยมในขณะนั้น ทำให้แบบจำลองที่ใช้คำศัพท์เป็นคุณลักษณะหลักในแบบจำลอง ไม่สามารถนำไปใช้ในช่วงเวลาที่ต่างกันได้ เนื่องจากจะเกิดกรณีที่ไม่มีคำศัพท์ใหม่ๆ อยู่ในคุณลักษณะของแบบจำลอง ทำให้ค่าประสิทธิภาพลดลงอย่างมาก จึงควรหาวิธีใช้คุณลักษณะอื่นๆ ที่ไม่ค่อยเปลี่ยนแปลงตามช่วงเวลา หรือเปลี่ยนไม่มาก เช่น คำหยุด ชนิดของคำ (อาจจะใช้แค่คำกริยา หรือคำสันธาน หรือคำอื่นๆ) เป็นต้น มาสร้างเป็นแบบจำลอง จะทำให้สามารถใช้ได้หลากหลายหัวข้อและช่วงเวลามากกว่าใช้คำศัพท์เป็นคุณลักษณะ

ข้อเสนอแนะ

1. ในงานวิจัยนี้มีการเลือกใช้อัลกอริทึมเพียง 3 อัลกอริทึม ซึ่งอาจจะมีอัลกอริทึมอื่นที่สามารถให้ค่าประสิทธิภาพที่ดีกว่า 3 อัลกอริทึมนี้

2. จากผลการวิจัยจะเห็นได้ว่า คำศัพท์ เป็นคุณลักษณะที่สามารถจำแนกเพศได้ดีที่สุด แต่เนื่องจากงานวิจัยนี้ได้ใช้ข้อมูลจากเว็บไซต์พันทิพซึ่งมีหลากหลายหัวข้อ ทำให้คำศัพท์มีหลากหลาย จึงต้องจำกัดจำนวนคำศัพท์ที่เป็นคุณลักษณะที่ใช้สร้างแบบจำลอง ทำให้เมื่อนำแบบจำลองมาทดสอบกับชุดข้อมูลทดสอบอาจจะเกิดกรณีที่ไม่มีคำศัพท์บางคำอยู่ในคุณลักษณะที่ใช้ในแบบจำลอง (out of vocabulary) ทำให้ค่าประสิทธิภาพลดลง ถ้าหากเปลี่ยนเป็นทำแบบจำลองเฉพาะหัวข้อใดหัวข้อหนึ่ง น่าจะทำให้ค่าประสิทธิภาพการจำแนกดีขึ้นอย่างมาก

บรรณานุกรม

Akhtyamova, L., Cardiff, J., & Ignatov, A. (2017). *Twitter Author Profiling Using Word Embeddings and Logistic Regression.* Paper presented at the CLEF 2017 Evaluation Labs and Workshop, Dublin, Ireland.

Bartle, A., & Zheng, J. (2015). *Gender Classification with Deep Learning*.

Chen, Z., Lu, X., Ai, W., Li, H., Mei, Q., & Liu, X. (2018). *Through a Gender Lens: Learning Usage Patterns of Emojis from Large-Scale Android Users*. Paper presented at the Proceedings of the 2018 World Wide Web Conference, Lyon, France.

Crosby, F., & Nyquist, L. (1977). The female register: an empirical study of Lakoff's hypotheses. *Language in Society, 6*(3), 313-322.

Dalal, M. K., & Zaveri, M. A. (2011). Automatic Text Classification: A Technical Review. *International Journal of Computer Applications, 28*, 37-40.

Gillett, A. Features of academic writing: Hedging. http://www.uefap.com/writing/feature/hedge.htm

Horsuwan, T., Kanwatchara, K., Vateekul, P., & Kijsirikul, B. (2019). A Comparative Study of Pretrained Language Models on Thai Social Text Categorization. *arXiv preprint arXiv:1912.01580*.

Jurafsky, D., & Martin, J. H. (2000). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*: Prentice Hall PTR.

Liddy, E. D. (2001). Natural Language Processing. In *Encyclopedia of Library and Information Science* (2nd ed.). NY: Marcel Decker, Inc.

Mukherjee, A., & Liu, B. (2010). *Improving Gender Classification of Blog Authors*. Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge, MA.

Müller, A. C., & Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientists* (1st): O'Reilly Media.

Poulston, A., Waseem, Z., & Stevenson, M. (2017). *Using TF-IDF n-gram and Word*

Embedding Cluster Ensembles for Author Profiling. Paper presented at the CLEF 2017 Evaluation Labs and Workshop, Dublin, Ireland.

Reynaldo, N., Goenawan, Chanrico, W., Suhartono, D., & Purnomo, F. (2019). Gender Demography Classification on Instagram based on User's Comments Section. *Procedia Computer Science, 157*, 64-71.

Smith, S. (2020). Hedging. https://www.eapfoundation.com/writing/skills/hedging/

Tsujii, J. (2011). *Computational linguistics and natural language processing*. Paper presented at the Proceedings of the 12th international conference on Computational linguistics and intelligent text processing - Volume Part I, Tokyo, Japan.

Veenhoven, R., Snijders, S., Hall, D. v. d., & Noord, R. v. (2018). *Word unigram weighing for author profiling.* Paper presented at the CLEF 2018 Evaluation Labs and Workshop, Avignon, France.

Virach, S., Naoto, T., & Hitoshi, I. (1999). Building a Thai Part-Of-Speech Tagged Corpus (ORCHID). *The Journal of the Acoustical Society of Japan (E), 20*(3), pp. 189-198.

เฟื่องขจรศักดิ์, ว. (2558). การใช้ถ้อยคำแสดงความลังเลของผู้หญิงและผู้ชายในบริบทการวิพากษ์วิจารณ์: The use of hedging utterance of women and men in critical context. วารสารมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยรังสิต ปีที่ *11, 19*, 44-60.

เศรษฐธัญการ, ช. (2562). การศึกษาเปรียบเทียบความสามารถในการใช้ภาษาระหว่างเพศหญิงและเพศชายที่สะท้อนให้เห็นภาพพจน์ทางเพศ วารสารมังรายสาร, ปีที่ *7* ฉบับที่ *2*, 16.

จิรวัฒน์กุล, อ. (2552). สถิติทางวิทยาศาสตร์สุขภาพเพื่อการวิจัย. กรุงเทพฯ: บริษัทวิทยพัฒน์ จำกัด.

จิระวิชิตชัย, น., สงวนสัตย์, ป., และ มีสัจ, พ. (2010). การจัดหมวดหมู่เอกสารภาษาไทยด้วยเครือข่ายฟังก์ชันฐานรัศมี: *Thai Document Categorization with Radial Basis Function Network.* Paper presented at the NCIT 2010 the National Conference on Information Technology : "IT Innovation for Global Awareness", Bangkok, Thailand.

ตาเมือง, ม. (2555). ความเหลื่อมล้ำในการใช้ภาษาเพื่อการอภิปรายกลุ่มในชั้นเรียนของนิสิตต่างเพศ: กรณีศึกษารายวิชาภาษา สังคมและวัฒนธรรม มหาวิทยาลัยนเรศวร. วารสารมนุษยศาสตร์ มหาวิทยาลัยนเรศวร ปีที่ *9*(ฉบับที่ 1 มกราคม - เมษายน 2555), 67-80.

ตาดทอง, น. (2558). เพศกับกลวิธีการใช้ภาษาบนเฟซบุ๊ก. (ปริญญานิพนธ์ กศ.ม. ภาษาศาสตร์การศึกษา). บัณฑิตวิทยาลัย มหาวิทยาลัยศรีนครินทรวิโรฒ, กรุงเทพมหานคร.

ทองพูล, พ. (2559). รูปเบี่ยงบังและหน้าที่ในแผ่นพับโฆษณาผลิตภัณฑ์เสริมอาหารลดน้ำหนัก : *HEDGES AND THEIR FUNCTIONS IN DIETARY SUPPLEMENT FOR WEIGHT LOSS ADVERTISING BROCHURES.* (ศิลปศาสตรมหาบัณฑิต). มหาวิทยาลัยธรรมศาสตร์, มหาวิทยาลัยธรรมศาสตร์. (วิทยานิพนธ์, สาขาวิชาภาษาไทย ภาควิชาภาษาไทยและภาษาวัฒนธรรมตะวันออก คณะศิลปศาสตร์).

นาคพันธุ์, ก. (2561). เฮดจ์ (Hedge). https://www.facebook.com/ThinkerTinkersLArtsTU/photos/a.365206150655618/367313343778232

ประสิทธิ์รัฐสินธุ์, อ. (2545). ภาษาในสังคมไทย ความหลากหลาย การเปลี่ยนแปลง และการพัฒนา. กรุงเทพมหานคร: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย.

ประวัติผู้เขียน

ชื่อ-สกุล	เอกภพ พูลสวัสดิ์
วัน เดือน ปี เกิด	1 มีนาคม 2529
สถานที่เกิด	นครปฐม
วุฒิการศึกษา	2564 วิทยาศาสตร์มหาบัณฑิต (เทคโนโลยีสารสนเทศ) มหาวิทยาลัยศรีนครินทรวิโรฒ 2552 วิทยาศาสตร์บัณฑิต (ธรณีวิทยา) จุฬาลงกรณ์มหาวิทยาลัย