การจำแนกเพศจากข้อความบนโซเชียลเน็ตเวิร์ก

Gender Classification From Text Data In Social Network

เอกภพ พูลสวัสดิ์
คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ
กรุงเทพมหานคร, ประเทศไทย
ekkapob.poonsawat@g.swu.ac.th

วีรยุทธ เจริญเรืองกิจ
คณะวิทยาศาสตร์
มหาวิทยาลัยศรีนครินทรวิโรฒ
กรุงเทพมหานคร, ประเทศไทย
werayuth@g.swu.ac.th

***บทคัดย่อ*—ในยุคที่ข้อมูลมีมากมายมหาศาลบนโลกของอินเตอร์เน็ต การที่เราสามารถเก็บข้อมูลของบุคคลกับความสนใจต่างๆ ได้ ทำให้เราสามารถใช้ประโยชน์จากข้อมูลเหล่านี้ได้อย่างเต็มที่ อาทิเช่น ถ้านักการตลาดรู้ว่าสินค้าของบริษัทชิ้นนี้อยู่ในความสนใจของคนกลุ่มไหน ไม่ว่าจะเป็นคนอายุน้อย คนอายุเยอะ เพศชาย เพศหญิง คนในกรุงเทพฯ หรือคนต่างจังหวัด จะทำให้สามารถวางแผนการตลาดเจาะกลุ่มบุคคลได้อย่างถูกต้อง ซึ่งเป็นการตลาดที่เรียกว่าการตลาดส่วนบุคคล (Personalized Marking) ทำให้เพิ่มประสิทธิภาพในการตลาดได้มากยิ่งขึ้น แต่ข้อมูลที่อยู่บนอินเตอร์เน็ตนั้นมักมาในรูปของข้อความที่ไม่มีโครงสร้างและไม่มีป้ายกำกับบอกสิ่งที่เราต้องการ ในงานวิจัยนี้ได้นำเสนอวิธีการจำแนกเพศของผู้เขียนข้อความ จากข้อความบนโซเชียลเน็ตเวิร์ก โดยการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติกับการสกัดคุณลักษณะ ร่วมกับการสร้างแบบจำลองการเรียนรู้ของเครื่อง ให้ค่าความแม่นยำในการจำแนกเพศ 79.04%**

***Keywords—Text classification, Gender classification, Social network,* การจำแนกเพศ, ความแตกต่างระหว่างเพศ**

I. บทนำ

จากการที่เทคโนโลยีมีการพัฒนาไปอย่างรวดเร็ว ทำให้เกิดการแข่งขันกันอย่างเข้มข้นในทางด้านการตลาดของหลายๆบริษัท ซึ่งการตลาดที่จะประสบความสำเร็จได้นั้น บริษัทจำเป็นต้องมีความเข้าใจถึงพฤติกรรมของผู้บริโภคเพื่อที่จะทำให้บริษัทมีความสามารถในแข่งขันกับบริษัทอื่นๆ ในตลาดได้ [1]

ความแตกต่างระหว่างพฤติกรรมของผู้บริโภคนั้นเกิดขึ้นได้จากหลากหลายปัจจัย หนึ่งในนั้นคือ ความแตกต่างระหว่างเพศของผู้บริโภค ซึ่งเป็นหนึ่งในทฤษฎีสำหรับการวิจัยทางด้านการตลาด [2] นักวิจัยทางด้านการตลาดพบว่าเพศหญิงและเพศชายนั้นมีพฤติกรรมการตัดสินใจในการใช้จ่ายแตกต่างกัน [3] เพราะฉะนั้นการสื่อสารไปยังผู้บริโภคในแต่ละเพศจึงเป็นสิ่งสำคัญที่จะเป็นตัวตัดสินว่าการตลาดนั้นจะประสบความสำเร็จหรือไม่ [4]

ด้วยเทคโนโลยีในปัจจุบันนั้น ทำให้บริษัทต่างๆ หันมาเก็บข้อมูลผู้บริโภคโดยการติดตามกิจกรรมต่างๆ ของผู้บริโภคที่เกิดขึ้นบนช่องทางโซเชียลเน็ตเวิร์กกันมากยิ่งขึ้น [5] ซึ่งข้อมูลส่วนใหญ่ที่ได้มานั้นมักเป็นข้อมูลแบบไม่มีโครงสร้าง เช่น ข้อความต่างๆ บนโซเชียลเน็ตเวิร์ก การที่สามารถนำเอาข้อมูลของผู้บริโภคออกมาจากข้อความเหล่านี้ได้ จะทำให้สามารถนำข้อมูลเหล่านี้ไปใช้ให้เกิดความได้เปรียบในการแข่งกันด้านการตลาด ซึ่งไม่เพียงแต่ด้านการตลาดเท่านั้น เรายังสามารถนำข้อมูลที่ได้ไปใช้ให้เกิดประโยชน์อื่นๆ อีกมากมาย เช่น การวิเคราะห์พฤติกรรมของคนในสังคมเพื่อให้เกิดการพัฒนาในด้านต่างๆ เป็นต้น สำหรับการจำแนกเพศของข้อความในภาษาไทยนั้นสามารถระบุได้จากคำลงท้ายประโยค เช่น ครับ ค่ะ หรือคำสรรพนามแทนตัว เช่น ผม ดิฉัน แต่ในความเป็นจริงแล้ว มีข้อความเพียง 30% เท่านั้นที่สามารถระบุเพศของผู้เขียนได้จากคำเหล่านี้ ยังเหลืออีก 70% ที่ไม่สามารถระบุได้

ในงานวิจัยนี้จึงนำเสนอวิธีการจำแนกเพศของผู้เขียนข้อความบนโซเชียลเน็ตเวิร์ก โดยการประยุกต์ใช้เทคนิคการประมวลผลภาษาธรรมชาติ (Natural Language Processing Techniques) ในการสกัดคุณลักษณะ (Features Extraction) จากข้อความ ร่วมกับการสร้างแบบจำลองสำหรับการจำแนก (Classification Model) โดยเลือกใช้อัลกอริทึมการเรียนรู้ของเครื่อง (Machine Learning Algorithms) 3 อัลกอริทึม ได้แก่ Logistic Regression, Naive Bayes และ Random Forest

II. เอกสารและงานวิจัยที่เกี่ยวข้อง

A. *ความแตกต่างของการใช้ภาษาระหว่างเพศหญิงและชาย*

เพศคือหนึ่งในหลากหลายปัจจัยที่ทำให้เกิดความแตกต่างของการใช้ภาษา เพศชายและเพศหญิงมีการใช้ภาษาที่แตกต่างกันอย่างมีนัยยะสำคัญในกรณีที่มีการใช้ภาษาอย่างเป็นทางการและมีความยาวของการใช้ภาษามากพอ กล่าวคือความแตกต่างเรื่องเพศในการใช้ภาษานั้นขึ้นอยู่กับบริบทของการใช้ภาษานั้นๆ [6]

สำหรับภาษาไทยนั้นความแตกต่างทางภาษาของเพศชายและเพศหญิงที่เห็นได้อย่างชัดเจนก็คือ สรรพนามที่ใช้เฉพาะเพศ เช่น ผม กระผม สำหรับเพศชาย และฉัน ดิฉัน หนู สำหรับเพศหญิง หรือคำลงท้าย ครับ สำหรับเพศชาย คะ ค่ะ สำหรับเพศหญิง เป็นต้น [7]

การใช้ภาษาของทั้งเพศชายและเพศหญิงมีความคล้ายคลึงกัน และมีแนวโน้มที่จะมีการใช้ภาษาแบบเสมอภาคกัน ส่วนที่มีความแตกต่างอย่างเห็นได้ชัดระหว่างการใช้ภาษาของเพศหญิงและเพศชาย คือ การใช้คำแสดงอารมณ์ความรู้สึก โดยเพศหญิงมักมีการใช้คำที่แสดงความเป็นกันเอง ส่วนเพศชายมักใช้คำที่แสดงความรุนแรง [8]

การใช้ถ้อยคำเพื่อการอธิบายสิ่งต่างๆ พบว่าเพศชายสามารถอธิบายสิ่งที่เป็นรูปธรรมได้มากกว่าเพศหญิง ในขณะที่เพศหญิงสามารถที่จะอธิบายสิ่งที่เป็นนามธรรมได้มากกว่าเพศชาย [9]

ส่วนในการใช้ถ้อยคำแสดงความลังเล (Hedging Words) นั้น เพศหญิงมีการใช้ถ้อยคำแสดงความมั่นใจมากกว่าเพศชาย ในขณะที่เพศชายจะใช้ถ้อยคำแสดงความไม่มั่นใจมากกว่าเพศหญิง [10]

B. *การจำแนกข้อความโดยใช้เทคนิคทางด้านการประมวลผลภาษาธรรมชาติ*

ขั้นตอนพื้นทางของงานวิจัยทางด้านการจำแนกข้อความโดยใช้เทคนิคทางด้านการประมวลผลภาษาธรรมชาติและเทคนิคการเรียนรู้ของเครื่อง ประกอบด้วย 5 ขั้นตอน ดังนี้ [11], [12]

1) ขั้นตอนการจัดเตรียมข้อมูล (Pre-Processing) เช่น การตัดคำ (Word Segmentation) การกำจัดคำหยุด (Stop-Word Removal) การหารากศัพท์ (Stemming)

2) ขั้นตอนการสกัดคุณลักษณะ (Extract Features) เป็นการเปลี่ยนข้อความให้เป็นคุณลักษณะโดยใช้เทคนิคต่างๆ เช่น Term Frequency – Inverse Document Frequency (TF-IDF), Word2Vec เป็นต้น

3) ขั้นตอนการเลือกอัลกอริทึมสำหรับการสร้างแบบจำลองการเรียนรู้ของเครื่องสำหรับการจำแนกข้อความ (Machine Learning Algorithms for Classification)

4) ขั้นตอนการฝึกสอนและประเมินประสิทธิภาพของแบบจำลอง (Model Training and Evaluation)

5) ขั้นตอนการนำแบบจำลองไปประเมินผลกับชุดข้อมูลทดสอบ (Model Testing)

C. *เทคนิคสำหรับการจำแนกที่ใช้ในงานวิจัย*

การจำแนกเป็นหนึ่งในเทคนิคการเรียนรู้ของเครื่องประเภทการเรียนรู้แบบมีผู้สอน (Supervised Learning) ซึ่งจำเป็นต้องมีคำตอบสำหรับให้อัลกอริทึมได้เรียนรู้ก่อนที่จะนำแบบจำลองที่ได้ไปใช้ในการจำแนกต่อไป สำหรับในงานวิจัยนี้ได้เลือกใช้อัลกอริทึมที่เป็นที่นิยมใช้สำหรับการจำแนก 3 อัลกอริทึม ได้แก่ Logistic Regression, Naive Bayes และ Random Forest [13]

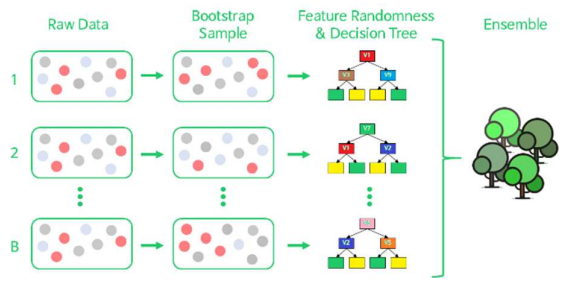
- Logistic Regression เป็นอัลกอริทึมสำหรับการจำแนกประเภท ซึ่งผลลัพธ์ของอัลกอริทึมนี้อยู่ในรูปของความน่าจะเป็น การทำงานของอัลกอริทึมนี้จะแบ่งเป็น 2 ขั้นตอนคือ ทำการคำนวณ Linear Combination ของ Input กับ Parameters ของ Model จากนั้นนำผลลัพธ์จากขั้นแรกมาคำนวณ Sigmoid Function ดังสมการ (1) ได้ผลลัพธ์สุดท้ายออกมาเป็นค่าความน่าจะเป็นที่ซึ่งมีค่าอยู่ระหว่าง 0 ถึง 1

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

- Naive Bayes อัลกอริทึมนี้จะใช้หลักการคำนวณความน่าจะเป็น โดยใช้สิ่งที่เรียกว่าทฤษฎีของเบย์ (Bayes Theorem) ดังสมการ (2)

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

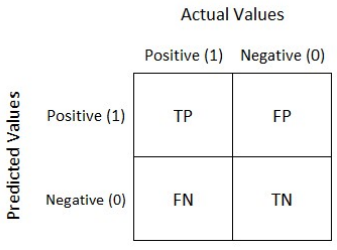
- Random Forest เป็นอัลกอริทึมที่พัฒนาขึ้นจากอัลกอริทึม Decision Tree โดยเป็นการนำแบบจำลอง Decision Tree หลายๆ แบบจำลองมาใช้ทำนายร่วมกัน ซึ่งแต่ละแบบจำลองนั้นจะใช้ข้อมูลและคุณลักษณะที่นำมาสร้างแบบจำลองไม่เหมือนกัน หลังจากนั้นจึงนำผลการทำนายที่ได้จากแบบจำลองมาทำการลงคะแนน (Voting) เพื่อให้ได้ผลลัพธ์สุดท้าย ดังแสดงในรูปที่ 1



รูปที่ 1 การทำงานของอัลกอริทึม Random Forest

D. วิธีการวัดประสิทธิภาพของการทดลองที่ใช้ประเมินผลการวิจัย

Confusion Matrix เป็นอีกหนึ่งวิธีที่ดีในการวัดประสิทธิภาพแบบจำลองการทำนายสำหรับปัญหาการจำแนก โดยมีแนวคิดคือการนับจำนวนครั้งสำหรับการจำแนกประเภทถูกและการจำแนกประเภทผิดของแต่ในละประเภท (Class) ดังรูปที่ 2 ซึ่ง Confusion Matrix นั้นจะให้ข้อมูลที่หลากหลายไม่เพียงแค่ค่าความถูกต้อง (Accuracy) แต่ยังมีค่าอื่นๆอีก โดยในงานวิจัยนี้นอกจากค่าความถูกต้องแล้ว จะใช้อีก 3 ค่าในการวัดประสิทธิภาพของแบบจำลองได้แก่ Precision, Recall และ F_1Score [14]



รูปที่ 2 Confusion Matrix ขนาด 2x2

- Accuracy คือ ค่าความถูกต้องของการจำแนกประเภท โดยการพิจารณารวมทุกประเภท สามารถคำนวณได้ ดังสมการ (3)

$$Accuracy = \frac{(TP+TN)}{(TP+TN+FP+FN)} \qquad (3)$$

- Precision คือ ค่าความแม่นยำของการจำแนกประเภท โดยการพิจารณาแยกทีละประเภท สามารถคำนวณได้ ดังสมการ (4)

$$Precision = \frac{TP}{(TP+FP)} \qquad (4)$$

- Recall คือ ค่าความถูกต้องของการจำแนกประเภท โดยการพิจารณาแยกทีละประเภท สามารถคำนวณได้ ดังสมการ (5)

$$Recall = \frac{TP}{(TP+FN)} \qquad (5)$$

- F_1Score คือ ค่าเฉลี่ยแบบฮาโมนิค (Harmonic Mean) ระหว่าง Precision และ Recall สามารถคำนวณได้ ดังสมการ (6)

$$F_1Score = 2 \times \frac{Precision \times Recall}{(Precision+Recall)} \qquad (6)$$

E. งานวิจัยเกี่ยวข้องกับการจำแนกเพศของผู้เขียนข้อความ

State-of-The-Art ของงานวิจัยทางด้านการจำแนกเพศของผู้เขียนข้อความในภาษาอังกฤษ โดยงานวิจัยนี้ได้ใช้ข้อมูลจากเว็บบล็อกทั้งสิ้น 3,100 เว็บบล็อก แบ่งเป็นเว็บบล็อกของเพศชาย 1,588 เว็บบล็อก และเว็บบล็อกของเพศหญิง 1,512 เว็บบล็อก ซึ่งมีความยาวเฉลี่ยของข้อความ 250 คำสำหรับข้อความของเพศชาย และ 330 คำสำหรับข้อความของเพศหญิง ในงานวิจัยนี้ได้นำเสนอเทคนิคใหม่ 2 เทคนิค ได้แก่ เทคนิคการสกัดคุณลักษณะรูปแบบของชนิดของคำ (Part-of-Speech Pattern Features) และเทคนิคการเลือกคุณลักษณะแบบองค์รวม (Ensemble of Feature Selection) ร่วมกับคุณลักษณะอื่นๆ อีก 4 แบบ คือ ค่าความถี่ชนิดของคำ (Frequency Measure) รูปแบบของคำ (Stylistic Features) รูปแบบของการใช้คำเฉพาะของแต่ละเพศ (Gender Preferential Features) และการวิเคราะห์ปัจจัยและประเภทของคำ (Factor Analysis and Word Classes) ส่วนการแปลงข้อความให้เป็นเวคเตอร์นั้นใช้วิธี TF (Term Frequency) และใช้อัลกอริทึม SVM regression ได้ค่าความถูกต้องสูงสุด 88.56% [15]

การจำแนกเพศของผู้เขียนข้อความบนทวิตเตอร์ ได้มีการนำเสนอเอาไว้ 2 วิธีคือ การใช้ TF-IDF n–grams กับอัลกอริทึม Logistic Regression และอีกวิธีคือ การใช้ Word Embedding Clusters กับอัลกอริทึม Gaussian Process Classifier พบว่าทั้ง 2 วิธีนี้ได้ค่าความถูกต้องของแบบจำลองการจำแนกเพศมากที่สุดอยู่ที่ภาษาโปรตุเกส โดยมีค่าความถูกต้องที่ 82.6% และ 83.9% ตามลำดับ [16] ส่วนอีกงานวิจัยหนึ่งได้นำเสนอการใช้ 3 คุณลักษณะ 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% [17]

นอกจากนี้ยังมีงานวิจัยที่เลือกใช้ข้อความที่มี Emoji อยู่ในข้อความ มาทำการจำแนกเพศของผู้เขียนข้อความ โดยใช้ข้อความที่เก็บจากโปรแกรมแป้นพิมพ์บนแอพลิเคชั่นบนสมาร์ทโฟนระบบปฏิบัติการแอนดรอยด์ จากผู้ใช้งานจำนวน 39,372 ผู้ใช้งาน ผู้ใช้งานละอย่างน้อย 100 ข้อความ จาก 58 ภาษารวมถึงภาษาไทย โดยใช้ Emoji เพียงอย่างเดียวมาทำการสกัดคุณลักษณะออกมาเป็นคุณลักษณะทั้งสิ้นจำนวน 1,370 คุณลักษณะ ได้ค่าความถูกต้องสูงสุดอยู่ที่ภาษาโปรตุเกส 84.1% ส่วนภาษาไทยนั้นอยู่ที่ 80.8% ด้วยอัลกอริทึม Gradient Boosting Classifier [18]

F. งานวิจัยอื่นๆ ที่เกี่ยวข้อง

งานวิจัยนี้ได้ทำการพัฒนาแบบจำลองทางภาษาก่อนการฝึกฝน (Pre-Train Language Modal) สำหรับภาษาไทย โดยได้ใช้ข้อมูลจากเว็บไซต์พันทิพ ซึ่งมีการพัฒนาพจนานุกรม (Dictionary) โดยการเพิ่มเติมคำศัพท์ที่ใช้กันในโซเชียลเน็ตเวิร์ก ลงในพจนานุกรมสำหรับใช้ตัดคำของไลบรารี pyThaiNLP ซึ่งเป็นไลบรารีประมวลผลภาษาธรรมชาติสำหรับภาษาไทยแบบโอเพนซอร์สในภาษาไพธอน (Python) [19] และวัดประสิทธิภาพการตัดคำกับวิธีปกติของไลบรารี pyThaiNLP พบว่าได้ค่าความถูกต้องเพิ่มขึ้นประมาณ 3.4% จากนั้นทาง

คณะผู้วิจัยได้ทำการสร้างแบบจำลองทางภาษา (Language Model) จาก State-of-The-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 โดยผู้เข้าแข่งกันส่วนใหญ่ของทั้งสองการแข่งขันนี้มักจะแบบจำลองทางภาษาก่อนการฝึกฝนที่ทำการฝึกฝนมาจากข้อมูล Thai Wiki Dump และตัดคำโดยใช้ไลบรารี pyThaiNLP ซึ่งกำลังเป็นที่นิยมกันในงานทางด้านการประมวลผลภาษาธรรมชาติของภาษาไทย [20]

III. การเก็บข้อมูลและการเตรียมข้อมูล

A. *การเก็บข้อมูล (Data Acquisition)*

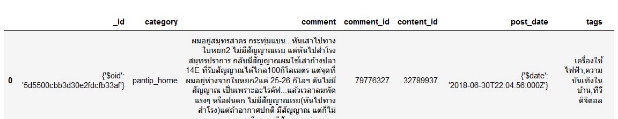
ข้อมูลที่ใช้ในงานวิจัยนี้ คือ ข้อความแสดงความคิดเห็นบนเว็บไซต์พันทิพดอทคอม (Pantip.com) ดังรูปที่ 3 ผู้วิจัยได้ทำการเก็บข้อมูลด้วยวิธี Scraping โดยใช้ไพธอนไลบรารีที่ชื่อว่า Selenium และ BeautifulSoup ตั้งแต่เดือน มิ.ย. 2562 ถึง ก.ค. 2562 มีจำนวนข้อความทั้งหมดก่อนเข้าสู่ขั้นตอนการทำความสะอาดข้อมูลทั้งสิ้น 854,472 ข้อความ

รูปที่ 3 ความคิดเห็นบนหน้าเว็บไซต์พันทิพ

ตารางที่ 1 รายละเอียดของคอลัมภ์ข้อมูลที่เก็บมาจากเว็บไซต์พันทิพดอทคอม

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

จากตารางที่ 1 แสดงให้เห็นถึงรายละเอียดของข้อมูลที่เก็บมาจากเว็บไซต์พันทิพดอทคอม โดยมีชื่อของคอลัมภ์ต่างๆ ของข้อมูลที่เก็บมาและรายละเอียดของแต่ละคอลัมภ์ว่าคือข้อมูลอะไร ดังรูปที่ 4 แสดงตัวอย่างข้อมูลเมื่อนำเข้าสู่ดาต้าเฟรม (Pandas DataFrame) บนโปรแกรมภาษาไพธอน



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

รูปที่ 4 ตัวอย่างข้อมูลที่เก็บมาจากเว็บไซต์พันทิพดอทคอมในดาต้าเฟรมบนโปรแกรมภาษาไพธอน

B. *การระบุประเภทข้อมูล (Data Labeling)*

เนื่องจากข้อมูลที่ได้มานั้น ยังไม่มีการระบุประเภทของข้อมูลว่าข้อความไหนผู้เขียนข้อความเป็นเพศหญิงหรือเพศชาย ดังนั้นจึงต้องมีการระบุประเภทของข้อมูลก่อน โดยการดูจากคำสรรพนามบุรุษที่ 1 และคำลงท้ายประโยค [7] ดังตารางที่ 2 แสดงคำที่ใช้สำหรับระบุเพศของผู้เขียนข้อความ โดยมีทั้งคำที่สะกดถูกต้องตามพจนานุกรม และสะกดผิด หลังจากนั้นจึงเลือกเฉพาะข้อความที่สามารถระบุเพศของผู้เขียนข้อความมาใช้สำหรับงานวิจัย ทั้งสิ้น 247,910 ข้อความ ซึ่งจะเห็นได้ว่ามีเพียงประมาณ 30% ของข้อมูลเท่านั้นที่สามารถระบุเพศได้

ตารางที่ 2 คำที่ใช้สำหรับระบุเพศของผู้เขียนข้อความ

คำที่ใช้	label
ผม, ผ ม, กระผม	male
ครับ, ครัช, ครัส, คับ, ค่าบ	male
ดิฉัน, ฉัน, หนู	female
คะ, ค่ะ, คั๊ะ, คร๊ะ, ค ะ	female
นะคะ, นะค่ะ	female
จ๊ะ, จ้า	female

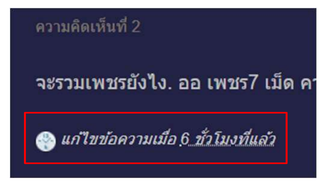
C. *การทำความสะอาดข้อมูล (Data Cleaning)*

ในขั้นตอนนี้จะเป็นการทำความสะอาดข้อมูล ซึ่งแบ่งออกได้เป็น 7 ขั้นตอนดังนี้

1) เลือกข้อมูลเฉพาะคอลัมภ์ที่ต้องการใช้ ได้แก่ comment

2) จัดการกับข้อความแสดงความคิดที่มีแต่การเคาะเว้นวรรคโดยไม่มีการพิมพ์ตัวหนังสือ ให้กลายเป็นข้อความว่าง

3) จัดการกับข้อความที่สร้างขึ้นโดยระบบของทางเว็บไซต์ ดังรูปที่ 5 แสดงข้อความที่ระบบสร้างขึ้นเมื่อมีการแก้ไขข้อความแสดงความคิดเห็น โดยการแทนที่ด้วยข้อความว่าง



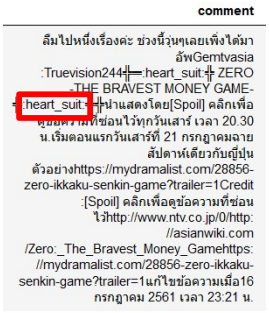
รูปที่ 5 ข้อความที่ระบบสร้างขึ้นเมื่อมีการแก้ไขข้อความแสดงความคิดเห็น

4) ลบข้อความที่แสดงความคิดเห็นที่ซ้ำกันออกไป จากสมมติฐานที่ว่าความคิดเห็นไม่ควรจะซ้ำกันทั้งประโยค ถ้ามีการซ้ำกันทั้งประโยคน่าจะเป็นโฆษณา แชทบอท หรือแอดมินของผลิตภัณฑ์ต่างๆ ที่คอยตอบคำถามอยู่ในเว็บไซต์พันทิพดอทคอม หรือคำที่สั้นๆ ที่เป็นคำสามัญซึ่งไม่สามารถบ่งบอกเพศได้ เมื่อตัดคำลงท้ายออกไป เช่น ข้อความที่มีแต่คำว่า "ขอบคุณครับ" เมื่อตัดคำว่า "ครับ" ออก (ซึ่งเป็นขั้นตอนการกำกับประเภทข้อความ) จะทำให้ไม่สามารถจำแนกเพศจากข้อความเหล่านี้ได้ ดังรูปที่ 6 แสดงตัวอย่างข้อความแสดงความคิดเห็นที่ซ้ำกันพร้อมกับจำนวนที่ซ้ำ

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

รูปที่ 6 ตัวอย่างข้อความแสดงความคิดเห็นที่ซ้ำกันพร้อมกับจำนวนที่ซ้ำ

5) จัดการกับ Emoji ซึ่งมาในรูปของตัวหนังสือ โดยการ Tag คำว่า emoji ลงไปที่ด้านหน้าเพื่อให้ทราบว่าเป็น Emoji ไม่ใช่ตัวข้อความ ดังรูปที่ 7 แสดงข้อความที่มี Emoji ที่มาในรูปแบบข้อความบนข้อความแสดงความคิดเห็น จะเห็นว่า Emoji ที่ได้มานั้น จะอยู่ในรูปของข้อความที่ทางเว็บไซต์พันทิพดอทคอมจัดทำขึ้นมาสำหรับใช้บนเว็บไซต์

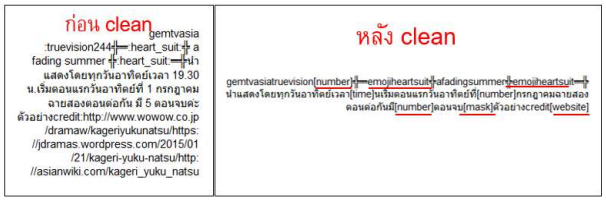


รูปที่ 7 ตัวอย่างข้อความที่มี Emoji ที่มาในรูปแบบข้อความบนข้อความแสดงความคิดเห็น

6) จัดการกับรูปแบบคำเฉพาะ เช่น ชื่อเว็บไซต์ วันที่ เวลา เบอร์โทรศัพท์ ที่เป็นตัวเลข โดยการเปลี่ยนให้เป็น Tag ของคำนั้นๆ

7) จัดการกับสัญลักษณ์ต่างๆ ที่ไม่มีความหมายของคำ โดยการเปลี่ยนเป็นข้อความว่าง

8) จัดการคำกับที่ใช้ระบุเพศของผู้เขียนข้อความโดยปกปิดคำ (Masking) ด้วยการใช้ [mask] แทนคำที่สามารถระบุเพศ



รูปที่ 8 เปรียบเทียบข้อความก่อนและหลังทำความสะอาด

D. การเตรียมข้อมูล (Data Pre-Processing)

1) การเตรียมพจนานุกรมคำศัพท์สำหรับใช้ในการตัดคำ

ในงานวิจัยนี้จะใช้พจนานุกรมตั้งต้นของงานวิจัย [20] และทำการเพิ่มคำศัพท์ที่ได้จากการศึกษางานวิจัยที่เกี่ยวข้อง ได้แก่ คำแสดงความลังเล คำหยุดและ Emoji ที่ได้มาในรูปแบบตัวหนังสือ

- คำแสดงความลังเล คือ คำที่ทำหน้าที่ลดความชัดเจนหรือความมั่นใจของผู้ใช้ถ้อยคำ มักถูกใช้เพื่อป้องกันการโจมตีจากผู้ฟังหรือผู้อ่านที่มีความรู้มากกว่า [21], [22] โดยใช้คำที่มาจาก [23], [24] ซึ่งเป็นคำในภาษาอังกฤษ และใช้โปรแกรมแปลภาษา Google Translate แปลเป็นภาษาไทย มีทั้งสิ้น 200 คำ ดังตารางที่ 3 แสดงตัวอย่างคำแสดงความลังเลในภาษาอังกฤษและภาษาไทยที่ได้จากโปรแกรมแปลภาษา

ตารางที่ 3 ตัวอย่างคำแสดงความลังเลในภาษาอังกฤษและภาษาไทย

Hedging Word	คำแสดงความลังเล
assume	สมมุติ
believe	เชื่อว่า
definitely	อย่างแน่นอน
doubt	สงสัยว่า
estimate	ประมาณ
estimate	ประมาณว่า
indicate	ระบุ
It could be the case that	อาจจะเป็นอย่างนั้นก็ได้
may	อาจจะ
perhaps	บางที
possibility	เป็นไปได้

- คำหยุด คือ คำทั่วๆไปที่สามารถพบได้บ่อยในประโยค เป็นคำที่ไม่มีนัยยะสำคัญต่อความหมายในประโยค [12] เมื่อตัดออกจะไม่ทำให้ใจความสำคัญของประโยคเปลี่ยนไป เช่นคำว่า นั่นไง เหตุนั้น มาก ช้านาน เพียงใด ฯลฯ เป็นต้น ซึ่งการกำจัดคำหยุดนั้นเป็นหนึ่งในขั้นตอนการจัดเตรียมข้อมูลสำหรับการประมวลผลภาษาธรรมชาติ [11] แต่ในงานวิจัยนี้ผู้วิจัยจะนำคำหยุดมาสร้างเป็นคุณลักษณะสำหรับการสร้างแบบจำลอง โดยใช้คำหยุดภาษาไทยที่มีอยู่ในไลบรารี่ pyThaiNLP ซึ่งมีจำนวนทั้งสิ้น 1,030 คำ
- Emoji คือ สัญลักษณ์แทนอารมณ์ ความรู้สึก หรือแทนสิ่งต่างๆ เช่น 😆 ☺ ♥ 🍔 ซึ่งหลังจากทำการเก็บข้อมูลจะได้มาเป็นข้อความแทนที่สัญลักษณ์ เช่น :heart_suit: แทนสัญลักษณ์ ♥ ดังรูปที่ 7 ดังนั้นผู้วิจัยจึงต้องทำการแยกคำปกติกับ Emojis ออกจากกัน โดยทำการสกัดข้อความ Emoji แล้วทำการรวบรวมเป็นพจนานุกรม Emoji โดยมีทั้งสิ้น 789 emojis เพื่อใช้สำหรับการตัดคำ และสร้างเป็นคุณลักษณะสำหรับการสร้างแบบจำลอง ซึ่งมีวิธีการสกัดข้อความ Emojis ดังรูปที่ 9

```
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',
```

รูปที่ 9 ตัวอย่างโค้ดไพธอนสำหรับการสร้างพจนานุกรม Emoji และตัวอย่างคำศัพท์ Emojis

2) การตัดคำ

การตัดคำคือการแบ่งคำ ซึ่งในภาษาไทยนั้นจะไม่เหมือนกับการตัดคำในภาษาอังกฤษ เพราะว่าภาษาอังกฤษมีการเว้นระหว่างคำ จึงสามารถใช้ช่องว่างในการตัดคำ ส่วนในภาษาไทยนั้นจะต้องมีวิธีการตัดที่ต่างออกไป

ในงานวิจัยนี้ผู้เขียนได้เลือกใช้ไลบรารีสำหรับการประมวลผลภาษาธรรมชาติในภาษาไทยที่ชื่อว่า pyThaiNLP ซึ่งมีโมดูลสำหรับการตัดคำภาษาไทยอยู่ ผู้วิจัยเลือกใช้วิธีตัดคำแบบใช้พจนานุกรมเป็นพื้นฐาน (Dictionary base) ร่วมกับวิธีการตัดคำแบบสอดคล้องมากที่สุด(Maximal Matching) โดยทำการตัดคำสองแบบ คือ แบบที่หนึ่งใช้เฉพาะพจนานุกรมที่มาจากงานวิจัย [20] เพื่อใช้สร้างเป็นแบบจำลองพื้นฐานสำหรับการวัดประสิทธิภาพ และแบบที่สองเพิ่มคำศัพท์ของผู้วิจัยเองอีก 2 ชุดคำศัพท์ลงในพจนานุกรม คือ คำแสดงความลังเลและคำศัพท์ Emoji นอกจากนี้ยังมี Tagging ต่างๆ ที่ใช้ในขั้นตอนการทำความสะอาดข้อมูลเพิ่มเข้าไปด้วย

ตารางที่ 4 เปรียบเทียบข้อความที่ทำความสะอาดแล้วก่อนตัดคำและหลังตัดคำทั้งสองแบบ

Clean Comment	Word Segmentation	Word Segmentation with new words
สาเหตุที่เที่ยงวันติดยาเพราะประชดพ่อแม่พ่อแม่ไม่มีเวลาให้ประมาณนี้หรือเปล่า[mask]หรือสาเหตุหลักมาจากอะไร[mask]	[สาเหตุ, ที่, เที่ยงวัน, ติดยา, เพราะ, ประชด, พ่อแม่, พ่อแม่, ไม่มีเวลา, ให้, ประมาณ, นี้, หรือ เปล่า, [mask], หรือ, สาเหตุ, หลัก , มาจาก, อะไร, [mask]]	[สาเหตุ, ที่, เที่ยงวัน, ติดยา, เพราะ, ประชด, พ่อแม่, พ่อแม่, ไม่มีเวลา, ให้, ประมาณนี้, หรือ เปล่า, [mask], หรือ, สาเหตุ, หลัก , มาจาก, อะไร, [mask]]
gemtvasiatruevision[number]╬═ emojiheartsuit╬survivalwedding	[gemtvasiatruevision, [number], ╬═, emojiheartsuit, ╬,	[gemtvasiatruevision, [number], ╬═, emojiheartsuit, ╬,

╫emojiheartsuit═╫นำแสดงโดยทุกวันพุธเวลา[time]นเริ่มตอนแรกวันพุธที่[number]กรกฎาคมเรื่องนี้ที่ญี่ปุ่นฉายตอนแรกวันเสาร์ที่[number]กรกฎาคม[mask]ข้อมูลเลยไม่ค่อยมีอะไรเลย credit[website]	survivalwedding, ╫, emojiheartsuit, ═╫, นำแสดง, โดย, ทุก, วันพุธ, เวลา, [time], น, เริ่ม, ตอนแรก, วันพุธ, ที่, [number], กรกฎาคม, เรื่อง, นี้, ที่, ญี่ปุ่น, ฉาย, ตอนแรก, วันเสาร์, ที่, [number], กรกฎาคม, [mask], ข้อมูล, เลย, ไม่ค่อย, มีอะไร, เลย, credit, [website]]	survivalwedding, ╫, emojiheartsuit, ═╫, นำแสดง, โดย, ทุก, วันพุธ, เวลา, [time], น, เริ่ม, ตอนแรก, วันพุธ, ที่, [number], กรกฎาคม, เรื่อง, นี้, ที่, ญี่ปุ่น, ฉาย, ตอนแรก, วันเสาร์, ที่, [number], กรกฎาคม, [mask], ข้อมูล, เลย, ไม่ค่อย, มีอะไร, เลย, credit, [website]]

3) การระบุประเภทของคำ (Part-of-Speech Tagging)

เป็นการระบุประเภทของคำตามหลักไวยกรณ์ โดยอิงหน้าที่ของคำเป็นหลัก ในงานวิจัยนี้ผู้เขียนได้ใช้โมดูลของ pyThaiNLP และเลือกคลังคำศัพท์ ORCHID [25] ในการระบุประเภทของคำ ดังตารางที่ 5 แสดงประเภทของคำและตัวอย่างคำในคลังคำศัพท์ ORCHID และตารางที่ 6 แสดงตัวอย่างข้อความที่ระบุประเภทของคำแล้ว

ตารางที่ 5 ประเภทของคำและตัวอย่างคำในคลังคำศัพท์ ORCHID

ตัวย่อ	ประเภทของคำ (Part-of-Speech tag)	ตัวอย่างคำ
NPRP	Proper noun	วินโดวส์ 95, โคโรน่า, ไค้ก
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
XVAE	Post-verb auxiliary	ไป, มา, ขึ้น
DDAN	Definite determiner, after noun without classifier in between	นี่, นั่น, โน่น, ทั้งหมด
DDAC	Definite determiner, allowing classifier in between	นี้, นั้น, โน้น, นู้น
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, ด้าน, แบบ, รุ่น
CMTR	Measurement classifier	กิโลกรัม, แก้ว, ชั่วโมง
CFQC	Frequency classifier	ครั้ง, เที่ยว
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	ไอ้บ, ไอ้, เออ, เอ๋, อ๊อ
FIXN	Nominal prefix	การทำงาน, ความสนุกสนาน
FIXV	Adverbial prefix	อย่างเร็ว
EAFF	Ending for affirmative sentence	จ๊ะ, จ๋ะ, ค่ะ, ครับ, นะ, น่า, เถอะ
EITT	Ending for interrogative sentence	หรือ, เหรอ, ไหม, มั้ย
NEG	Negator	ไม่, มิได้, ไม่ได้, มิ
PUNC	Punctuation	(,), “, ,, ;

ตารางที่ 6 เปรียบเทียบข้อความที่ทำความสะอาด กับข้อความที่ระบุประเภทของคำแล้ว

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)

หลังจากที่ระบุประเภทของคำได้แล้ว จะทำการปกปิดคำสรรพนามแทนตัวด้วยการใช้ [PPRS] แทนคำสรรพนามแทนตัวที่ได้จากการระบุประเภทของคำ เพื่อป้องกันการ Overfitting จากคำเหล่านี้ ดังตารางที่ 7 แสดงข้อความหลังการปกปิดคำสรรพนามแทนตัวแล้ว

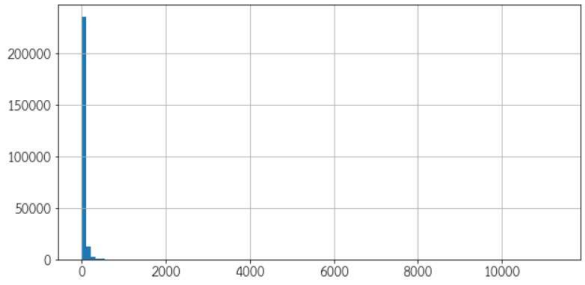
ตารางที่ 7 เปรียบเทียบข้อความที่ทำความสะอาด กับข้อความที่ตัดคำและทำการปกปิดคำสรรพนามแล้ว

Clean Comment	Word Segmentation with [PPRS]
ใช่[mask]เราว่ามีเสน่ห์มากจริงๆบอเล่นเรื่องไหนก็เข้าถึงดีบทแตกหมดนะเราว่าแต่เรื่องนี้เราไม่ค่อยได้เห็นไงเลยรู้สึกเซอร์ไพรสมากและชอบบทนี้ของบิ๊กเอ็มมากจริงๆ[mask]	[ใช่, [mask], [PPRS], ว่า, มีเสน่ห์, มาก, จริงๆ, บอ, เล่น, เรื่อง, ไหน, ก็, เข้าถึง, ดีบทแตก, 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]]

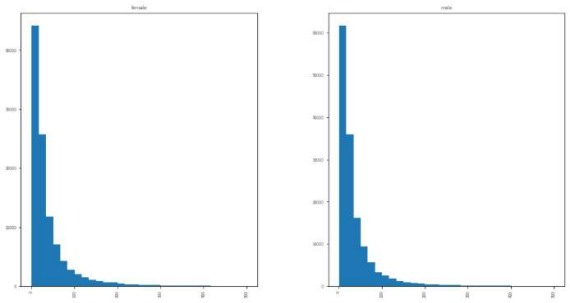
E. การสำรวจข้อมูลเบื้องต้น (Exploratory Data Analysis)

เมื่อทำการสำรวจการกระจายตัวของความยาวของข้อความเป็นจำนวนคำในข้อความทั้งหมดที่จะนำมาใช้ในการจำแนก พบว่าข้อความกระจุกตัวกันที่อยู่ที่ความยาวไม่มาก ดังรูป

ที่ 10 และรูปที่ 11 โดยมีค่าสถิติแสดงดังตารางที่ 8 จะเห็นได้ว่าเพศชายและเพศหญิงมีการกระจายตัวของข้อมูลค่อนข้างใกล้เคียงกันมาก เมื่อทำการทดสอบข้อมูลทางสถิติ T-Test พบว่าทั้งเพศชายและเพศหญิงมีค่าเฉลี่ยของความยาวของข้อความไม่แตกต่างกันอย่างมีค่านัยสำคัญทางสถิติ และทำการทดสอบ F-Test พบว่ามีค่าความแปรปรวนของความยาวของข้อความไม่แตกต่างกันอย่างมีค่านัยสำคัญทางสถิติ ซึ่งมีผลการทดสอบทางสถิติดังรูปที่ 12



รูปที่ 10 การกระจายตัวของจำนวนคำของข้อความทั้งหมด



รูปที่ 11 การกระจายตัวของจำนวนคำของข้อความที่มีจำนวนคำไม่เกิน 500 คำ เพศหญิง (ซ้าย) เพศชาย (ขวา)

ตารางที่ 8 เปรียบเทียบค่าสถิติการกระจายตัวของข้อความของผู้เขียนข้อความในเพศชายและเพศหญิง

	เพศชาย	เพศหญิง	รวม
count	143,006	104,904	247,910
mean	40.4	43.3	41.6
STD	77.9	98.8	87.4
min	1	1	1
25%	11	11	11
50%	21	22	21
75%	43	46	45
max	2,793	8,850	8,850

```
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)
```

รูปที่ 12 ผลการทดทางสถิติ T-Test และ F-Test ของจำนวนคำในข้อความระหว่างเพศชายและเพศหญิง

F. การแบ่งตัวอย่างข้อมูลสำหรับการสร้างแบบจำลอง

อ้างอิงจากงานวิจัย [6] ที่กล่าวว่าความแตกต่างของการใช้ภาษาระหว่างเพศหญิงและชายจะแตกต่างอย่างมีนัยยะก็ต่อเมื่อมีความยาวของการใช้ภาษามากพอ ผู้วิจัยจึงได้ทำการแบ่งข้อมูลออกตามจำนวนคำของข้อความ ออกเป็น 4 กลุ่ม ดังตารางที่ 9 โดยใช้ข้อมูลจากการกระจายตัวของข้อความที่ได้จากขั้นตอนการสำรวจข้อมูลดังนี้

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

หลังจากนั้นจึงแบ่งตัวอย่างในแต่ละกลุ่มออกเป็นชุดข้อมูลสำหรับฝึกฝน (Training dataset) และชุดข้อมูลสำหรับทดสอบ (Testing dataset) โดยมีอัตราส่วน 80:20

ตารางที่ 9 จำนวนข้อความของแต่ละกลุ่มตัวอย่างแยกเพศชายและเพศหญิง

ความยาวข้อความ (คำ)	ความยาวเฉลี่ย (คำ)	จำนวนข้อความ		
		เพศชาย	*เพศหญิง*	*รวม*
< 11	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

IV. การสกัดคุณลักษณะ

เป็นการเปลี่ยนข้อมูลที่มีให้กลายเป็นคุณลักษณะสำหรับการสร้างแบบจำลองการจำแนกผู้วิจัยได้ทำการสกัดคุณลักษณะโดยใช้เทคนิค TF-IDF (n-gram 1,2) ได้ดังนี้

- สกัดคุณลักษณะจากข้อความปกติ ได้คุณลักษณะจำนวน 5,000 คุณลักษณะ ดังรูปที่ 13 แสดงตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-gram 1,2) กับคำทั่วไป

[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	...

รูปที่ 13 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับคำทั่วไป

- สกัดคุณลักษณะจาก Emoji ได้คุณลักษณะจำนวน 789 คุณลักษณะ ดังรูปที่ 14 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	

รูปที่ 14 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับ emojis

- สกัดคุณลักษณะจากคำแสดงความลังเลได้คุณลักษณะจำนวน 200 คุณลักษณะ ดังรูปที่ 15 แสดงตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-gram 1,2) กับคำแสดงความลังเล

ก็	ก็ตามที	กระจ่างแจ้ง	กระจ่างชัด	กะจ่ายเลย	กิจวัตร	ขอให้	คาหนังคาเขา	คง	คงจะ	...	อะไรอย่างงี้	อะไรอย่างนั้น	อะไรอย่างนี้
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0
0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0

รูปที่ 15 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับคำแสดงความลังเล

- สกัดคุณลักษณะจากคำหยุดได้คุณลักษณะจำนวน 1,030 คุณลักษณะ ดังรูปที่ 16 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

รูปที่ 16 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับคำหยุด

- สกัดคุณลักษณะจากประเภทของคำได้คุณลักษณะจำนวน 1,239 คุณลักษณะ ดังรูปที่ 17 แสดงตัวอย่างคุณลักษณะที่ได้จากการสกัดคุณลักษณะด้วยวิธี TF-IDF (n-gram 1,2) กับประเภทของคำ

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

รูปที่ 17 คุณลักษณะที่สกัดได้จากเทคนิค TF-IDF (n-gram 1,2) กับประเภทของคำ

V. แบบจำลองการจำแนก

ในขั้นตอนการสร้างแบบจำลอง ผู้วิจัยได้ทำการเลือกอัลกอริทึมสำหรับงานประเภทการจำแนกจำนวน 3 อัลกอริทึม คือ Logistic Regression, Naive Bayes และ Random Forest โดยจะใช้คุณลักษณะทั้ง 5 ประเภทที่สกัดได้ สลับไปมาเพื่อทดลองว่าคุณลักษณะใดและอัลกอริทึมใดที่ให้ประสิทธิภาพในการจำแนกได้มากที่สุด และยังใช้อัลกอริทึมสำหรับการเลือกคุณลักษณะอีก 2 อัลกอริทึม คือ SelectKBest และ Recursive Feature Elimination with Cross-Validation (RFECV) โดยได้ทำการวัดประสิทธิภาพของแบบจำลองโดยใช้ cross-validation score บนชุดข้อมูลฝึกฝน และ Testing Score บนชุดข้อมูลทดสอบ โดยมีชุดคุณลักษณะที่ใช้ทดสอบดังนี้

- emoji
- คำแสดงความลังเล
- คำหยุด
- ประเภทของคำ
- TF-IDF (สำหรับค่าประสิทธิภาพพื้นฐาน(Baseline))
- TF-IDF+คำศัพท์ใหม่
- TF-IDF+คำศัพท์ใหม่, ประเภทของคำ
- TF-IDF+คำศัพท์ใหม่, คำหยุด
- TF-IDF+คำศัพท์ใหม่, ประเภทของคำ, คำหยุด
- TF-IDF+คำศัพท์ใหม่, emoji
- TF-IDF+คำศัพท์ใหม่, emoji, ประเภทของคำ
- TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด
- TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด, ประเภทของคำ
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, ประเภทของคำ
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, คำหยุด
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, ประเภทของคำ, คำหยุด
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, ประเภทของคำ
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด
- TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ

VI. ผลลัพธ์การทดลองและสรุปผลการทดลอง

ผลลัพธ์ที่ได้จากการทดสอบแบบจำลองการจำแนกกับชุดข้อมูลทดสอบ วัดประสิทธิภาพของแบบจำลองโดยใช้ค่า Accuracy, Precision, Recall และ F_1 Score ได้ผลลัพธ์ออกมาดังตารางที่ 10 ตารางที่ 11 ตารางที่ 12 และตารางที่ 13 ตามกลุ่มข้อมูลที่มีความยาวข้อความจากน้อยไปมากตามลำดับ

ตารางที่ 10 เปรียบเทียบค่าประสิทธิภาพพื้นฐาน กับประสิทธิภาพของแบบจำลองที่ดีที่สุด 3 ชุดคุณลักษณะ ในแต่ละอัลกอริทึม ที่สร้างบนชุดข้อมูลกลุ่มที่ 1 ที่มีความยาวน้อยกว่า 11 คำ

กลุ่มที่ 1 ข้อความที่มีจำนวนคำน้อยกว่าเปอร์เซนไทล์ที่ 25 (น้อยกว่า 11 คำ)					
อัลกอริทึม	***คุณลักษณะ***	***Accuracy***	***Precision***	***Recall***	***F1 Score***
Logistic Regression	TF-IDF (สำหรับ baseline)	0.6564	0.6506	0.6564	0.6428
	TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด, ประเภทของคำ	0.6713	0.6662	0.6713	0.6636
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6702	0.6651	0.6702	0.6626
Logistic Regression + SelectKBest	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6723	0.6673	0.6723	0.6642
Naive Bayes	TF-IDF (สำหรับ baseline)	0.6560	0.6518	0.6560	0.6376
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด	0.6694	0.6666	0.6694	0.6532
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6679	0.6633	0.6679	0.6558
Naive Bayes + SelectKBest	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6683	0.6637	0.6683	0.6561
Random Forest	TF-IDF (สำหรับ baseline)	0.5888	0.7441	0.5888	0.4413
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด	0.5892	0.7393	0.5892	0.4424
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.5861	0.7577	0.5861	0.4350
Random Forest + RFECV	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6096	0.6216	0.6096	0.5257

ตารางที่ 11 เปรียบเทียบค่าประสิทธิภาพพื้นฐาน กับประสิทธิภาพของแบบจำลองที่ดีที่สุด 3 ชุดคุณลักษณะ ในแต่ละอัลกอริทึม ที่สร้างบนชุดข้อมูลกลุ่มที่ 2 ที่มีความยาว 11-96 คำ

กลุ่มที่ 2 ข้อความที่มีจำนวนคำมากกว่าเปอร์เซนไทล์ที่ 25 ถึง Q3+1.5IQR (จำนวนคำระหว่าง 11-96)					
อัลกอริทึม	***คุณลักษณะ***	***Accuracy***	***Precision***	***Recall***	***F1 Score***
	TF-IDF (สำหรับ baseline)	0.7450	0.7438	0.7450	0.7412

Logistic Regression	TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด, ประเภทของคำ	0.7662	0.7649	0.7662	0.7640
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7655	0.7643	0.7655	0.7633
Logistic Regression + SelectKBest	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7657	0.7644	0.7657	0.7635
Naïve Bayes	TF-IDF (สำหรับ baseline)	0.7308	0.7304	0.7308	0.7248
	TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด	0.7391	0.7407	0.7391	0.7319
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7518	0.7521	0.7518	0.7502
Naïve Bayes + SelectKBest	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7518	0.7521	0.7518	0.7502
Random Forest	TF-IDF (สำหรับ baseline)	0.6040	0.7157	0.6040	0.4848
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด	0.6141	0.7282	0.6141	0.5053
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6114	0.7333	0.6114	0.4986
Random Forest + RFECV	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6564	0.7162	0.6564	0.5955

ตารางที่ 12 เปรียบเทียบค่าประสิทธิภาพพื้นฐาน กับประสิทธิภาพของแบบจำลองที่ดีที่สุด 3 ชุดคุณลักษณะ ในแต่ละอัลกอริทึม ที่สร้างบนชุดข้อมูลกลุ่มที่ 3 ที่มีความยาว 97-200 คำ

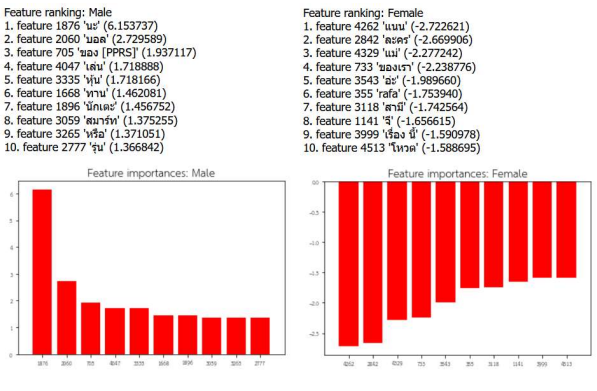
กลุ่มที่ 3 ข้อความที่มีจำนวนคำมากกว่า Q3+1.5IQR ถึง 200 คำ (จำนวนคำระหว่าง 97-200)					
อัลกอริทึม	*คุณลักษณะ*	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Logistic Regression	TF-IDF (สำหรับ baseline)	0.7752	0.7750	0.7752	0.7745
	TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด, ประเภทของคำ	0.7895	0.7893	0.7895	0.7890
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7889	0.7886	0.7889	0.7884
Logistic Regression + SelectKBest	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7904	0.7903	0.7904	0.7898
Naïve Bayes	TF-IDF (สำหรับ baseline)	0.7582	0.7577	0.7582	0.7577
	TF-IDF+คำศัพท์ใหม่, emoji, คำหยุด	0.7588	0.7587	0.7588	0.7577
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7366	0.7374	0.7366	0.7298
Naïve Bayes + RFECV	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7366	0.7374	0.7366	0.7298
Random Forest	TF-IDF (สำหรับ baseline)	0.6907	0.7364	0.6907	0.6638
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, คำหยุด	0.7085	0.7554	0.7085	0.6848
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6961	0.7429	0.6961	0.6699
Random Forest + RFECV	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7255	0.7460	0.7255	0.7134

ตารางที่ 13 เปรียบเทียบค่าประสิทธิภาพพื้นฐาน กับประสิทธิภาพของแบบจำลองที่ดีที่สุด 3 ชุดคุณลักษณะ ในแต่ละอัลกอริทึม ที่สร้างบนชุดข้อมูลกลุ่มที่ 3 ที่มีความมากกว่า 200 คำ

กลุ่มที่ 4 ข้อความที่มีจำนวนคำมากกว่า 200 คำ					
อัลกอริทึม	*คุณลักษณะ*	*Accuracy*	*Precision*	*Recall*	*F1 Score*
Logistic Regression	TF-IDF (สำหรับ baseline)	0.7604	0.7609	0.7604	0.7587
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, ประเภทของคำ, คำหยุด	0.7903	0.7911	0.7903	0.7889
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7894	0.7903	0.7894	0.7880
Logistic Regression + SelectKBest	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7852	0.7863	0.7852	0.7835
Naïve Bayes	TF-IDF (สำหรับ baseline)	0.7298	0.7298	0.7298	0.7276
	TF-IDF+คำศัพท์ใหม่, คำหยุด	0.7451	0.7466	0.7451	0.7422
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7434	0.7466	0.7434	0.7395
Naïve Bayes + SelectKBest	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7434	0.7464	0.7434	0.7396
Random Forest	TF-IDF (สำหรับ baseline)	0.6777	0.7027	0.6777	0.6557
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด	0.6897	0.7188	0.6897	0.6682
	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.6820	0.7161	0.6820	0.6567
Random Forest + RFECV	TF-IDF+คำศัพท์ใหม่, คำแสดงความลังเล, emoji, คำหยุด, ประเภทของคำ	0.7153	0.7285	0.7153	0.7045

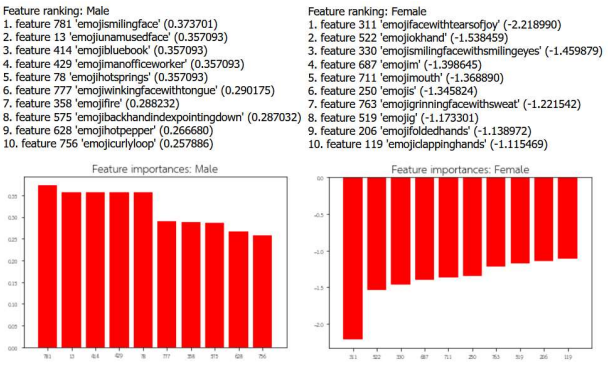
จากตารางผลการทดลองจะเห็นว่า ค่าประสิทธิภาพของแบบจำลองการจำแนกเพศของผู้เขียนข้อความสูงสุดมีค่า Accuracy 79.04%, Precision 79.03%, Recall 79.04% และ F1 Score 78.98% โดยใช้คุณลักษณะที่สกัดจากคำศัพท์ใหม่ Emoji คำแสดงความลังเล คำหยุด และประเภทของคำ ร่วมกับอัลกอริทึม Logistic Regression และใช้อัลกอริทึม SelectKBest ในการเลือกคุณลักษณะที่ดีที่สุด บนข้อความที่มีความยาว 97-200 คำ

คำที่มีผลต่อการจำแนกเพศสามารถดูได้จากค่าความสำคัญของคุณลักษณะ (Feature Importance) เมื่อดูจากแบบจำลองที่ให้ค่าประสิทธิภาพสูงสุด ซึ่งใช้อัลกอริทึม Logistic Regression สามารถดูจากค่าสัมประสิทธิ์ของคุณลักษณะ (Coefficient of the features) โดยคุณลักษณะที่มีค่าสัมประสิทธิ์สูง จะมีผลต่อการจำแนกเพศชาย ในขณะที่คุณลักษณะที่มีค่าสัมประสิทธิ์ต่ำจะมีผลต่อการจำแนกเพศหญิงที่ ดังรูปที่ 18 โดยคำที่มีผลต่อการจำแนก 10 อันดับแรกในเพศชาย ได้แก่ นะ, บอล, ของ, เล่น, หุ้น, ทาน, นักเตะ, สมาร์ท, หรือ, รุ่น และ 10 อันดับแรกในเพศหญิง ได้แก่ แนน, ละคร, แม่, ของเรา, อ่ะ, rafa, สามี, จี, เรื่องนี้, โหวต

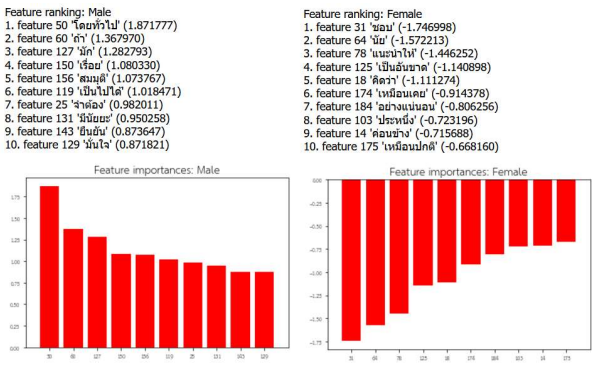


รูปที่ 18 ค่าความสำคัญของคุณลักษณะแบบรวมทุกคุณลักษณะที่มีผลต่อการจำแนกเพศ 10 อันดับแรก สำหรับเพศชาย (ซ้าย) และเพศหญิง (ขวา)

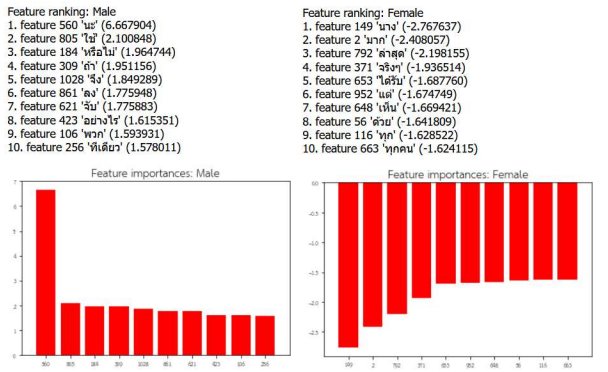
สำหรับ 10 อันดับคุณลักษณะที่มีผลต่อการจำแนกเพศชายและเพศหญิงนอกจากแบบรวมทุกคุณลักษณะที่แสดงดังรูปที่ 18 แล้ว สามารถดู 10 อันดับคุณลักษณะที่มีผลต่อการจำแนกแยกตามประเภทคุณลักษณะ ได้แก่ Emoji คำแสดงความสังเล คำหยุด และประเภทของคำ แสดงดังรูปที่ 19 รูปที่ 20 รูปที่ 21 และรูปที่ 22 ตามลำดับ



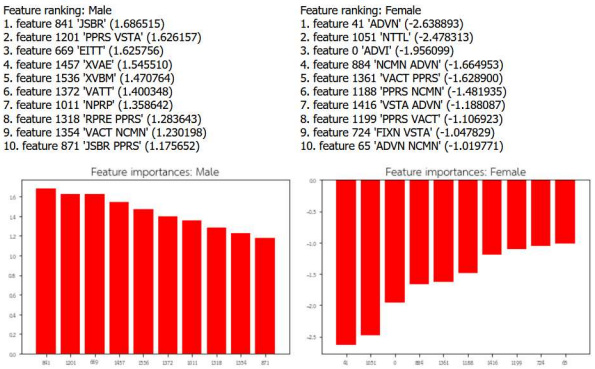
รูปที่ 19 ค่าความสำคัญของคุณลักษณะเฉพาะ Emoji ที่มีผลต่อการจำแนกเพศ 10 อันดับแรก สำหรับเพศชาย (ซ้าย) และเพศหญิง (ขวา)



รูปที่ 20 ค่าความสำคัญของคุณลักษณะเฉพาะคำแสดงความสังเลที่มีผลต่อการจำแนกเพศ 10 อันดับแรก สำหรับเพศชาย (ซ้าย) และเพศหญิง (ขวา)



รูปที่ 21 ค่าความสำคัญของคุณลักษณะเฉพาะคำหยุดที่มีผลต่อการจำแนกเพศ 10 อันดับแรก สำหรับเพศชาย (ซ้าย) และเพศหญิง (ขวา)



รูปที่ 22 ค่าความสำคัญของคุณลักษณะเฉพาะประเภทของคำที่มีผลต่อการจำแนกเพศ 10 อันดับแรก สำหรับเพศชาย (ซ้าย) และเพศหญิง (ขวา)

VII. อภิปรายผล

ค่าประสิทธิภาพสูงสุดของแบบจำลองการ มีค่ามากกว่าค่าประสิทธิภาพพื้นฐานที่ใช้คุณลักษณะที่สกัดจากพจนานุกรมจากงานวิจัยเดิม [20] 1.52% บนชุดข้อมูลและอัลกอริทึมเดียวกัน และเมื่อเปรียบเทียบกับ State-of-The-Art ในภาษาอังกฤษ [15] จะได้ค่าประสิทธิภาพของแบบจำลองน้อยกว่า 9.52% อาจเป็นเพราะข้อความที่นำมาใช้ในงานวิจัยนี้ ยังมีการระบุเพศของผู้เขียนข้อความได้ไม่ถูกต้องเท่าที่ควร ในขณะที่งานวิจัย State-of-The-Art ใช้ข้อมูลส่วนตัวของผู้เขียนข้อความ และ/หรือรูปโปรไฟล์ของผู้เขียนข้อความในการระบุเพศ และยังมีการระบุเพศที่ละข้อความ ทำให้มีความถูกต้องของข้อมูลที่นำมาใช้ในการวิจัยเป็นอย่างมาก

ในภาพรวมค่าประสิทธิภาพจะเพิ่มขึ้นตามความยาวของข้อความ เมื่อเปรียบเทียบในอัลกอริทึมเดียวกันและชุดคุณลักษณะเดียวกัน ซึ่งสอดคล้องกับ [6] ที่กล่าวว่าความแตกต่างของการใช้ภาษาระหว่างเพศหญิงและชายจะแตกต่างอย่างมีนัยยะก็ต่อเมื่อมีความยาวของการใช้ภาษามากพอ คือเมื่อข้อความที่นำมาวิเคราะห์มีความยาวไม่เพียงพอ จะทำให้ค่าประสิทธิภาพที่ได้มีค่าไม่มาก (ไม่มีความแตกต่างอย่างมีนัยยะ)

เมื่อพิจารณาคำที่มีผลต่อการจำแนกเพศ จะเห็นได้ว่าสามารถบ่งบอกถึงสิ่งที่แต่ละเพศสนใจได้เป็นอย่างดี อาทิเช่น เพศชายสนใจในเรื่องกีฬาฟุตบอล หุ้นและโทรศัพท์มือถือ (สมาร์ทโฟน) ส่วนเพศหญิงสนใจในเรื่องความบันเทิง เช่น ละคร ดารา การประกวดต่างๆ เป็นต้น

VIII. งานวิจัยในอนาคต

นำงานวิจัยนี้ไปพัฒนาแบบจำลอง โดยทำการเพิ่มข้อมูลการฝึกฝน หาวิธีระบุเพศให้มีความถูกต้องมากยิ่งขึ้น ทดลองเปลี่ยนอัลกอริทึมทั้งอัลกอริทึมการตัดคำ การสกัดคุณลักษณะและอัลกอริทึมที่ใช้สร้างแบบจำลอง อาจใช้แบบจำลองทางภาษาก่อนการฝึกฝนและการเรียนรู้เชิงลึก (Deep Learning) ร่วมด้วย เพื่อเพิ่มประสิทธิภาพของแบบจำลอง

อ้างอิง

[1] J. Fahy, and D. Jobber, “Understanding Customer Behavior,” in Foundations of marketing, 5th ed. Maidenhead, UK: McGraw-Hill Education, 2015, ch. 1, sec. 3, pp. 129-165.

[2] Z. Arsel, K. E., and J. Moisander, Gendering Theory in Marketing and Consumer Research. Routledge, 2017.

[3] Knowledge@Wharton. (2007, Nov 28), “Men Buy, Women Shop”: The Sexes Have Different Priorities When Walking Down the Aisles. [Online]. Available: https://knowledge.wharton.upenn.edu/article/men-buy-women-shop-the-sexes-have-different-priorities-when-walking-down-the-aisles/

[4] E. Koc, “The impact of gender in marketing communications: The role of cognitive and affective cues,” Journal of Marketing Communications, vol. 8, pp. 257-275, 2002.

[5] D. Fridh, and T. D., “A consumer perspective of personalized marketing: An exploratory study on the consumer perception of personalized marketing and how it affects the consumer decision-making process,” B.Sc. in BA thesis, Kristianstad Univ., Sweden, 2019.

[6] F. Crosby, and L. Nyquist, “The female register: an empirical study of Lakoff’s hypotheses,” Language in Society, vol. 6, issue 3, pp. 313-322, 1977.

[7] อ. ประสิทธิ์รัฐสินธุ์, ภาษาในสังคมไทย ความหลากหลาย การเปลี่ยนแปลง และการพัฒนา. กรุงเทพมหานคร: สำนักพิมพ์แห่งจุฬาลงกรณ์มหาวิทยาลัย, 2545.

[8] น. ดาดทอง, "เพศกับกลวิธีการใช้ภาษาบนเฟซบุ๊ก,". ปริญญานิพนธ์ กศ.ม. ภาษาศาสตร์การศึกษา, บัณฑิตวิทยาลัย, มหาวิทยาลัยศรีนครินทรวิโรฒ, กรุงเทพมหานคร, 2558

[9] ช. เศรษฐธัญการ, "การศึกษาเปรียบเทียบความสามารถในการใช้ภาษาระหว่างเพศหญิงและเพศชายที่สะท้อนให้เห็นภาพพจน์ทางเพศ," วารสารมังรายสาร, มหาวิทยาลัยราชภัฏพิบูลสงคราม, ปีที่ 7, ฉบับที่ 2, หน้า 17-31, 2562.

[10] ว. เฟื่องขจรศักดิ์, "การใช้ถ้อยคำแสดงความลังเลของผู้หญิงและผู้ชายในบริบทการวิพากษ์วิจารณ์: The use of hedging utterance of women and men in critical context," วารสารมนุษยศาสตร์และสังคมศาสตร์ มหาวิทยาลัยรังสิต, ปีที่ 11, ฉบับที่ 19, หน้า 44-60, 2558.

[11] M. K. Dalal, and M. A. Zaveri, "Automatic Text Classification: A Technical Review," International Journal of Computer Applications, vol. 28, pp. 37-40, 2011.

[12] น. จิระวิชิตชัย, ป. สงวนสัตย์, และ พ. มีสัจ, "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", กรุงเทพมหานคร, ประเทศไทย, 2553, pp. 302-307.

[13] A. C. Müller, and S. Guido, Introduction to Machine Learning with Python: A Guide for Data Scientists, 1st ed. O'Reilly Media, 2016, pp. 58-90.

[14] A. Géron, Hands-On Machine Learning with Scikit-Learn and TensorFlow, 1st ed. O'Reilly Media, 2017, pp. 86-89.

[15] A. Mukherjee, and B. Liu, "Improving Gender Classification of Blog Authors," Paper presented at the Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing, Cambridge: Association for Computational Linguistics, 2010, pp. 207-217.

[16] A. Poulston, Z. W., and M. Stevenson, "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, 2017

[17] R. Veenhoven, S. S., D. D. Hall, and R. Noord, "Word unigram weighing for author profiling," Paper presented at the CLEF 2018 Evaluation Labs and Workshop, Avignon, France, 2018

[18] Z. Chen, X. Lu, W. Ai, H. Li, Q. Mei, and X. Liu, "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, 2018.

[19] pyThaiNLP. [Online]. Available: https://pythainlp.github.io/

[20] T. Horsuwan, K. Kanwatchara, P. Vateekul, and B. Kijsirikul, A Comparative Study of Pretrained Language Models on Thai Social Text Categorization, arXiv, 2019.

[21] พ. ทองพูล, "รูปเบี่ยงบังและหน้าที่ในแผ่นพับโฆษณาผลิตภัณฑ์เสริมอาหารลดน้ำหนัก : HEDGES AND THEIR FUNCTIONS IN DIETARY SUPPLEMENT FOR WEIGHT LOSS ADVERTISING BROCHURES,". ปริญญานิพนธ์ ศศ.ม. ภาษาไทย, คณะศิลปะศาสตร์, มหาวิทยาลัยธรรมศาสตร์, กรุงเทพมหานคร, 2559

[22] น. กฤษณา. (2561, พ.ค. ๑). เฮดจ์ (Hedge). [Online]. Available: https://www.facebook.com/ThinkerTinkersLArtsTU/photos/a.365206150655618/367313343778232

[23] S. Smith. (2020, Dec 29). Hedging. [Online]. Available: https://www.eapfoundation.com/writing/skills/hedging/

[24] Features of academic writing: Hedging. [Online]. Available: http://www.uefap.com/writing/feature/hedge.htm

[25] S. Virach, T. Naoto, and I Hitoshi, "Building a Thai Part-Of-Speech Tagged Corpus (ORCHID)," The Journal of the Acoustical Society of Japan (E), vol. 20, no. 3, pp. 189-198, 1999.